Tracking Semantic Change in Cognate Sets for English and Romance Languages

Ana Sabina Uban[♠,♡,♢] **Alina Maria Cristea**[♡] **Anca Dinu**[♣,♡]
Liviu P. Dinu[♠,♡] **Simona Georgescu**[♣,♡] **Laurențiu Zoicaș**[♣,♡]

[♡]Human Languages Technologies Research Center, University of Bucharest
[♠] Faculty of Mathematics and Computer Science, University of Bucharest
[♣]Faculty of Foreign Languages and Literatures, University of Bucharest
[♢]PRHLT Research Center, Universitat Politècnica de València

ana.uban+acad@gmail.com, alina.cristea@fmi.unibuc.ro, anca.dinu@lls.unibuc.ro
ldinu@fmi.unibuc.ro, simona.georgescu@lls.unibuc.ro, laurentiu.zoicas@lls.unibuc.ro

Abstract

Semantic divergence in related languages is a key concern of historical linguistics. We cross-linguistically investigate the semantic divergence of cognate pairs in English and Romance languages, by means of word embeddings. To this end, we introduce a new curated dataset of cognates in all pairs of those languages. We describe the types of errors that occurred during the automated cognate identification process and manually correct them. Additionally, we label the English cognates according to their etymology, separating them into two groups: old borrowings and recent borrowings. On this curated dataset, we analyse word properties such as frequency and polysemy, and the distribution of similarity scores between cognate sets in different languages. We automatically identify different clusters of English cognates, setting a new direction of research in cognates, borrowings and possibly false friends analysis in related languages.

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, most of which anchored in the speakers' experiences, encyclopedic knowledge and cognitive mechanisms (Rousseau, 2000). Words are continuously changing, with new senses emerging all the time. Campbell (1998) presents 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 analyse the principles and statistical laws governing semantic change, such as the law of parallel change and the law of differentiation (Xu and Kemp, 2015), the law of conformity and the law of innovation (Hamilton et al., 2016), or the law of prototypicality (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) and Tahmasebi et al. (2018).

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

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 which also have a common meaning (e.g. Ro. *mână*, It. *mano*, Fr. *main*, Es. *mano*, Pt. *mão* 'hand'), and *deceptive cognates* or *false friends* to refer to cognate pairs which do not have the same meaning (anymore) (e.g. Ro. *pleca* 'to leave' / Fr. *plier* 'to fold' / Es. *llegar* 'to arrive', all of them originated from Lat. *plicare* 'to fold').

Most linguists found psychological and structural factors to be the main cause of semantic change (Meillet, 1906; Coseriu, 1958), but the evolution of technology and socio-cultural changes are not to be omitted. Moreover, when

Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021, pages 64–74

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-linguistic, cultural and historical factors (for details concerning the semantic fields most permeable to borrowing, in accordance with the socio-cultural circumstances, cf. Tadmor (2009)). The evolution of cognate words in different languages can be seen as a collection of different parallel histories of the proto-word from entering the new languages to its current state. Based on this view, we rely on a different framework 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.

A comprehensive list of cognates and false friends for every language pair is difficult to find or manually build – this is why applications often rely on automatically identifying them. Related to our task, there have been a number of previous studies attempting to automatically extract pairs of true cognates and false friends. 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 cognates, usually using simple methods on only one or two pairs of languages (Torres and Aluísio, 2011; Castro et al., 2018).

Uban et al. (2019a) propose a method for identifying and correcting false friends, as well as define a measure of their "falseness", using cross-lingual word embeddings and automatically extracted cognate sets (Uban et al., 2019b; Uban and Dinu, 2020; Uban et al., 2021). Expanding upon the direction proposed there, we create a new curated dataset of cognate sets in English and Romance languages. Additionally, we label the cognate sets according to their etymology and the period they entered the language, separating them into two distinct groups: *old borrowings* and *recent borrowings*. On this dataset, we investigate patterns related to the distribution of frequency, polysemy and cross-lingual semantic similarity across cognates, and show that the similarity distributions of English words show a specific bimodal pattern. We provide qualitative analyses and extensive linguistic interpretations for all our findings.

We bring several contributions to the computational study of semantic change and cognate words. To the best of our knowledge, we are the first to approach the problem of dating cognates based on their semantic content. Analysing the formal properties of cognates (i.e. their word form) is a method that is well-known in computational historical linguistics to gauge how language families have evolved (Ciobanu and Dinu, 2015). Computational approaches to analyse changes in meanings of cognate sets in order to investigate language contact settings have not been considered in historical computational linguistics research. Additionally, we publish a novel electronically readable dataset with high quality annotations regarding the period a word entered the English language, for a selection of cognates in English and Romance languages. To our knowledge, it is the first of its kind, and we hope it can help further research into computer-assisted analysis of cognate words.

1.1 Preliminaries

Cognates are words in sister languages (languages descending from a common ancestor) with a common proto-word. For example, the Spanish word *paz* and the French word *paix* are cognates, as they both descend from the Latin word *pacem* (N. *pax*, meaning *peace*) – see Figure 1.

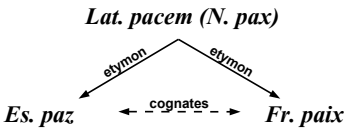


Figure 1: Example of cognates and their common ancestor: *peace*.

An important distinction is to be made between inherited words and borrowings: we speak of *inherited words* when referring to those lexemes that have been preserved from the ancestor language in the vernacular languages by uninterrupted oral usage, thus taking part in the process of language formation; by *borrowing* (also known as *loanword*), on the contrary, we understand any word that has been adopted in a language *A* from a language *B* after the language *A* has passed through its ba-

sic formation period (Reinheimer Ripeanu, 2001, 2004). According to Hall (1960), there is no such thing as a "pure language" – a language "without any borrowing from a foreign language". The process in which words enter one language from another is called *linguistic borrowing*. The average borrowing rate, reaching 24.2% (Tadmor, 2009), turns the borrowing process into one of the main resorts of lexical enrichment. The result of the borrowing process depends on numerous factors, such as the length and intensity of the contact and the extent to which the populations in question are bilingual (Campbell, 1998). Although admittedly regarded as relevant factors in the history of a language (McMahon et al., 2005), borrowings bias the genetic classification of the languages, characterizing them as being closer than they actually are (Minett and Wang, 2003). Thus, the need for discriminating between cognates and borrowings emerged (Ciobanu and Dinu, 2019). Heggarty (2012) acknowledged the necessity and difficulty of the task, emphasizing the role of the "computerized approaches" (Ciobanu and Dinu, 2015; Tsvetkov et al., 2015).

The concept of "Latin inherited word" can only be applied to the Romance languages, as these are the only languages whose ancestor is Latin. The descendants of the same Latin word in various (if not all) Romance languages are called "cognates" (ex. Ro. *drept* "right", It. *dritto*, Fr. *droit*, Es. *derecho*, Pt. *direito* are cognates, as they are all inherited from Lat. *directus*). On the other hand, the Romance languages have also experimented a period of "relatinization" (starting as early as the 13th century in Western Europe), when they massively borrowed words, through a cultural, written channel, from the same language from which they originate: in this case, Latin does not play the role of ancestor language any more, but it represents a non-contemporary source of lexical enrichment (Reinheimer Ripeanu, 2004). To give an example, the same Latin word *directus* has been borrowed in Ro. *direct* "direct", It. *diretto*, Fr. *direct*, Es. *directo*, Pt. *directo*, in a period that varies from the 13th century for French, to the 19th century for Romanian.

In order to maintain the distinction between the two possible channels (oral vs written) through which Latin words entered the Romance lexica (inherited word vs borrowing), and at the same time to highlight the genetic relation between the Romance lexemes in either case, we have adopted a twofold terminology: we shall use the concept of *real cognate* to refer to the relationship between inherited words that come from a common ancestor (Ro. *drept* "right", It. *dritto*, Fr. *droit*, Es. *derecho*, Pt. *direito*), and *virtual cognate* to denote the connection between words that have been borrowed from the same Latin etymon (Ro. *direct* "direct", It. *diretto*, Fr. *direct*, Es. *directo*, Pt. *directo*).

When it comes to English, we can only use the term "borrowing" whenever we refer to a word of Latin origin. Given that the accuracy of our dataset analysis involves a clear distinction between the two main historical stages when clusters of words of Latin origin were integrated in the English lexicon, we established an internal differentiation between *"old borrowings"* (that penetrated English through Old French, that means anytime before the first half of the 15th century) and *"recent borrowings"* (taken directly from Latin, from the second half of the 15th century to the present day).

It is easily understandable that the Latin lexical thesaurus has offered to the English language more or less the same lexical items that it disseminated in the Romance languages (either by inheritance or by cultural transmission). In this case, the English borrowing will equally be considered a *"virtual cognate"* of the Romance lexical items coming from the same Latin etymon, regardless if these are inherited or borrowed (e.g. En. *direct* vs Ro. *drept/direct*, It. *dritto/diretto*, Fr. *droit/direct*, Es. *derecho/directo*, Pt. *direito/directo*).

2 Cognates 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. The dictionary-based approach for identifying cognates, described in detail in (Ciobanu and Dinu, 2013), comprises two steps: firstly, the etymological information is extracted from electronic dictionaries; secondly, the etymologies are matched: words with the same language of origin and the same etymon are considered to be cognates. This approach answers the question raised by Swadesh (1954): "Given a small collection of likely-looking cognates, how can one

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

Table 1: Example of a cognate set: *architect*.

definitely determine whether they are really the residue of common origin and not the workings of pure chance or some other factor?", as the analysis is performed only on words that share a common etymology. We augment the dataset with the corresponding cognate in English (in the broad sense, since these are borrowings) for a subset of 305 of these cognate sets, using the same approach that was used for building the original dataset.[1] Considering a Romance cognate set and an English cognate candidate, both with Latin etymology, we compare their etymons. If they match, we identify the English word as being part of the cognate set. One complete example of a cognate set in Romance languages and English for the word *architect* is represented in Table 1.

We curate the obtained cognate sets and include high-quality annotations separating them into two groups according to their etymology (*old borrowings* and *recent borrowings*), provided by experts in linguistics. Out of the total 305 cognate pairs, we find 105 old borrowings and 135 recent borrowings (while the rest cannot be assigned a clear label or represent errors). We provide more details on data curation and evaluation in the following section.

2.1 Dataset Evaluation and Manual Curation

Our approach needs not be totally automated, nor completely manual, but rather computer-assisted.

The corpus was built by extracting the basic information from electronic dictionaries of Romance languages, as described in detail in (Ciobanu and Dinu, 2014), as well as the *Collins Dictionary*[2] for English, followed by a detailed curation of the lexical sets obtained, with the aid of the following dictionaries:

- for English: *Online Etymology Dictionary*[3]; *The Oxford Dictionary of English Etymology* (Onions et al., 1994); *Merriam-Webster*[4];
- for Romanian: *Dicționar Explicativ Român* (DEX[5]), *Dicționarul Etimologic al Limbii Române* (Ciorănescu, 2002)[6];
- for Italian: *Il Nuovo De Mauro*[7];
- for French: *Trésor de la Langue Française Informatisé*[8], *Dictionnaire historique de la langue française* (Rey, 2011); *Le Grand Robert* (CD);
- for Spanish: *Diccionario de Uso del Español* (Moliner, 2007); *Diccionario de la Lengua Española*[9];
- for Portuguese: *Dicionario Priberam*[10].

The annotations made by the expert linguists for the English cognates had to give account of the following data: on the one hand, the way they entered the English language (either as direct borrowings from Latin or via French), and, on the other hand, the period when they were first attested (before the first half of the 15th century or after). By using these two criteria, we could decide whether a cognate is an old or recent borrowing.

To evaluate our dataset, we consider a cognate set to be correct if all cognates in the set were correctly identified for each language. We evaluate not only the automatic extraction, but also the etymological information from the electronic dictionaries. We ought to mention that we classified as an error any type of distancing from the standard version we were expecting (e.g. a conjugated form of the verb instead of its infinitive, for instance *admits* instead of *admit*, or, when it comes to Romance languages, the feminine form of a noun or adjective instead of the standard masculine variant). Thus, the resulted overall accuracy was 53% (161 correctly identified cognate sets out of the 305 automatically extracted ones). The overall accuracy represents the percentage of cognate sets in which the comprising cognates are correct for

[1] The dataset size was reduced when including English mainly because of two reasons: 1) we did not identify etymologies for all English cognate candidates; 2) some cognate sets from the initial dataset might not have a corresponding cognate in English.

[2] https://www.collinsdictionary.com/

[3] https://etymonline.com/

[4] https://www.merriam-webster.com/

[5] https://dexonline.ro/

[6] cf. https://dexonline.ro/

[7] https://dizionario.internazionale.it/

[8] http://atilf.atilf.fr/

[9] https://dle.rae.es/

[10] https://dicionario.priberam.org/

all languages. In other words, if at least one cognate was incorrect, we considered the cognate set to be incorrectly identified. Per language, we obtained the following accuracy values: 70.8% (English), 82.6% (Spanish), 77.3% (French), 80.0% (Italian), 79.6% (Portuguese), 81.3% (Romanian). In this case, we computed accuracy individually for each language, without looking at the entire cognate sets. As expected, the accuracy values per language were higher than the accuracy per cognate sets. We have also computed the average normalized edit distance (Levenshtein, 1965) between the correct cognates and those extracted automatically, as a way to assess the degree of minor errors in word forms as opposed to entirely incorrect cognate associations, obtaining the following values: 0.20 (English), 0.14 (Spanish), 0.17 (French), 0.14 (Italian), 0.16 (Portuguese), 0.14 (Romanian). Thus, to obtain an accurate dataset, a second stage of manual curation and error removal was necessary. We observed several types of errors that generally occur due to the interference between similar forms, which the machine cannot discriminate, but also due to lack of information in the source of the data (dictionaries). Most of those errors consisted in a missing cognate or an incorrect one, while some were incorrect associations of words that had no common etymology. Particularly, a common type of error is the selection of different grammatical categories from one language to another (En. *cause* – that can be either verb or noun – is placed next to Es. *causar* – verb –, but It. *causa*, Ro. *cauza* – noun).

Another inaccuracy – not fully mistaken and at the same time very interesting from an etymological point of view – is the identification of an English lexical item that only has a distant etymological connection to the Romance words selected as its cognates: for instance, next to Es. *fuego* "fire", It. *fuoco*, Fr. *feu*, etc. – all inherited from Lat. *focus* –, the machine placed En. *fuel*, that, although not directly derived from Lat. *focus*, was borrowed from the Old French descendant of a derivative of *focus*, namely *focale* (Fr. *fouaille*). Another case of placing at the same level different strata of virtual cognates is that of En. *brave* (borrowed from It. or Es. *bravo*, at their turn inherited from Lat. *barbarus* "barbarian") that appeared next to the loanwords It. *barbaro* "barbarian", Es. *bárbaro*, etc. Intrinsically related to this inaccuracy was the lack of dating of the exact period when the words entered a language. As a particular observation, we found that the errors generated by the automatic processing sometimes coincide with the cases where speakers themselves misinterpret the origin of a word (a linguistic process known as "folk etymology" or "paretymology", i.e. the false connection between two similar words that etymologically have nothing in common, leading to a change of one of them either in form or in meaning (Schweickard, 2008)).

We report the results on the curated dataset, which we make available publicly[11].

3 Measuring Cognate Divergence

3.1 Methods

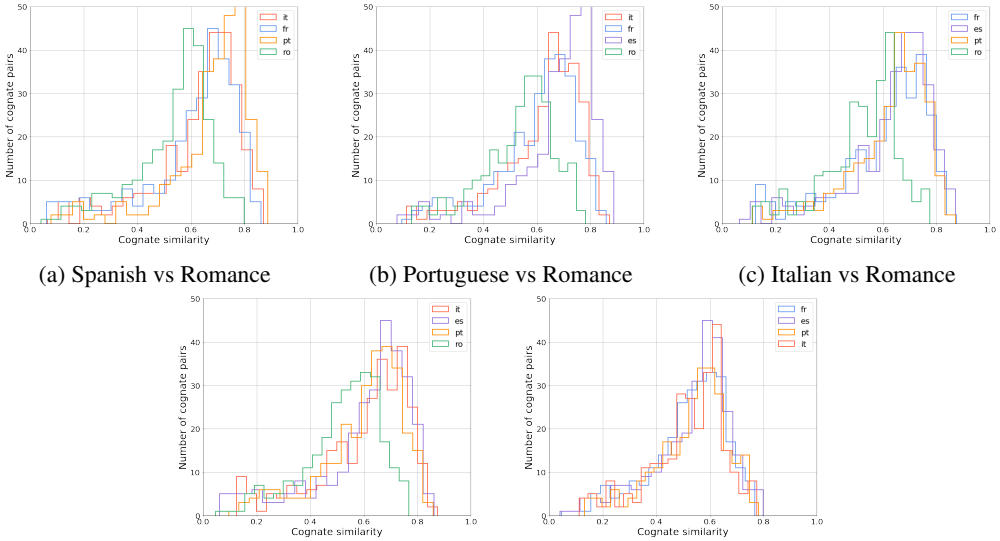
Word embeddings have become a standard method for measuring lexical semantic similarity in the field of computational analysis of semantic change.

In our study, we make use of word embeddings computed using the FastText algorithm, pretrained 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. In our cross-lingual setup, we make use of cross-lingual word embeddings in order to compute semantic similarities between words in different languages. Obtaining cross-lingual word embeddings entails training word embedding spaces for each language separately, then applying an alignment algorithm across the obtained vector spaces in order to create a common space.

This is accomplished through 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 (for which it is known that they have the same meaning), based on a small bilingual dictionary. For our purposes, we use the publicly available multilingual alignment matrices that were published by Smith et al. (2017). Finally, we compute semantic similarities for each pair of cognate words using the cosine similarity between their corresponding vectors in the shared embedding space.

We separately extract word frequency scores for all words in the dataset. For measuring fre-

[11] https://nlp.unibuc.ro/projects/cotohili.html



(a) Spanish vs Romance (b) Portuguese vs Romance (c) Italian vs Romance

(d) French vs Romance (e) Romanian vs Romance

Figure 2: Distributions of cross-lingual similarity scores between cognates.

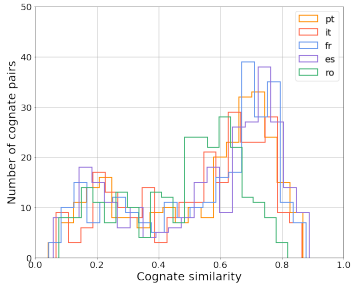


Figure 3: Distribution of similarities for automatically extracted English cognate sets according to the proposed algorithm.

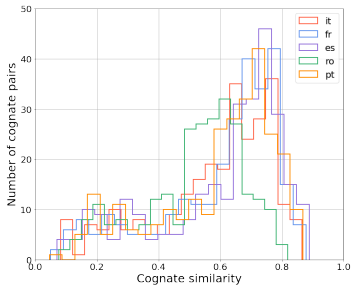


Figure 4: Distribution of similarities for curated English cognate sets according to the proposed algorithm.

quency, we use the multilingual Wordfreq Python library (Speer et al., 2018), which estimates word frequency based on multiple corpora (such as Wikipedia and Twitter). For most of the languages we consider, we are able to extract frequency scores for the majority of words in our cognate sets, with a coverage of at least 92% of the words in our cognate sets for every language considered, except for Romanian, which has a poorer coverage of only 60%. The library provides a log-normalized frequency score, ranging between 0 and 10 on a logarithmic scale, with higher scores corresponding to more frequent words.

We additionally measure word polysemy, making use of Open Multilingual WordNet (OMW) (Bond and Paik, 2012). In this way, the polysemy of a word can be defined as the number of synsets that it is part of in WordNet. We have to exclude Romanian from this analysis, since it is not supported in OMW.

Given these data, we performed several experiments to compare the two groups of English borrowings according to our annotations: comparing their frequencies, polysemy scores as well as their average similarity scores across languages. We report the obtained results in the following section.

3.2 Results

From the common vector space of the curated dataset, we obtained the cosine similarity score (between 0 and 1) for all pairs of cognates and for all pairs of languages. The distribution of these similarities is depicted in Figure 2, for each Romance language versus all other Romance languages. One notices that the distribution is unimodal, skewed to the right, with a mean similarity around 0.7. One possible explanation for the longer left tail is the inherent noise present in the relatively small dataset, which results in a bulk of less similar cognate sets.

An interesting case is the distribution of similarity between English and Romance languages cog-

nate pairs, which seems bimodal, indicating two groups: a low similarity group, with a mean of around 0.2 similarity score, for the left curve and a high similarity one, with a mean of around 0.7 similarity score, for the right curve. In Figure 2 and Figures 3 and 4 one can observe the difference between the distribution of English versus Romance languages on the automatically generated dataset and on the curated dataset, respectively. After eliminating the errors from the dataset, the curve for the low similarity group flattened, probably because many of the eliminated errors resulted in low similarities between pairs of cognates. Still, the two distinct groups for the English cognate similarities remain visible, which demands an explanation. Our hypothesis was that the low similarity group could represent the old English borrowings from Latin, while the high similarity group could represent the recent English borrowings. To test this hypothesis, we used the manual labels for the English cognate words as old or recent borrowings and used a Mann-Whitney U Test on the two sets, to check whether the means of the two groups are actually different, shown in Table 2. It turned out that the mean differences between the two groups are not statistically significant. It might be that the bimodality is a result of noise or chance, or that there is another explanation that we have missed.

We additionally tested other hypotheses related to the difference between the two groups of English cognates, and compared the average frequency and polysemy for the two groups, which showed some statistically significant patterns. We note that the distribution of word properties such as frequency and polysemy for cognate sets have been studied before (Uban et al., 2019b, 2021) on automatically extracted cognate sets: in our study, we perform the analysis based on the curated cognate sets (providing more reliable results), as well as analyse them in relation to the two groups of English borrowings according to our annotations.

	Recent Borrw.	Old Borrw.	p-val
FR	.63	.62	.35
ES	.64	.61	.47
IT	.61	.60	.45
PT	.62	.61	.21
RO	.53	.53	.44

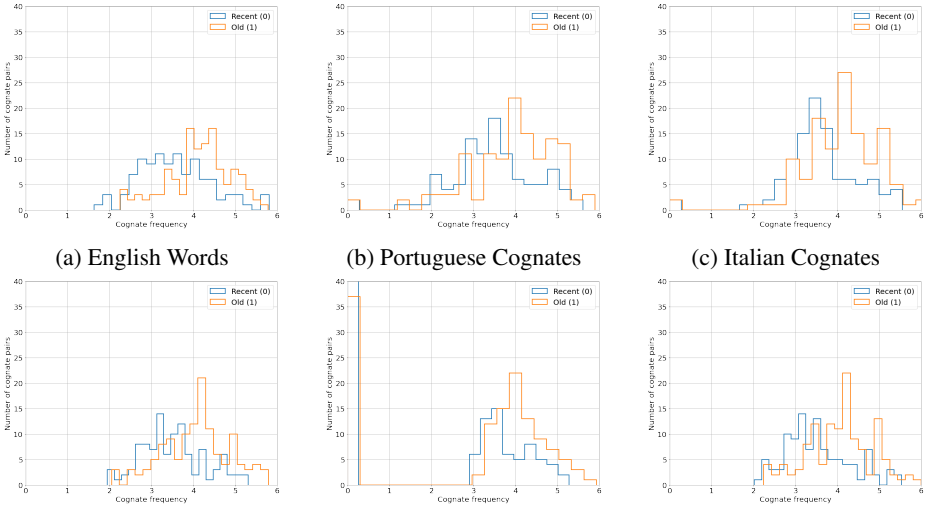
Table 2: Average cognate similarities for old and recent English borrowings.

In Table 3 we show the average log-frequencies of the two groups, as well as the statistical significance of their difference. The difference in frequency is statistically significant, with very low p-values. We can see the histograms representing the frequency distributions in Figure 5. Table 4 shows the average polysemy scores for the two groups, which show a similar pattern: old borrowings have higher polysemy than recent borrowings, and so do their cognates in Romance languages. We note here that given the known dependence between frequency and polysemy (frequent words tend to be more polysemous), more experimentation is needed to confirm whether the noticed effects with regards to frequency and polysemy still manifest independently.

	Recent Borrw.	Old Borrw.	p-val
EN	3.55	4.12	4.28e-08
FR	3.52	4.06	7.33e-08
ES	3.54	4.05	2.37e-07
IT	3.59	4.10	5.04e-07
PT	3.45	3.92	3.72e-05
RO	2.33	2.93	1.36e-03

Table 3: Average (log-)frequencies for old and recent English borrowings and their cognates.

Despite the lack of a statistically significant difference between the old and the recent borrowings, we can still extract various socio-historical features that may characterize each of the two groups. Thus, the first stratum of borrowings is usually represented by concepts of primary necessity in communication, adopted through direct contact between two contemporary languages (Franco-Norman / Old French – Old English), hence become part of the fundamental lexical core of the language (e.g. *eagle*, *anchor*, *peace*, etc.). On the other hand, the more recent Latin borrowings are adopted through a cultural channel, either as lexical units circumscribed to the acrolect – often as a mere consequence of the prestige of the source language – (e.g. *celestial*, *diurnal*, *aphorism*, etc.), or as specialized terms restricted to a particular professional domain (e.g. *diameter*, *apostasy*, *atrophy*, etc.). Although most of them may be included in the category of catachrestic borrowings (according to the differentiation between catachrestic and non-catachrestic borrowings drawn by Onysko and Winter-Froemel (2011)) – as they entered the language together with the concept they designate –, the great majority of these recent Latin borrowings did not reach the shared lexicon, as a consequence of their ab-



(a) English Words (b) Portuguese Cognates (c) Italian Cognates

(d) French Cognates (e) Romanian Cognates (f) Spanish Cognates

Figure 5: Distributions of (log-)frequencies for English old borrowings vs recent borrowings, and their corresponding cognates in the Romance languages.

sence in the average speaker discourse.

	Recent Borrw.	Old Borrw.	p-val
EN	3.43	5.42	6.89e-05
FR	4.38	7.04	0.002
ES	2.47	3.58	0.001
IT	2.07	3.35	7.49e-08
PT	3.16	4.00	0.02

Table 4: Average polysemy scores for old and recent English borrowings and their cognates.

4 Qualitative Analysis and Interpretation

While the initial hypothesis of an effect of the older versus more recent borrowings on semantic similarity was not supported by mean evidence, we tried to deepen our investigation by researching in detail a sample of cognate sets. We aimed to observe whether the fluctuation in the degree of similarity between the English virtual cognates, on the one hand, and their Romance correspondents, on the other hand, could be more related to the transmission channel through which they became part of the modern languages' lexica.

As we previously mentioned, the Latin borrowings in English can date from very different periods of time: some of them go back to the period of direct contact between Germanic and Latin speakers (e.g. *fork*), many of them are borrowed via Old French – thus having as a starting point in their semantic evolution the French meaning (e.g. *camp*) –, while a more recent cluster consists of loanwords taken directly from Latin (e.g. *precocious*). We shall detail the particularities of each category by highlighting the degree of semantic similarity and by interpreting the causes of either divergence or closeness.

The example of En. *fork* is illustrative for the semantic divergence that affected the relationship between an early loanword in Celtic taken directly from Latin and its Romance virtual cognates. Lat. *furca* designated "an instrument with two arms or prongs", as well as any "Y -shaped piece of wood used as a support", including the "gallows". The distribution of meanings varied from one area of the Roman Empire to another, according to the prevalent socio-cultural domain in which a *furca* was used: the Romanian descendent of *furca* designates the instrument used in agriculture, in Spanish and Portuguese it was specialized as an instrument used for punishment, "gallows", while in English it semantically evolved to designate a refined instrument for eating. The semantic similarity between the English word and its Romance correspondents is thus very low (between 0.1 and 0.4 [French]), as the cognates reflect different semantic trajectories based on concrete socio-cultural realities.

For the second category (lexical items inherited in Old French, that were later on borrowed in English), we shall approach the case of En. *powder* "fine, dry particles produced by the grinding, crushing, or disintegration of a solid substance", borrowed from O. Fr. *poldre* "finely ground and pounded substance" (registered with this mean-

ing as early as the 12th century), inherited from Lat. *puluerem* "dust". Contrastingly, the other Romance languages inherited the original meaning of "dust" as their main significate (Ro. *pulbere*, It. *polvere*, Es. *polvo*, Pt. *pólvora*). That explains why the degree of similarity between English and French is higher than between English and the other Romance languages (0.75 vs 0.5). Another significant example would be that of En. *camp* "a place with temporary accommodation of huts, tents, or other structures, typically used by soldiers, refugees, or travelling people", highly divergent from its Romance virtual cognates (Ro. *câmp*, It. *campo*, Fr. *champ*, Es. *campo*, Pt. *campo*, all of them real (and true) cognates sharing the meaning of "field"). In this case, the English word is a borrowing from Fr. *camp*, in its turn borrowed from Italian, that doubled the inherited form *champ*. As it was borrowed in French as a military term – in contrast to its virtual cognates specialized in the agricultural area – it continued the same line once it penetrated the English lexicon. The degree of similarity between En. *camp* and its Romance virtual cognates is, thus, as low as 0.1 (or even lower for Portuguese).

The dataset we obtained also allows us to draw specific conclusions concerning the semantic fields where the degree of similarity is higher, regardless of the difference between real and virtual cognates, as well as of the channel through which they penetrated in English. Thus, we may observe that the terms denoting concrete or at least experimentable elements, be they animals (e.g. En. *eagle*, Ro. *acvilă*, It. *aquila*, etc.), specific materials (En. *marble*, Ro. *marmura*, It. *marmo*, etc.), or seasons (En. *autumn*, Ro. *toamnă*, etc.), show a very high degree of similarity (with the average value of 0.75), as a consequence either of their frequency (as postulated by the *law of conformity*, cf. (Hamilton et al., 2016)), either of the lack of change in the referent or in the speakers' attitude towards the referent. Equally similar from a semantic point of view are the abstract terms that designate a very particular concept which could either be circumscribed to a restricted (scientific) domain (e.g. *astronomy*, *industry*, *diameter*, *identity*, *liquid*, etc.), or did not experience any polysemic developments (thus complying with the *law of innovation*, cf. (Hamilton et al., 2016)) (e.g. *avarice*, *circumstance*, *convince*, *irony*, *presence*, etc.).

We should also draw attention to the words that were borrowed from Latin in order to cover modern concepts, absent from the source culture. It is the case of En. *consul* "an official appointed by a government to reside in a foreign country to represent the commercial interests of citizens of the appointing country", Ro. *consul*, It. *console*, etc., which show one of the highest degrees of similarity: although the word in itself existed in Classical Latin, it referred to a different political position, designating "one of the two highest magistrates at Rome".

Parallelly, it is easily understandable why words modernly created in a determined scientific domain from Latin roots have almost no semantic divergence from one language to another: once created in a contemporary language, they were naturally spread in the other languages, along with the concept newly invented. It is the case of En. *nihilism*, *optimism*, *exhaustive*, etc.

5 Conclusions

We constructed a common vector space for English and Romance languages cognate sets to analyse their similarity and thus track their semantic divergence. We analysed their similarity distribution and proposed some linguistic and historical hypotheses to explain their behaviour, especially for English cognates.

An important byproduct of our work is the curated dataset, which can be employed in other work related to semantic analysis of cognates, borrowings or false friends.

We plan to extend this study, as part of future work, to cognate similarity based on phonetic transcription and compare it to the current orthographic dataset. Moreover, we will investigate more in-depth the automatic identification of the date a word entered a language. To this end, we need to obtain a dataset that contains this information. We intend to use and adapt (Dinu, 1996) to approximate the "age" of words.

Acknowledgements

We would like to thank the reviewers for their comments. All authors contributed equally to this work. This research is supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI – UEFISCDI, *CoToHiLi* project, project number 108, within PNCDI III.

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 Kyonghee Paik. 2012. A Survey of WordNets and their Licenses. *Small*, 8(4):5.

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 *LREC 2014*, pages 1038–1043.

Alina Maria Ciobanu and Liviu P. Dinu. 2015. Automatic Discrimination between Cognates and Borrowings. In *Proceedings of ACL 2015, Volume 2: Short Papers*, pages 431–437.

Alina Maria Ciobanu and Liviu P. Dinu. 2019. Automatic Identification and Production of Related Words for Historical Linguistics. *Computational Linguistics*, 45(4):667–704.

Alina Maria Ciobanu and Liviu Petrisor Dinu. 2013. A Dictionary-Based Approach for Evaluating Orthographic Methods in Cognates Identification. In *Proceedings of RANLP 2013*, pages 141–147.

Alexandru Ciorănescu. 2002. *Dicţionarul etimologic al limbii române*. Saeculum, Bucharest.

Eugenio Coseriu. 1958. *Sincronía, diacronía e historia*.

Mihai Dinu. 1996. *Personalitatea Limbii Romane*. Cartea Romaneasca.

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 EMNLP 2017*, pages 1136–1145.

Robert Anderson Hall. 1960. *Linguistics and Your Language*. Doubleday, New York.

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.

Paul Heggarty. 2012. Beyond Lexicostatistics: How to Get More out of "Word List" Comparisons. In *Quantitative Approaches to Linguistic Diversity: Commemorating the Centenary of the Birth of Morris Swadesh*, pages 113–137. Benjamins.

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.

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

Vladimir I. Levenshtein. 1965. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10:707–710.

April McMahon, Paul Heggarty, Robert McMahon, and Natalia Slaska. 2005. Swadesh Sublists and the Benefits of Borrowing: an Andean Case Study. *Transactions of the Philological Society*, 103(2):147–170.

Antoine Meillet. 1906. Comment les Mots Changent de Sens. In *Linguistique historique et linguistique générale*. Champion, Paris.

James W. Minett and William S.-Y. Wang. 2003. On Detecting Borrowing: Distance-based and Character-based Approaches. *Diachronica*, 20(2):289–331.

María Moliner. 2007. *Diccionario de Uso del Español*. Gredos, Madrid.

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.

Charles Talbut Onions, George Washington Salisbury Friedrichsen, and Robert William Burchfield. 1994. *The Oxford Dictionary of English Etymology*. Clarendon Press, Oxford.

Alexander Onysko and Esme Winter-Froemel. 2011. Necessary loans–luxury loans? Exploring the pragmatic dimension of borrowing. *Journal of pragmatics*, 43(6):1550–1567.

Sanda Reinheimer Ripeanu. 2001. *Lingvistica Romanica: Lexic, Morfologie, Fonetica.* . BIC ALL, Bucuresti.

Sanda Reinheimer Ripeanu. 2004. *Les emprunts latins dans les langues romanes.* Editura Universităţii din Bucureşti.

Alain Rey. 2011. *Dictionnaire historique de la langue française*. Le Robert, Paris.

André Rousseau. 2000. L'évolution lexico-sémantique: explications traditionnelles et propositions nouvelles. *Théories contemporaines du changement sémantique*, pages 11–30.

Wolfgang Schweickard. 2008. Le sirene degli etimologi nel mare onomastico. le reinterpretazioni paretimologiche. *D'Achille, Paolo/Caffarelli, Enzo (edd.), Lessicografia e onomastica*, 2:83–95.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*.

Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. LuminosoInsight/wordfreq: v2.2.

Adam St Arnaud, David Beck, and Grzegorz Kondrak. 2017. Identifying cognate sets across dictionaries of related languages. In *Proceedings of EMNLP 2017*, pages 2519–2528.

Morris Swadesh. 1954. Perspectives and Problems of Amerindian Comparative Linguistics. *WORD*, 10(2-3):306–332.

Uri Tadmor. 2009. Loanwords in the world's languages: Findings and results. In Martin Haspelmath and Uri Tadmor, editors, *Loanwords in the World's Languages*, chapter 3, pages 55–75. De Gruyter Mouton, Berlin.

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*.

Yulia Tsvetkov, Waleed Ammar, and Chris Dyer. 2015. Constraint-Based Models of Lexical Borrowing. In *Proceedings of NAACL-HLT 2015*, pages 598–608.

Ana Sabina Uban, Alina Ciobanu, and Liviu Dinu. 2019a. A Computational Approach to Measuring the Semantic Divergence of Cognates. In *Proceedings of CICLing 2019*. In press.

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 (In press)*. Berlin: Language Science Press.

Ana Sabina Uban, Alina Maria Ciobanu, and Liviu P Dinu. 2019b. 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.

Yang Xu and Charles Kemp. 2015. A Computational Evaluation of Two Laws of Semantic Change. In *CogSci*.