Studying Laws of Semantic Divergence across Languages using Cognate Sets

Ana Sabina Uban, Alina Maria Ciobanu, and Liviu P. Dinu

Faculty of Mathematics and Computer Science,
Human Language Technologies Research Center,
University of Bucharest

ana.uban@gmail.com, alina.ciobanu@my.fmi.unibuc.ro,
liviu.p.dinu@gmail.com

Abstract

Semantic divergence in related languages is a key concern of historical linguistics. Intralingual semantic shift has been previously studied in computational linguistics, but it can only provide a limited picture of the evolution of word meanings, which often develop in a multilingual environment. In this paper we investigate semantic change across languages by measuring the semantic distance of cognate words in multiple languages. By comparing current meanings of cognates in different languages, we hope to uncover information about their previous meanings, and about how they diverged within their respective languages from their common original etymon. We further study the properties of the semantic divergence of cognates, by analyzing how features of the words, such as frequency and polysemy, are related to their shift in meaning, and thus take the first steps towards formulating laws of cross-lingual semantic change.

1 Introduction and Related Work

Semantic change – that is, change in the meaning of individual words (Campbell, 1998) – is a continuous, inevitable process stemming from numerous reasons and influenced by various factors. Words are continuously changing, with new senses emerging all the time. Campbell (1998) presents no less than 11 types of semantic change, that are generally classified in two wide categories: narrowing and widening.

In recent years, multiple computational linguistic studies have focused on the issue of semantic change, tracking the shift in the meaning of words by looking at their usage across time in corpora dating from different time periods. More than this, computational linguists have also tried to systematically analyze the principles describing semantic change hypothesized by linguists (such as the law of parallel change and the law of differentiation (Xu and Kemp, 2015)), or even proposed new statistical laws of semantic change, based on empirical observations, such as the law of conformity (stating that polysemy is positively correlated with semantic change), the law of innovation (according to which word frequency is negatively correlated with semantic change) (Hamilton et al., 2016), or the law of prototypicality (according to which prototypicality is negatively correlated with semantic change) (Dubossarsky et al., 2015). More recently, Dubossarsky et al. (2017) revisited some of the semantic change laws proposed in previous literature, claiming that a more rigorous consideration of control conditions when modelling these laws leads to the conclusion that they are weaker or less reliable than reported. More extensive surveys of computational studies relating to semantic change have been conducted by Kutuzov et al. (2018); Tahmasebi et al. (2018).

All previous computational studies on lexical semantic change have, to our knowledge, only looked at the semantic change of the words within one language. However, words do not evolve only in their own language in isolation, but are rather inherited and borrowed between and across languages.

Cognates are words in sister languages (languages descending from a common ancestor) with a common proto-word. For example, the Romanian word *victorie* and the Italian word *vittoria* are cognates, as they both descend from the Latin word *victoria* (meaning *victory*) – see Figure 1. In most cases, cognates have preserved similar meanings across languages, but there are also exceptions. These are called deceptive cognates or, more commonly, false friends. Here we use the definition of cognates that refers to words with similar appearance and some common etymology, and use *true cognates* to refer to cognates

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 161–166
Florence, Italy, August 2, 2019.

which also have a common meaning, and *deceptive cognates* or *false friends* to refer to cognate pairs which do not have the same meaning (anymore).

Dominguez and Nerlich (2002) distinguish between *chance false friends*, which have similar form but different etymologies as well as different meanings in different languages, and *semantic false friends*, which share the etymological origin, but their meanings differ (to some extent) in different languages. In this study we focus on the latter, which we consider more relevant from the point of view of semantic change since, in principle, they begin with a common meaning then diverge, to a lower or higher degree, while often preserving some common meaning, whereas *chance false friends* usually have entirely distinct meanings.

Most linguists found structural and psychological factors to be the main cause of semantic change, but the evolution of technology and cultural and social changes are not to be omitted. Moreover, when a word enters a new language, features specific to that particular language can affect the way it is used and contribute to shaping its meaning through time: existing words in the same language, as well as socio-cultural and historical factors etc. The evolution of cognate words in different languages can be seen as a collection of different parallel histories of the proto-word from its entering the new languages to its current state. Based on this view, we propose a novel approach for studying semantic change: instead of comparing *monolingual* texts from *different time periods* as ways to track meanings of words at different stages in time - we compare *present meanings* of cognate words across *different languages*, viewing them as snapshots in time of each of the word's different histories of evolution.

Related to our task, 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 have been 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).

Uban et al. (2019) propose a method for identifying and correcting false friends, as well as define a measure of their "falseness", using cross-lingual word embeddings. We base our study on the method proposed here, and take it further by analyzing the properties of semantic divergence as they relate to different properties of the words, across five Romance languages, as well as English. Similarly to how Hamilton et al. (2016) formulate statistical laws of semantic change within one language, we propose studying the same laws cross-lingually, from the point of view of cognate semantic divergence.

In the following sections, we first present the method for measuring cognate semantic distance in Section 2, then in Section 3 provide details on our experiments for characterizing the properties of semantic change across languages using cognates.

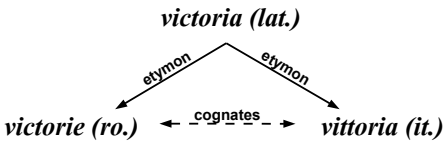

Figure 1: Example of cognates and their common ancestor.

2 Semantic Divergence of Cognates

2.1 Cross-lingual Word Embeddings

Word embeddings are vectorial representations of words in a continuous space, built by training a model to predict the occurrence of a target word in a text corpus given its context, and can be used as representations of word meaning: words that are similar semantically appear close together in the embedding space.

In our study we make use of word embeddings computed using the FastText algorithm, pre-trained on Wikipedia for the six languages in question. The vectors have 300 dimensions, and were obtained using the skip-gram model described by Bojanowski et al. (2016) with default parameters. These pre-trained embeddings are suitable for our study since: they are trained on large amounts of text, which minimizes the amount of noise in the vectors, making them good approximators of word meanings; and they are trained on text that is relatively uniform in style and topic - ensuring

Romanian	French	Italian	Spanish	Portuguese	Latin ancestor
arhitect	architecte	architetto	arquitecto	arquiteto	architectus

Table 1: An example of a cognate set: "architect" in Romance languages.

any differences in the structure of the embedding spaces of different languages is dependent on the language, rather than an artifact of topic or genre. Nevertheless, even high quality embeddings can be noisy or biased and this should be kept in mind when interpreting the results of our experiments.

To compute the semantic divergence of cognates across sister languages, we need to obtain a multilingual semantic space, which is shared between the cognates. Having the representations of both cognates in the same semantic space, we can then compute the semantic distance between them using their vectorial representations in this space. For a given pair of languages among the six considered, we can then accomplish this following the steps below:

Step 1. Obtain word embeddings for each of the two languages.

Step 2. Obtain a shared embedding space, common to the two languages. This is accomplished using an alignment algorithm, which consists of finding a linear transformation between the two spaces that on average optimally transforms each vector in one embedding space into a vector in the second embedding space, minimizing the distance between a few seed word pairs (which are assumed to have the same meaning), based on a small bilingual dictionary. The linear nature of the transformation guarantees distances between words in the original spaces (within each language) are preserved. For our purposes, we use the publicly available FastText multilingual word embeddings pre-aligned in a common vector space (Conneau et al., 2017).[1]

Step 3. Compute the semantic distance for the pair of cognates in the two languages, using a vectorial distance (we chose cosine distance) on their corresponding vectors in the shared embedding space.

2.2 Dataset

As our data source, we use the list of cognate sets in Romance languages proposed by Ciobanu and Dinu (2014). It contains 3,218 complete cognate sets in Romanian, French, Italian, Spanish and Portuguese, along with their Latin common ancestors, extracted from online etymology dictionaries. A subset of 305 of these sets also contains the corresponding cognate (in the broad sense, since these are mostly borrowings) in English.

One complete example of a cognate set for the word "architect" in the Romance languages is illustrated in Table 1.

2.3 Deceptive Cognates and Falseness

The multilingual embedding spaces as defined above can be used to measure the semantic distances between cognates in order to detect pairs of false friends, which are simply defined as pairs of cognates which do not share the same meaning. More specifically, following the false friends detection and correction algorithm of Uban et al. (2019), we consider a pair of cognates to be a false friend pair if in the shared semantic space, there exists a word in the second language which is semantically closer to the original word than its cognate in that language (in other words, the cognate is not the optimal translation). The arithmetic difference between the semantic distance between these words and the semantic distance between the cognates will be used as a measure of the *falseness* of the false friend.

	Accuracy	Precision	Recall
EN-ES	76.58	63.88	88.46
ES-IT	75.80	41.66	54.05
ES-PT	82.10	40.0	42.85
EN-FR	77.09	57.89	94.28
FR-IT	74.16	32.81	65.62
FR-ES	73.03	33.89	69.96
EN-IT	73.07	33.76	83.87
IT-PT	76.14	29.16	43.75
EN-PT	77.25	59.81	86.48

Table 2: Performance for all language pairs using WordNet as gold standard.

Uban et al. (2019) also perform an evaluation of the introduced false friends detection algorithm using multilingual WordNet as a gold standard. In order to provide more context for the method that we employ in our study, we briefly reiterate their results. A pair of words with common etymology are considered true cognates if they belong to the

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

same WordNet synset (are synonyms), and false friends if they are not synonyms. Using this gold standard, the obtained measured accuracy falls between 74% and 82%, depending on the language pair considered. Table 2 presents a breakdown of the obtained performance per language pair considered (limited to languages available in multilingual WordNet).

We select a few results of the algorithm to show in Table 3, containing examples of extracted false friends, along with the suggested correction and the computed degree of falseness. Each row in the table contains a pair of false cognates, among which one is chosen as a reference, and corrected so as to obtain its true translation in the second language using the correction algorithm.

Cognate	**False Friend**	**Correction**	**Falseness**
long (FR)	luengo (ES)	largo	0.50
face (FR)	faz (ES)	cara	0.39
change(FR)	caer (ES)	cambia	0.46
stânga (RO)	stanco (IT)	destra	0.52
tânăr (RO)	tenero (IT)	giovane	0.41
inimă (RO)	anima (IT)	cuore	0.13
amic (RO)	amico (IT)	amichetto	0.04

Table 3: Extracted false friends and falseness.

3 Laws of Cross-lingual Semantic Divergence

We use the measure of falseness of a deceptive cognate pair to quantify the semantic shift between the meanings of a word derived from the same etymon in different languages. We further propose analyzing how the properties of frequency and polysemy of a word relate to semantic shift, and, analogously to what Hamilton et al. (2016) do for monolingual semantic change, we aim to move towards uncovering statistical laws of semantic change across languages.

We first define a measure of the frequency of a word, as well as a measure of its polysemy. Further, we try to correlate these measures of frequency and polysemy with the falseness measure defined in the previous sections. At this step, we discard all cognate pairs that, according to the false friend detection algorithm, are true cognates, and focus only on the deceptive cognates. On average across all language pairs, 37% of the cognate pairs in our dataset are found as deceptive cognates. Moreover, we validate these results using multilingual WordNet, and further select only pairs which are confirmed to be deceptive cognates as such: two cognates are considered to be true cognates if they are synonyms according to WordNet, and are considered to be deceptive cognates otherwise. It should be noted that having to use WordNet limits us to languages for which WordNet is available (excluding Romanian).

Although our approach is very similar to the one proposed by Hamilton et al. (2016), an important difference should be noted: while the authors of the monolingual study correlate the magnitude of the shift of meaning in a word to its frequency and polysemy *prior* to the change in meaning, our method looks at the properties of words *after* the meaning shift has already occurred, presumably from the original meaning of the proto-word they derive from to their current meanings in their respective languages.

	ES	PT	IT	FR	EN
ES	-	-23.4	-31.5	-39.8	-20.9
PT	-42.0	-	-37.7	-34.2	-31.4
IT	-29.5	-28.5	-	-33.9	-36.2
FR	-25.9	-16.3	-23.3	-	-31.9
EN	-27.7	-39.3	-39.7	-39.2	-

Table 4: Correlations of frequency with falseness.

3.1 Word Frequency and Semantic Divergence

For measuring **frequency**, we use the rankings of words based on their frequency in the corpus used to build the embeddings, which are readily available in the FastText embeddings that we use out of the box. The most frequent words will be associated with the lowest ranks. We normalize the absolute rank of a word dividing by the total number of words in its language, obtaining a relative rank ranging from 0 to 1 (with 0 corresponding to the most frequent words and 1 to the rarest).

For each pair of languages in a cognate set, we compute the Spearman correlation between the frequency rank of the first word in the cognate pair and the falseness of the deceptive cognate. Since frequency and polysemy are correlated, we need to control for polysemy in order to observe the marginal effect of frequency on semantic divergence. To this effect, we compute partial correlations, using polysemy as a covariate variable. Similarly, when computing correlations for polysemy, we set frequency as a covariate.

The results showing the correlations for each language pair are reported in Table 4. The values

	ES	PT	IT	FR	EN
ES	-	56.2	47.3	26.5	12.1
PT	20.2	-	34.5	28.8	4.2
IT	18.6	15.0	-	6.2	2.1
FR	14.2	26.0	16.4	-	-5.4
EN	-9.1	-11.2	-16.5	-14.0	-

Table 5: Correlations of polysemy with falseness.

are considerable for most language pairs, suggesting that the frequency of the word does play a role in the way its meaning shifts.

We also further try to understand the type of relationship between frequency and falseness. Following the results of Hamilton et al. (2016) showing that frequency relates to semantic shift according to a power law, we verify this in our setup by plotting the log of the frequency against the falseness degree, and then the log of polysemy against the falseness degree, confirming a similar type of relationship in our case, as shown for Spanish-Portuguese in Figure 2.

It is interesting to compare our results with those of Hamilton et al. (2016), where the authors observe an inverse correlation between frequency and meaning shift: the more frequent words tend to change their meaning more slowly. Our experiments show the opposite effect: even though the correlation values are negative, here we use frequency ranks rather than raw counts, so a negative correlation indicates a positive relation: more frequent words have diverged more in meaning. This may be related to the fact that we measure frequency *a posteriori*: the cognates we compared had *already* diverged in meaning before we measured their frequency, which may lead to a different effect than the one observed by Hamilton et al. (2016).

3.2 Word Polysemy and Semantic Divergence

For **polysemy**, we make use of WordNet, a semantic network organized in synsets which represent concepts - where each word is part of as many synsets as concepts it designates. In this way, the polysemy of a word can be defined as the number of synsets that it is part of in WordNet.

We perform similar experiments for polysemy, correlating the degree of polysemy of the first word in a cognate pair to the falseness of the pair. The results, shown in Table 5, are noteworthy for most language pairs here as well, though less pronounced than for frequency. Figure 2 shows the relationship between log-polysemy and falseness,

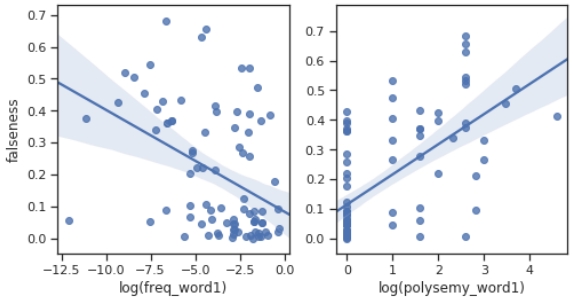

Figure 2: Falseness correlation with log-frequency and log-polysemy for Spanish-Portuguese.

which displays a clear linear trend. More than that, it is interesting to see that the correlations are higher for languages which are known to be more closely related: the strongest effects are observed for Spanish and Portuguese, which are the closest, geographically, of all Romance languages and may have evolved together for parts of their history. English, as the only non-Romance language, also stands out for showing the weakest effects of polysemy on falseness for most language pairs, and for some even shows an inversed effect of negative correlation with falseness with Romance languages.

For Romance languages, polysemy proves to be positively correlated with falseness, confirming the results on monolingual experiments in previous studies: more polysemantic words seem to suffer more semantic shift – or rather, in our case, words which have undergone more semantic shift tend to be more polysemantic.

4 Conclusions

We have proposed in this paper a new perspective for studying semantic change: comparing meaning of cognate words across languages.

We have shown how frequency and polysemy relate to semantic shifts of cognates across languages, demonstrating that both the frequency and polysemy of cognates positively correlate with their cross-lingual semantic shift, taking the first steps towards formulating statistical laws of cross-lingual semantic change. In the future, including the proto-word in the analysis where available (in this case, the Latin etymon) may give further insight into how cognates change their meaning. Additionally, it would be interesting to further explain these correlations, as well as study other hypothesized laws of semantic change in a multilingual setting.

References

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

Lyle Campbell. 1998. *Historical Linguistics. An Introduction*. MIT Press.

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*.

Alina Maria Ciobanu and Liviu P. Dinu. 2014. Building a Dataset of Multilingual Cognates for the Romanian Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1038–1043.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

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.

Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In *NetWordS*, pages 66–70.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1136–1145.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 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 the International Conference Recent Advances in Natural Language Processing*, volume 9, pages 251–257.

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

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 the International Conference RANLP-2009*, pages 292–298.

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. 2018. Survey of computational approaches to diachronic conceptual change. *arXiv preprint arXiv:1811.06278*.

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 Dinu. 2019. A computational approach to measuring the semantic divergence of cognates. In *Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing*. To be published.

Yang Xu and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *CogSci*.