Caveats of Measuring Semantic Change of Cognates and Borrowings using Multilingual Word Embeddings

Clémentine Fourrier **Syrielle Montariol**
Inria
firstname.lastname@inria.fr

Abstract

Cognates and borrowings carry different aspects of etymological evolution. In this work, we study semantic change of such items using multilingual word embeddings, both static and contextualised. We underline caveats identified while building and evaluating these embeddings. We release both said embeddings and a newly-built historical words lexicon, containing typed relations between words of varied Romance languages.

1 Introduction

Languages are in constant evolution over time; words appear, disappear, and their syntactic form and semantic function evolve (Blank and Koch, 1999). However, languages evolutions can be closely inter-related, following phenomena of interactions and inheritance. Cognates and borrowings, which are the targets of our study, are direct consequences of these phenomena. **Cognates** are words which descend from the same ancestor word (their proto-form) belonging to a shared common direct parent language. For example, the French word *chat* 'cat' is cognate with Spanish *gatto* and Romanian *cătușă*, as they all descend from Latin *cattus* 'cat', a direct ancestor of these three languages. When a word is an evolution of a form which does not come from a direct ancestor, it is called a **borrowing**. English *cat* also comes from *cattus*,[1] but as Latin is not a direct ancestor of English, it is therefore a borrowing of English to Latin. We consider the relation between *cat* and *chat* to be of 'borrowing' type by extension. Borrowings mostly occur to designate 'realities that were unknown before the adopting speech community got in contact with the "giving" culture and its language' or to replace already existing meanings by the word of the related dominant culture (Krefeld, 2013). To study semantic variation, we look at our words' **glosses**, which are expressions of their meaning, here as their English translations or definitions. In our previous example, while the French and Spanish cognates both retained the original sense 'cat', the Romanian cognate went through a semantic change and is translated as 'handcuff'.

Semantic change studies historically relied on specific word relations, cognates and 'borrowings' (Durkin, 2015), found through the comparative method (formalised by Osthoff and Brugmann (1878)). The last few years have seen the emergence of new tools such as contextualised embeddings to study semantic variation (Martinc et al., 2020), enabling the comparison of word senses across domains, periods and languages. We join both approaches and expand on the work of Uban et al. (2021), who use 'static' (non-contextualised) embeddings to study semantics of cognates and borrowings in contemporary Romance languages and English. In this work, we use static as well as contextualised embedding to study the semantic evolution of cognates and borrowings, for both contemporary and older Romance languages, as well as English. To this end, we first create a dataset of cognates and borrowings from the widely studied Romance family (contemporary: Spanish, French, Italian, Portuguese, Romanian, old: Latin, Old Spanish, Middle French), to which we add English.[2] Then, we compare several methods to tackle the issue of obtaining, for low-resource historical languages, embeddings spaces aligned with the ones of contemporary languages. Both dataset and embeddings are released with the paper.[3] Lastly, we use these embeddings to study semantic shift for both diachronic (between parent and child) and synchronic (between children) cognates or borrowing

[1] Latin *cattus* is, that we know of, the most plausible origin of the proto-Germanic reconstructed word **kattuz*, ancestor of English *cat*

[2] The language codes are the following: Spanish (ES), French (FR), Italian (IT), Portuguese (PT), Romanian (RO), Latin (LA), Old Spanish (OSP), Middle French (FRM), English (EN).

[3] github.com/clefourrier/historical-semantic-change

Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, pages 97 - 112

relations, and find that contextualised embeddings allow us to reach more accurate conclusions. At each step, we highlight the possible pitfalls.

2 Related works

Cognates and borrowings transcribe different aspects of their languages history, and are often studied through the lens of orthographic (Ciobanu and Dinu, 2015, 2019) or phonetic combined with semantic variation (Kondrak, 2001). Uban et al. (2021), which we extend, study semantic variation in modern Romance languages between cognates and borrowings by considering their modern-day embeddings as a 'snapshot in time' of their meaning. As their dataset is not available, we can not use as a benchmark; however, like several public etymological databases, among which CogNet (Batsuren et al., 2019), containing cognates and borrowings without differentiating between both relation types, and EtymDB2 (Fourrier and Sagot, 2020), too small for our needs but which differentiate between both types, we build a dataset using the Wiktionary[4] as etymological source.

Semantic change across languages is actively researched in the linguistic and sociology research communities (Boberg, 2012), as it offers valuable information for sociological and historical analysis. In the NLP domain, many authors apply diachronic embeddings models to more than one language (Hamilton et al., 2016; Schlechtweg et al., 2020), but without considering their interactions. Some work studies variations between languages or dialects, diachronically (Martinc et al., 2020; Montariol and Allauzen, 2021) or synchronically (Hovy and Purschke, 2018; Beinborn and Choenni, 2020). However, although several annotated datasets are available to evaluate diachronic semantic change detection methods (Schlechtweg et al., 2020), cross-lingual semantic change does not have such resource and cognates and borrowings seem like a promising proxy for evaluating these methods.

3 Datasets and Corpora Construction

We create a dataset of cognates and borrowings in all languages under study. To complement it, we need corpora in each language to train or extract word embeddings; such corpora are publicly available for highly studied languages. We use a sample of the OSCAR corpus (Ortiz Suarez et al., 2019; Abadji et al., 2021) for contemporary languages and Latin. For Middle French and Old Spanish, we use less well-known resources.

3.1 Reference dataset construction

From the latest version of the Wiktionary, our goal is to construct a simple relational set of triplets (lang, lexeme, gloss) to other triplets for cognates and borrowings.

Parsing and general information extraction (for lexeme, language, relations) is described in Appendix A.[5] As extracting glosses proved less straightforward, we detail it here. We encountered three types of problem. 1) In the Wiktionary, some words have English translations as glosses, while others have English definitions: for example, the first definition of 'eau' (water) is 'Water, a liquid that is transparent, colorless, odorless and tasteless in its pure form, the primary constituent of lakes, rivers, seas and oceans', while for 'fort' (strong) it is 'strong; powerful' and 'skilled, proficient, successful, ...', a translation. Splitting glosses on punctuation to store the different semantic aspects as words is therefore indispensable in translation cases, but introduces mistakes when definitions are present. These cases were manually checked, but some mistakes might still remain. 2) All English words are defined (which makes sense, as the Wiktionary technically is an English multilingual dictionary). In order to have an homogeneous base, and as we try to keep translations only, we therefore make the choice to use English lexemes as their own 'translation' to English. 3) Some words (especially in Latin) are only defined as inflections or derivations of other words (e.g. capitum, only defined as 'genitive plural of caput'). In those cases, the gloss is not retained. After cleaning (also detailed in App. A), we construct our database, looking only at inheritance relations (App. A.2). Though cognate-typed relations exist in the Wiktionary, we deliberately choose to ignore them, as they can induce noise for our task: to define cognacy, we stood so far on the side of historical linguistics, but the term can sometimes more broadly refer to words with shared form and meaning, regardless of etymology (Frunza and Inkpen, 2009). This underlines the attention to sources which needs to be paid when constructing one's own database.

Statistics by language are detailed in App. A.4,

[4] The Wiktionary is a user-built free multilingual dictionary, found at en.wiktionary.org

[5] github.com/clefourrier/historical-semantic-change

Table 3. The cognate set contains a total of 34,574 word pairs, linking 8,334 unique words from all languages except English, which only has cognates to itself, as it does not descend from Latin and therefore cannot have cognates with any of the Romance languages. The borrowing set contains a total of 5,042 word pairs, linking 2,925 unique words. Here, most relations include English, with less than 100 pairs in relations without English.

3.2 Historical languages datasets

For **Middle French** (FRM, 1340–1610), we collect data from several datasets (see App. C.1): LEM17, a linguistically annotated corpus of modern French; MCVF 1.0/2.0 and PPCHF 1.0, parsed historical French data; OpenMedFr, plain versions of Middle French texts; and BFM2019, annotated Middle French texts. We manually filter these datasets to select all texts in the correct time period and clean them (see App. C.2).

For **Old Spanish** (OSP, 10th to 15th century), we extract data from the Digital Library of Old Spanish Texts[6], then clean it using the transcription norms described on the website.

After preprocessing, we obtain FRM/OSP datasets of 3.1M/4.7M words respectively.

4 Cross-lingual embeddings

We compare the semantic function of words in cognates and borrowings pairs. To this end, we explore various ways of obtaining aligned word embeddings in all languages (multilingual embeddings), using static and contextualised embeddings. The former are trained using FastText (Bojanowski et al., 2016) and aligned a posteriori, while the latter are extracted using the multilingual language model mBERT (Devlin et al., 2019) from corpora in all the languages under study. Trained embeddings and language models can be found for all our contemporary languages. However, historical languages such as Middle French and Old Spanish suffer from a scarcity of resources that we have to address.

4.1 Static embeddings

Available FastText embeddings. They were trained on Wikipedia data and either already cross-lingually aligned for our contemporary languages (Bojanowski et al., 2016), or available unaligned for Latin (Grave et al., 2018).

Training FastText embeddings. OSP and FRM do not have available embeddings: we therefore train some, using default subword tokenisation and an embedding size of 300.[7] However, we expect the quality of these new embeddings to be lower than the pre-trained ones, as 1) the imposed embedding size is likely too big with respect to the training data size, which could affect embedding ability to store relevant information, and 2) we were not able to define an adapted preprocessing.[8]

Aligning all embeddings spaces. Alignment is needed to obtain a coherent representation space between languages, and can be done either in a supervised or unsupervised way (Lample et al., 2017; Conneau et al., 2017). Preliminary experiments of unsupervised alignment (Alaux et al., 2018) led to extremely poor results. Consequently, we use bilingual lexicons[9] to supervise the alignment of Latin embeddings with Spanish, with around 2k bilingual word pairs used for supervision. Having no such dictionary for OSP/FRM, we use transparent words with their closest language (respectively SP/FR) to perform a supervised alignment, extracting for each language a bilingual lexicon of around 8k transparent words.

Extracting embeddings To build word embeddings, we had to manage un-homogeneous data with respect to diacritic: many cognates and borrowings seem absent from the embeddings vocabulary, especially for languages with diacritics (FR, RO, ES) or spelling variations (FRM) not homogenised in the embedding training corpora. We define a set of rules to extract embeddings despite word form variations. To embed word glosses (when made up of several words/sentences), we remove stopwords and compute the mean of all sequence word embeddings.

4.2 Contextualised embeddings

For contemporary languages and Latin, we use a sample of the OSCAR corpus (Ortiz Suarez et al., 2019; Abadji et al., 2021) to build our contextualised embeddings, as, given its very large

[6] http://www.hispanicseminary.org/t&c/nar/index-en.htm

[7] We use the same parameters as the pre-trained embeddings, to be able to align them together.

[8] OSP and FRM are too different from their descendants (e.g strong spelling variations inside FRM) to just use their languages preprocessing as such, and very few resource exist for these languages (e.g lists of stopwords).

[9] github.com/clefourrier/CopperMT/blob/master/inputs/raw_data/romance_bilingual

size (e.g. for Spanish, more than 25 billion tokens), working on the whole corpus would be time-intensive. For the other languages, we use the corpora descibed in Section 3.2.

We use an mBERT [10] model trained on 104 languages, including Latin and all our contemporary languages, from the `transformers` library (Wolf et al., 2020). Its training on Wikipedia data, allows for fairer comparison with FastText embeddings.

Massive multilingual pre-trained language models have been shown to perform well on new languages in a zero-shot fashion (Muller et al., 2021), especially those closely related to already seen high resource languages. Thus, we expect mBERT to perform well on OSP and FRM, but we also compare fine-tuning it on our FRM and OSP corpora using the masked language modelling task. We study cognates and borrowings representations *in context*, by computing the average embedding across all target word occurrences in corpora of their respective languages (Martinc et al., 2020). We compute word embeddings as the sum of the last 4 encoder layers of the model. When a word is divided into sub-words, we take the average of the sub-word embeddings

For word gloss embeddings (that we see as a representation of meaning), as we often have several words or a sentence as definition, we can directly generate their embeddings without contextualisation in a corpus. When the gloss is composed of several words, we try both averaging the representations of all tokens in the gloss, and using the embedding of the CLS representation. To compare them, we compute the cosine similarity between the target word embedding and the embedding of its associated gloss. Taking the CLS embedding leads to a similarity of 0.61 on average, while the average of all token embeddings leads to 0.67; we choose the latter to represent word meanings.

5 Results

Our metric is cosine similarity, commonly used in semantic change detection (Kutuzov et al., 2018). Our results are summarised in Table 1 (full results in App. B). Language pairs are split into parent to child (with LA, FRM, or OSP), and child to child (between contemporary languages) relations. We also differentiate cognates and borrowing pairs whose meaning stayed the same (un-shifted, equal gloss between the two items) or changed between the two languages (shifted, different gloss between the two items).[11]

We display similarity (across all our languages) between cognates / borrowings and their counterparts in an un-shifted (line 1) or shifted (l. 2) pair. We also display the average difference between these two scores (l. 3), this time computed per language pair: we expect it to be a measure of the models ability to capture semantic shift. The last two lines show the average embedding similarity between an item and its meaning,[12] which should be constant on average for a given language pair, since it reflects embedding alignment distance between the languages of interest and English.[13]

Embedding space quality. For **FastText** , the average similarity between item and meaning (l. 4 and 5) varies considerably from one language pair to another, which indicates variation in embedding alignment quality between English and other languages. This score also varies inside a given language pair (between borrowing/cognates or shifted/un-shifted words), which could further indicate embedding space quality problems. Indeed, an item embeddings and the embedding of its English gloss should always be relatively similar when using properly aligned embeddings spaces. We also observe that, contrary to expectations, publicly available pre-aligned embeddings (child-to-child) often have even higher variance and lower item-meaning similarity (therefore a worst alignment to English) than our aligned low-resource historical embeddings (parent-to-child). On the other hand, for **mBERT** embeddings, this similarity score is constant (with a slight variation between cognates and borrowings, likely explained by the fact that language pairs distribution between cognates and borrowings is different), which reflects a high embedding alignment quality. One should therefore be wary of conclusions drawn from the aligned FastText embeddings, even publicly available pre-aligned ones, which might lead to incorrect assumptions by introducing hidden factors into play. We will therefore draw conclusions only us-

[10] `bert-base-multilingual-cased`

[11] Semantic change would normally be seen as more of a continuum than a binary, but this was the more feasible approach with respect to our data.

[12] The item is the word form, where its meaning is the word English gloss.

[13] Note that even though we use definition embeddings, they should be comparable with word embeddings (Bosc and Vincent, 2018).

Relation		FastText				mBERT			
		Parent to child		Child to child		Parent to child		Child to child	
		cog	bor	cog	bor	cog	bor	cog	bor
item(a) ↔ item(b)	un-shifted	50±20	38±18	14±18	1±10	84± 9	86± 9	86±10	82± 9
	shifted	35±20	14±16	21±17	1± 9	79± 9	77± 9	79±10	76± 8
	difference	16± 7	16± 8	3± 6	-1± 3	4± 3	4± 7	2± 2	5± 4
item ↔ gloss	un-shifted	35±17	67±34	22±22	47±49	67± 5	72± 6	69± 6	71± 6
	shifted	29±24	62±40	16±16	49±48	69± 5	72± 6	69± 6	72± 6

Table 1: Aggregated results of cosine similarity (%) and standard deviation, for both FastText and mBERT embeddings. cog stands for cognate, bor for borrowing.

	FR-ES		FR-IT	
	cog	bor	cog	bor
% for un-shifted	84±8	92±4	84±8	87±5
% for shifted	79±9	84±8	80±9	83±8
#items	1884	22	1740	36

Table 2: item(a) ↔ item(b) mBERT similarity (%).

ing mBERT.

We also compared vanilla and fine-tuned OSP and FRM mBERT embeddings (Tables 8 and 9 in App. B); fine-tuning shows no significant improvement, though for some edge cases, it seems to increase semantic shift sensitivity slightly while decreasing similarity with other embedding spaces; consequently, we keep the simplest approach, the vanilla mBERT model. When working on historical data, it is interesting to study whether fine-tuning results justify its cost, or if zero-shot transfer can directly provide good enough results.

Global comparison Using mBERT embeddings, the only difference in similarity scores for items occurs between un-shifted and shifted word embeddings, with un-shifted pairs similarity being on average 4 points higher than shifted pairs (not necessarily statistically significant). Some outliers cases can be found in the per-language tables (see Table 6 in App. B), where shifted cognates have higher intra-pair similarity compared to un-shifted cognates for the same language pair. However, this situation only happens for languages with less than 20 cognates examples of shifted or un-shifted pairs (e.g. OSP-FRM, 12 shifted cognates), and are likely not significant.

There is virtually no difference between cognates or borrowings embeddings similarity. As a side note, FastText embeddings would have shown that cognates are more similar than borrowings, and a word is more similar to its parent than to its siblings: a hasty analysis using bad quality embeddings could have lead us to draw seductive but erroneous conclusions from the FastText embeddings.

Focus In order to investigate differences at the language pair level for the mBERT embeddings, we focus on two language pairs which have at least 20 samples for both shifted and un-shifted pairs of cognates and of borrowings: FR-ES and FR-IT (Table 2).[14] Both present a trend where borrowings are more similar than cognates and un-shifted words more similar than shifted words (as expected).

6 Conclusion

In this work, we create a cognate and borrowing dataset for English and Romance languages from different periods, as well as two aligned embeddings sets for all languages. When assessing embedding quality and alignment, we show that FastText embeddings, even when already pre-trained and aligned, are poorer than the mBERT ones on all respects. We therefore use the latter to study semantic change between cognates and borrowings: as expected, un-shifted word pairs are on average more similar than shifted ones. Furthermore, we observe a trend between cognates and borrowings, the latter being seemingly more similar than the former. Further analysis would be needed to determine whether this difference can be confirmed, by looking at chosen cognate and borrowings of similar histories in more languages. In summary, properly designed embeddings can be used to support historical lexicographic studies, while well-understood phenomena underlying cognates and borrowings can help design and evaluate cross-lingual word embeddings.

[14] There is a difference in data size of two orders of magnitude between small borrowing sets and bigger cognate sets, therefore conclusions must be taken with a pinch of salt.

References

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event), pages 1 – 9, Mannheim. Leibniz-Institut für Deutsche Sprache.

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2018. Unsupervised hyperalignment for multilingual word embeddings. *CoRR*, abs/1811.01124.

Khuyagbaatar Batsuren, Gabor Bella, and Fausto Giunchiglia. 2019. CogNet: A large-scale cognate database. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145, Florence, Italy. Association for Computational Linguistics.

Lisa Beinborn and Rochelle Choenni. 2020. Semantic drift in multilingual representations. *Computational Linguistics*, 46(3):571–603.

Andreas Blank and Peter Koch. 1999. Why do new meanings occur? a cognitive typology of the motivations for lexical semantic change. In *Historical Semantics and Cognition*, page 61–89.

Charles Boberg. 2012. English as a minority language in quebec. *World Englishes*, 31.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Tom Bosc and Pascal Vincent. 2018. Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1532, Brussels, Belgium. Association for Computational Linguistics.

Alina Maria Ciobanu and Liviu P. Dinu. 2015. Automatic discrimination between cognates and borrowings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 431–437, Beijing, China. Association for Computational Linguistics.

Alina Maria Ciobanu and Liviu P. Dinu. 2019. Automatic identification and production of related words for historical linguistics. *Computational Linguistics*, 45(4):667–704.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Philip Durkin. 2015. Etymology.

Clémentine Fourrier and Benoît Sagot. 2020. Methodological aspects of developing and managing an etymological lexical resource: Introducing EtymDB-2.0. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3207–3216, Marseille, France. European Language Resources Association.

Oana Frunza and Diana Inkpen. 2009. Identification and disambiguation of cognates, false friends, and partial cognates using machine learning techniques. *International Journal of Linguistics*, 1(1):1–37.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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 (Volume 1: Long Papers)*, pages 1489–1501. Association for Computational Linguistics.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

Grzegorz Kondrak. 2001. Identifying cognates by phonetic and semantic similarity. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Thomas Krefeld. 2013. Cognitive ease and lexical borrowing: the recategorization of body parts in romance. In *Cognitive ease and lexical borrowing: the recategorization of body parts in Romance*, pages 259–278. De Gruyter Mouton.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New

Mexico, USA. Association for Computational Linguistics.

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

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Syrielle Montariol and Alexandre Allauzen. 2021. Measure and evaluation of semantic divergence across two languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1247–1258, Online. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Pedro Javier Ortiz Suarez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

Hermann Osthoff and Karl Brugmann. 1878. *Morphologische Untersuchungen auf dem Gebiete der indogermanischen Sprachen*. Hirzel, Leipzig, Germany.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Ana Sabina Uban, Alina Maria Cristea, Anca Dinu, Liviu P. Dinu, Simona Georgescu, and Laurentiu Zoicas. 2021. Tracking semantic change in cognate sets for English and Romance languages. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 64–74, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Extracting cognates and borrowings data

A.1 Extraction

Parsing the Wiktionary The Wiktionary dumps mixes several formatting types, mostly HTML for the page tags and a pseudo-markdown for the internal structure of each article, which is not homogeneous between entries. The first step of processing was 1) to cut the Wiktionary by page, by literally cutting it on page HTML tags, and 2) at the same time, to only keep the title (lexeme) using HTML title tags and the text (core of the page) without the rest of the HTML using HTML text tags. Some pages were automatically discarded, if containing "Wiktionary", "App." or "Thesaurus" in their titles, as they are out of scope for the database.

Storing words and relations Once each page was cut, we cleaned the text, by extracting lexeme (first line), langs (second level pseudo markdown separation), and associated information (third levels pseudo markdown separations). The associated information was then cleaned using regexes, to find meanings (lines starting with an enumeration marker), descendants (using 'desc', 'desctree', and 'bor=1' as markers), ascendants (using 'inh' and 'root' as markers),[15] and supposed cognates (using 'cog' as marker). Lexemes were normalized using unicodedata. This allowed us to construct a list of Word objects, storing lexeme, lang, gloss, parent words, children words, and plausible cognates. (Related words were stored as "word_lang" in order to filter them). For each word, we added to its ancestor the set of its ancestors' ancestors, and we converted gloss for English lexemes to the English lexeme itself.

A.2 Constructing our cognates and borrowing sets

Lastly, we converted this list to our cognate and borrowing sets. For each word, we first stored indirect parents as borrowing relations (borrowing set) and direct parents as cognate relations (cognate set), for parent languages in our languages of interest. Then, we looked at each direct ancestor's children (no matter the direct ancestor language): if a given child was direct, both its relation to the parent and to the initial word were stored as cognates (for language pairs of interest). Else, we stored both relations in our 'borrowings' set (id.). In other terms, two words are kept if they share a common proto-form. If their ancestor is direct, we save them as cognates, else borrowings. We use an extended version of the notions of cognacy and borrowing defined in the introduction, and consider that the proto-words are also both cognates with their direct descendants, and in a borrowing relationship with their indirect descendants.

A.3 Cleaning

Extraction problems Splitting the document on HTML page limits was sometimes linked to pages not being cut at the right place, and the title tag not being recognised: some lexemes were stored as '<tag>' (they were removed). Some irregularities in meaning definitions appeared, such as #English not being removed, or some reference urls being accidentally added to the English meanings. All these were manually managed.

Special characters Some symbols were not homogeneous in the Wiktionary originally, and appeared under several forms, such as 'l' for 'or', '&lt' for '<', '&gt' for '>', '&amp' for '&', among others. They were manually removed to ensure consistency.

A.4 Results

Our most doted language pairs usually contain relations between generally higher-resourced contemporary languages (FR-ES, IT-ES, PT-ES, FR-IT, IT-PT, more than 1,000 pairs), as well as, surprisingly, the FR-FRM pair. Pairs with Latin and other contemporary languages follow, with our least doted language pairs being Middle French or Old Spanish to any language other than French or Spanish, and most languages to themselves (word pairs including two different descendants from a common origin word in the same language).

[15] The 'from' marker was too noisy and therefore ignored.

Cognates											
Lang	#words	#uniq	Pair								
Total	34574	8334	EN	ES	FR	FRM	IT	LA	OSP	PT	RO
EN	896	498	448	0	0	0	0	0	0	0	0
ES	6156	1403	0	270	1047	225	1222	763	263	1255	841
FR	6062	1377	0	1047	230	1208	958	660	84	952	693
FRM	2253	630	0	225	1208	13	200	202	21	198	173
IT	5363	1058	0	1222	958	200	141	696	101	1080	824
LA	3573	1309	0	763	660	202	696	0	78	668	506
OSP	710	188	0	263	84	21	101	78	2	91	68
PT	5451	1103	0	1255	952	198	1080	668	91	209	789
RO	4110	768	0	841	693	173	824	506	68	789	108
borrowings											
Lang	#words	#uniq	Pair								
Total	5042	2925	EN	ES	FR	FRM	IT	LA	OSP	PT	RO
EN	2456	873	0	418	711	226	399	0	40	405	257
ES	435	354	418	0	12	4	0	1	0	0	0
FR	756	574	711	12	0	0	18	0	1	11	3
FRM	242	177	226	4	0	0	6	0	1	4	1
IT	424	348	399	0	18	6	0	1	0	0	0
LA	4	1	0	1	0	0	1	0	0	1	1
OSP	42	36	40	0	1	1	0	0	0	0	0
PT	421	341	405	0	11	4	0	1	0	0	0
RO	262	221	257	0	3	1	0	1	0	0	0

Table 3: Cognate and borrowings pairs relations

B Full results tables

The tables contain the number of cognate pairs kept for each language pairs, as well as an embedding similarity score between 1) both cognates/borrowings of a given pair, 2) both glosses of a given pair, 3) each cognate/borrowing to its gloss. Results are split by language pair and category (meaning shift or no meaning shift).

Cognates shifted in meaning	EN-EN	ES-EN	ES-ES	ES-FRM	ES-IT	ES-LA	ES-OSP	ES-RO
cognate (a) ↔ cognate (b)	27 ± 18		32 ± 18	4 ± 8	45 ± 21	35 ± 16	38 ± 10	28 ± 16
meaning (a) ↔ meaning (b)	27 ± 18		42 ± 19	52 ± 24	55 ± 24	59 ± 20	69 ± 19	50 ± 23
cognate ↔ meaning	100 ± 0		26 ± 17	13 ± 18	26 ± 16	19 ± 16	22 ± 16	23 ± 16
#items	706	0	474	304	2002	1172	276	1230
Cognates similar in meanings	EN-EN	ES-EN	ES-ES	ES-FRM	ES-IT	ES-LA	ES-OSP	ES-RO
cognate (a) ↔ cognate (b)			49 ± 26	3 ± 8	62 ± 17	41 ± 15	40 ± 8	43 ± 17
meaning (a) ↔ meaning (b)			100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
cognate ↔ meaning			22 ± 22	24 ± 27	41 ± 16	27 ± 17	29 ± 20	34 ± 18
#items	0	0	14	64	286	86	136	230
Cognates shifted in meaning	FR-EN	FR-ES	FR-FR	FR-FRM	FR-IT	FR-LA	FR-OSP	FR-RO
cognate (a) ↔ cognate (b)		39 ± 20	32 ± 17	5 ± 7	41 ± 21	26 ± 11	25 ± 11	27 ± 16
meaning (a) ↔ meaning (b)		52 ± 23	33 ± 18	63 ± 24	54 ± 23	56 ± 21	54 ± 25	49 ± 24
cognate ↔ meaning		25 ± 16	25 ± 16	13 ± 17	24 ± 16	18 ± 15	20 ± 14	22 ± 15
#items	0	1690	410	1194	1512	1026	108	1024
Cognates similar in meanings	FR-EN	FR-ES	FR-FR	FR-FRM	FR-IT	FR-LA	FR-OSP	FR-RO
cognate (a) ↔ cognate (b)		53 ± 20	66 ± 29	3 ± 8	57 ± 16	31 ± 9	32 ± 8	40 ± 16
meaning (a) ↔ meaning (b)		100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
cognate ↔ meaning		35 ± 17	31 ± 20	19 ± 24	38 ± 15	28 ± 15	27 ± 18	34 ± 16
#items	0	256	10	706	250	74	30	198
Cognates shifted in meaning	FRM-EN	FRM-FRM	IT-EN	IT-FRM	IT-IT	IT-LA	IT-OSP	LA-EN
cognate (a) ↔ cognate (b)		42 ± 30		5 ± 9	33 ± 17	30 ± 12	33 ± 10	
meaning (a) ↔ meaning (b)		35 ± 12		53 ± 25	41 ± 21	60 ± 21	61 ± 25	
cognate ↔ meaning		-0 ± 6		13 ± 18	24 ± 16	19 ± 15	23 ± 14	
#items	0	14	0	270	236	1088	134	0
Cognates similar in meanings	FRM-EN	FRM-FRM	IT-EN	IT-FRM	IT-IT	IT-LA	IT-OSP	LA-EN
cognate (a) ↔ cognate (b)		64 ± 21		5 ± 8	47 ± 30	30 ± 12	34 ± 10	
meaning (a) ↔ meaning (b)		100 ± 0		100 ± 0	100 ± 0	100 ± 0	100 ± 0	
cognate ↔ meaning		-8 ± 8		24 ± 25	24 ± 21	25 ± 18	30 ± 17	
#items	0	8	0	68	12	74	42	0
Cognates shifted in meaning	LA-FRM	LA-LA	LA-OSP	OSP-EN	OSP-FRM	OSP-OSP	RO-EN	RO-FRM
cognate (a) ↔ cognate (b)	3 ± 7		28 ± 9		2 ± 6	76 ± 0		4 ± 8
meaning (a) ↔ meaning (b)	53 ± 22		66 ± 18		70 ± 30	23 ± 0		44 ± 24
cognate ↔ meaning	6 ± 10		13 ± 9		7 ± 12	22 ± 10		11 ± 15
#items	274	0	96	0	12	2	0	170
Cognates similar in meanings	LA-FRM	LA-LA	LA-OSP	OSP-EN	OSP-FRM	OSP-OSP	RO-EN	RO-FRM
cognate (a) ↔ cognate (b)	-1 ± 7		27 ± 11		2 ± 7			4 ± 7
meaning (a) ↔ meaning (b)	100 ± 0		100 ± 0		100 ± 0			100 ± 0
cognate ↔ meaning	12 ± 14		18 ± 14		13 ± 15			17 ± 19
#items	30	0	20	0	20	0	0	102
Cognates shifted in meaning	RO-IT	RO-LA	RO-OSP	RO-RO				
cognate (a) ↔ cognate (b)	30 ± 17	21 ± 10	18 ± 10	29 ± 24				
meaning (a) ↔ meaning (b)	52 ± 23	57 ± 21	54 ± 26	38 ± 16				
cognate ↔ meaning	23 ± 15	15 ± 13	16 ± 14	19 ± 15				
#items	1210	682	64	136				
Cognates similar in meanings	RO-IT	RO-LA	RO-OSP	RO-RO				
cognate (a) ↔ cognate (b)	39 ± 17	26 ± 10	26 ± 8	40 ± 18				
meaning (a) ↔ meaning (b)	100 ± 0	100 ± 0	100 ± 0	100 ± 0				
cognate ↔ meaning	32 ± 17	22 ± 14	26 ± 13	20 ± 8				
#items	230	44	38	6				

Table 4: Cognate results for fasttext embeddings

Borrowings shifted in meaning	EN-EN	ES-EN	ES-ES	ES-FRM	ES-IT	ES-LA	ES-OSP	ES-RO
borrowing (a) ↔ borrowing (b)		13 ± 16		-2 ± 4		29 ± 0		
meaning (a) ↔ meaning (b)		38 ± 21		62 ± 4		47 ± 0		
borrowing ↔ meaning		64 ± 38		10 ± 16		14 ± 14		
#items	0	704	0	4	0	2	0	0
Borrowings similar in meanings	EN-EN	ES-EN	ES-ES	ES-FRM	ES-IT	ES-LA	ES-OSP	ES-RO
borrowing (a) ↔ borrowing (b)		37 ± 15		-0 ± 8				
meaning (a) ↔ meaning (b)		100 ± 0		100 ± 0				
borrowing ↔ meaning		69 ± 33		31 ± 30				
#items	0	40	0	4	0	0	0	0
Borrowings shifted in meaning	FR-EN	FR-ES	FR-FR	FR-FRM	FR-IT	FR-LA	FR-OSP	FR-RO
borrowing (a) ↔ borrowing (b)	14 ± 17	46 ± 15			43 ± 13			38 ± 10
meaning (a) ↔ meaning (b)	42 ± 24	51 ± 16			63 ± 22			78 ± 5
borrowing ↔ meaning	61 ± 40	22 ± 15			28 ± 16			38 ± 13
#items	1066	10	0	0	28	0	0	4
Borrowings similar in meanings	FR-EN	FR-ES	FR-FR	FR-FRM	FR-IT	FR-LA	FR-OSP	FR-RO
borrowing (a) ↔ borrowing (b)	39 ± 17	63 ± 11			60 ± 18		43 ± 0	36 ± 0
meaning (a) ↔ meaning (b)	100 ± 0	100 ± 0			100 ± 0		100 ± 0	100 ± 0
borrowing ↔ meaning	69 ± 33	39 ± 19			47 ± 13		39 ± 19	36 ± 15
#items	206	12	0	0	8	0	2	2
Borrowings shifted in meaning	FRM-EN	FRM-FRM	IT-EN	IT-FRM	IT-IT	IT-LA	IT-OSP	LA-EN
borrowing (a) ↔ borrowing (b)	-1 ± 7		15 ± 16	4 ± 11		32 ± 0		
meaning (a) ↔ meaning (b)	34 ± 21		39 ± 21	42 ± 16		60 ± 0		
borrowing ↔ meaning	51 ± 50		63 ± 39	3 ± 16		18 ± 18		
#items	278	0	702	10	0	2	0	0
Borrowings similar in meanings	FRM-EN	FRM-FRM	IT-EN	IT-FRM	IT-IT	IT-LA	IT-OSP	LA-EN
borrowing (a) ↔ borrowing (b)	-1 ± 8		33 ± 19					
meaning (a) ↔ meaning (b)	100 ± 0		100 ± 0					
borrowing ↔ meaning	49 ± 51		67 ± 36					
#items	72	0	36	0	0	0	0	0
Borrowings shifted in meaning	LA-FRM	LA-LA	LA-OSP	OSP-EN	OSP-FRM	OSP-OSP	RO-EN	RO-FRM
borrowing (a) ↔ borrowing (b)				8 ± 10			9 ± 12	
meaning (a) ↔ meaning (b)				36 ± 21			31 ± 17	
borrowing ↔ meaning				58 ± 42			60 ± 42	
#items	0	0	0	56	0	0	384	0
Borrowings similar in meanings	LA-FRM	LA-LA	LA-OSP	OSP-EN	OSP-FRM	OSP-OSP	RO-EN	RO-FRM
borrowing (a) ↔ borrowing (b)				3 ± 7	12 ± 0		22 ± 7	
meaning (a) ↔ meaning (b)				100 ± 0	100 ± 0		100 ± 0	
borrowing ↔ meaning				52 ± 49	7 ± 5		61 ± 39	
#items	0	0	0	6	2	0	14	0
Borrowings shifted in meaning	RO-IT	RO-LA	RO-OSP	RO-RO				
borrowing (a) ↔ borrowing (b)		24 ± 0						
meaning (a) ↔ meaning (b)		49 ± 0						
borrowing ↔ meaning		21 ± 20						
#items	0	2	0	0				
Borrowings similar in meanings	RO-IT	RO-LA	RO-OSP	RO-RO				
borrowing (a) ↔ borrowing (b)								
meaning (a) ↔ meaning (b)								
borrowing ↔ meaning								
#items	0	0	0	0				

Table 5: Borrowings results for fasttext embeddings

Cognates shifted in meaning	EN-EN	ES-EN	ES-ES	ES-FRM	ES-IT	ES-LA	ES-OSP	ES-RO
cognate (a) ↔ cognate (b)	73 ± 8		79 ± 6	77 ± 8	84 ± 10	74 ± 8	87 ± 8	77 ± 8
meaning (a) ↔ meaning (b)	84 ± 6		81 ± 6	81 ± 6	83 ± 7	83 ± 7	84 ± 5	81 ± 7
cognate ↔ meaning	73 ± 5		70 ± 5	69 ± 5	70 ± 5	68 ± 6	74 ± 5	69 ± 5
#items	620	0	450	322	1962	906	300	1200
Cognates similar in meanings	EN-EN	ES-EN	ES-ES	ES-FRM	ES-IT	ES-LA	ES-OSP	ES-RO
cognate (a) ↔ cognate (b)			76 ± 6	81 ± 7	89 ± 8	77 ± 10	86 ± 8	81 ± 9
meaning (a) ↔ meaning (b)			100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
cognate ↔ meaning			65 ± 7	67 ± 6	68 ± 5	65 ± 7	72 ± 7	67 ± 5
#items	0	0	10	64	270	46	150	216
Cognates shifted in meaning	FR-EN	FR-ES	FR-FR	FR-FRM	FR-IT	FR-LA	FR-OSP	FR-RO
cognate (a) ↔ cognate (b)		79 ± 9	77 ± 7	90 ± 8	80 ± 9	72 ± 8	77 ± 7	76 ± 8
meaning (a) ↔ meaning (b)		83 ± 7	79 ± 6	83 ± 7	82 ± 7	82 ± 7	82 ± 5	81 ± 6
cognate ↔ meaning		70 ± 5	69 ± 6	69 ± 6	69 ± 5	68 ± 6	74 ± 6	68 ± 6
#items	0	1632	388	1314	1500	824	112	974
Cognates similar in meanings	FR-EN	FR-ES	FR-FR	FR-FRM	FR-IT	FR-LA	FR-OSP	FR-RO
cognate (a) ↔ cognate (b)		84 ± 8	78 ± 3	91 ± 8	84 ± 8	75 ± 8	79 ± 8	80 ± 7
meaning (a) ↔ meaning (b)		100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
cognate ↔ meaning		68 ± 5	70 ± 5	69 ± 6	68 ± 5	66 ± 6	72 ± 6	66 ± 5
#items	0	252	8	780	240	56	34	198
Cognates shifted in meaning	FRM-EN	FRM-FRM	IT-EN	IT-FRM	IT-IT	IT-LA	IT-OSP	LA-EN
cognate (a) ↔ cognate (b)		82 ± 5		79 ± 8	78 ± 6	77 ± 8	82 ± 8	
meaning (a) ↔ meaning (b)		73 ± 8		81 ± 7	81 ± 7	83 ± 7	84 ± 5	
cognate ↔ meaning		66 ± 8		69 ± 5	69 ± 5	68 ± 6	74 ± 6	
#items	0	10	0	292	236	874	134	0
Cognates similar in meanings	FRM-EN	FRM-FRM	IT-EN	IT-FRM	IT-IT	IT-LA	IT-OSP	LA-EN
cognate (a) ↔ cognate (b)		84 ± 3		80 ± 8	86 ± 9	81 ± 9	84 ± 6	
meaning (a) ↔ meaning (b)		100 ± 0		100 ± 0	100 ± 0	100 ± 0	100 ± 0	
cognate ↔ meaning		64 ± 9		67 ± 5	66 ± 4	66 ± 5	73 ± 5	
#items	0	10	0	76	10	52	42	0
Cognates shifted in meaning	LA-FRM	LA-LA	LA-OSP	OSP-EN	OSP-FRM	OSP-OSP	RO-EN	RO-FRM
cognate (a) ↔ cognate (b)	74 ± 8		77 ± 7		83 ± 6	84 ± 0		74 ± 7
meaning (a) ↔ meaning (b)	79 ± 7		82 ± 6		89 ± 3	87 ± 0		80 ± 6
cognate ↔ meaning	67 ± 6		72 ± 7		74 ± 5	78 ± 4		67 ± 6
#items	272	0	90	0	12	2	0	172
Cognates similar in meanings	LA-FRM	LA-LA	LA-OSP	OSP-EN	OSP-FRM	OSP-OSP	RO-EN	RO-FRM
cognate (a) ↔ cognate (b)	76 ± 8		78 ± 6		79 ± 5	79 ± 0		78 ± 6
meaning (a) ↔ meaning (b)	100 ± 0		100 ± 0		100 ± 0	100 ± 0		100 ± 0
cognate ↔ meaning	66 ± 6		70 ± 6		71 ± 7	72 ± 1		67 ± 5
#items	24	0	22	0	26	2	0	110
Cognates shifted in meaning	RO-IT	RO-LA	RO-OSP	RO-RO				
cognate (a) ↔ cognate (b)	79 ± 8	73 ± 8	76 ± 5	77 ± 7				
meaning (a) ↔ meaning (b)	81 ± 7	81 ± 7	84 ± 4	80 ± 6				
cognate ↔ meaning	68 ± 5	66 ± 6	73 ± 6	69 ± 6				
#items	1218	548	66	144				
Cognates similar in meanings	RO-IT	RO-LA	RO-OSP	RO-RO				
cognate (a) ↔ cognate (b)	81 ± 9	75 ± 10	79 ± 6	83 ± 13				
meaning (a) ↔ meaning (b)	100 ± 0	100 ± 0	100 ± 0	100 ± 0				
cognate ↔ meaning	66 ± 4	64 ± 5	71 ± 6	62 ± 7				
#items	224	34	48	6				

Table 6: Cognates results for BERT embeddings, using the last 4 layers

Borrowings shifted in meaning	EN-EN	ES-EN	ES-ES	ES-FRM	ES-IT	ES-LA	ES-OSP	ES-RO
borrowing (a) ↔ borrowing (b)		75 ± 8		83 ± 2				
meaning (a) ↔ meaning (b)		79 ± 6		74 ± 5				
borrowing ↔ meaning		72 ± 5		70 ± 5				
#items	0	636	0	4	0	0	0	0
Borrowings similar in meanings	EN-EN	ES-EN	ES-ES	ES-FRM	ES-IT	ES-LA	ES-OSP	ES-RO
borrowing (a) ↔ borrowing (b)		85 ± 7		92 ± 1				
meaning (a) ↔ meaning (b)		100 ± 0		100 ± 0				
borrowing ↔ meaning		73 ± 5		67 ± 5				
#items	0	40	0	4	0	0	0	0
Borrowings shifted in meaning	FR-EN	FR-ES	FR-FR	FR-FRM	FR-IT	FR-LA	FR-OSP	FR-RO
borrowing (a) ↔ borrowing (b)	78 ± 10	84 ± 8			83 ± 8			83 ± 7
meaning (a) ↔ meaning (b)	80 ± 6	81 ± 10			84 ± 6			83 ± 1
borrowing ↔ meaning	72 ± 6	71 ± 5			71 ± 5			70 ± 2
#items	974	10	0	0	28	0	0	4
Borrowings similar in meanings	FR-EN	FR-ES	FR-FR	FR-FRM	FR-IT	FR-LA	FR-OSP	FR-RO
borrowing (a) ↔ borrowing (b)	87 ± 8	92 ± 4			87 ± 5		92 ± 0	73 ± 0
meaning (a) ↔ meaning (b)	100 ± 0	100 ± 0			100 ± 0		100 ± 0	100 ± 0
borrowing ↔ meaning	72 ± 6	69 ± 5			70 ± 5		79 ± 4	69 ± 1
#items	200	12	0	0	8	0	2	2
Borrowings shifted in meaning	FRM-EN	FRM-FRM	IT-EN	IT-FRM	IT-IT	IT-LA	IT-OSP	LA-EN
borrowing (a) ↔ borrowing (b)	75 ± 8		77 ± 8	81 ± 6				
meaning (a) ↔ meaning (b)	82 ± 6		79 ± 6	82 ± 5				
borrowing ↔ meaning	71 ± 6		72 ± 5	67 ± 5				
#items	274	0	632	10	0	0	0	0
Borrowings similar in meanings	FRM-EN	FRM-FRM	IT-EN	IT-FRM	IT-IT	IT-LA	IT-OSP	LA-EN
borrowing (a) ↔ borrowing (b)	82 ± 8		85 ± 9					
meaning (a) ↔ meaning (b)	100 ± 0		100 ± 0					
borrowing ↔ meaning	71 ± 6		74 ± 5					
#items	82	0	34	0	0	0	0	0
Borrowings shifted in meaning	LA-FRM	LA-LA	LA-OSP	OSP-EN	OSP-FRM	OSP-OSP	RO-EN	RO-FRM
borrowing (a) ↔ borrowing (b)				76 ± 8			74 ± 8	
meaning (a) ↔ meaning (b)				84 ± 4			80 ± 6	
borrowing ↔ meaning				77 ± 3			70 ± 6	
#items	0	0	0	56	0	0	340	0
Borrowings similar in meanings	LA-FRM	LA-LA	LA-OSP	OSP-EN	OSP-FRM	OSP-OSP	RO-EN	RO-FRM
borrowing (a) ↔ borrowing (b)				75 ± 11	82 ± 0		75 ± 10	
meaning (a) ↔ meaning (b)				100 ± 0	100 ± 0		100 ± 0	
borrowing ↔ meaning				74 ± 6	72 ± 2		70 ± 5	
#items	0	0	0	6	2	0	16	0
Borrowings shifted in meaning	RO-IT	RO-LA	RO-OSP	RO-RO				
borrowing (a) ↔ borrowing (b)								
meaning (a) ↔ meaning (b)								
borrowing ↔ meaning								
#items	0	0	0	0				
Borrowings similar in meanings	RO-IT	RO-LA	RO-OSP	RO-RO				
borrowing (a) ↔ borrowing (b)								
meaning (a) ↔ meaning (b)								
borrowing ↔ meaning								
#items	0	0	0	0				

Table 7: Borrowings results for BERT embeddings, using the last 4 layers

C Corpora for Embeddings Training

C.1 Data collection sources

All datasets are under open, CC BY, or CC BY-NC-SA licences, and our chosen subset will be released with the paper. LEM17 is found at https://github.com/e-ditiones/LEM17, MCVF 1.0/2.0 and PPCHF 1.0 at https://github.com/beatrice57/mcvf-plus-ppchf, OpenMedFr at https://github.com/OpenMedFr/texts, BFM2019 at http://txm.ish-lyon.cnrs.fr/bfm/?path=/BFM2019, and the Digital Library of Old Spanish Texts at http://hispanicseminary.org/t&c/nar/index-en.htm.

C.2 FRM preprocessing

The LEM files were in csv format for UD, and only the words (first column) were extracted. The BFM2019 and MCVF v1 files were in XML format, and the div containing text were selected. The MCVF v2 and PPCHF files were in text format, parsed, and text was extracted from the correct lines. Lastly, the OpenMedFr were already in raw text format, and we only had to remove the comment lines and page indications. Then, all files were automatically separated on end of sentence punctuation mark (full stop, exclamation mark, question mark), then manually on indicators of dialogue (dashes, quotation marks) to keep one sentence per line. The line creation process could have introduced some noise. One specificity of FRM is the presence of extremely long sentences divided into sub-sentences with commas. Thus, we perform a secondary split around commas when the sentences are too long to ease the model fine-tuning and embeddings extraction steps.

C.3 Fine-tuning experiments

Cognates un-shifted in meanings	OSP-ft	OSP$_{ft}$-FR	OSP-ES	OSP$_{ft}$-ES	OSP-RO	OSP$_{ft}$-RO
cognate (a) ↔ cognate (b)	79 ± 8	75 ± 7	86 ± 8	79 ± 7	79 ± 6	74 ± 8
meaning (a) ↔ meaning (b)	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
cognate ↔ meaning	72 ± 6	67 ± 5	72 ± 7	68 ± 5	71 ± 6	66 ± 5
Cognates shifted in meaning	OSP-ft	OSP$_{ft}$-FR	OSP-ES	OSP$_{ft}$-ES	OSP-RO	OSP$_{ft}$-RO
cognate (a) ↔ cognate (b)	77 ± 7	72 ± 8	87 ± 8	79 ± 7	76 ± 5	70 ± 6
meaning (a) ↔ meaning (b)	82 ± 5	82 ± 5	84 ± 5	84 ± 5	84 ± 4	84 ± 4
cognate ↔ meaning	74 ± 6	69 ± 5	74 ± 5	69 ± 5	73 ± 6	69 ± 5
Shift measure	1	3	-0	1	3	4
Cognates un-shifted in meanings	OSP-IT	OSP$_{ft}$-IT	OSP-LA	OSP$_{ft}$-LA	OSP-FRM	OSP$_{ft}$-FRM
cognate (a) ↔ cognate (b)	84 ± 6	79 ± 4	78 ± 6	72 ± 7	79 ± 5	71 ± 7
meaning (a) ↔ meaning (b)	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
cognate ↔ meaning	73 ± 5	69 ± 3	70 ± 6	66 ± 4	71 ± 7	67 ± 5
Cognates shifted in meaning	OSP-IT	OSP$_{ft}$-IT	OSP-LA	OSP$_{ft}$-LA	OSP-FRM	OSP$_{ft}$-FRM
cognate (a) ↔ cognate (b)	82 ± 8	76 ± 8	77 ± 7	69 ± 8	83 ± 6	74 ± 7
meaning (a) ↔ meaning (b)	84 ± 5	84 ± 5	82 ± 6	82 ± 6	89 ± 3	89 ± 3
cognate ↔ meaning	74 ± 6	69 ± 5	72 ± 7	68 ± 5	74 ± 5	70 ± 3
Shift measure	2	3	1	3	-4	-3
Cognates un-shifted in meanings	FRM-ft	FRM$_{ft}$-FR	FRM-ES	FRM$_{ft}$-ES	FRM-RO	FRM$_{ft}$-RO
cognate (a) ↔ cognate (b)	91 ± 8	84 ± 6	81 ± 7	78 ± 7	78 ± 6	75 ± 7
meaning (a) ↔ meaning (b)	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
cognate ↔ meaning	69 ± 6	68 ± 6	67 ± 6	67 ± 6	67 ± 5	67 ± 5
Cognates shifted in meaning	FRM-ft	FRM$_{ft}$-FR	FRM-ES	FRM$_{ft}$-ES	FRM-RO	FRM$_{ft}$-RO
cognate (a) ↔ cognate (b)	90 ± 8	83 ± 7	77 ± 8	74 ± 7	74 ± 7	70 ± 7
meaning (a) ↔ meaning (b)	83 ± 7	83 ± 7	81 ± 6	81 ± 6	80 ± 6	80 ± 6
cognate ↔ meaning	69 ± 6	68 ± 6	69 ± 5	68 ± 5	67 ± 6	66 ± 6
Shift measure	2	2	4	4	5	5
Cognates un-shifted in meanings	FRM-IT	FRM$_{ft}$-IT	FRM-LA	FRM$_{ft}$-LA	FRM-OSP	FRM$_{ft}$-OSP
cognate (a) ↔ cognate (b)	80 ± 8	77 ± 6	76 ± 8	71 ± 7	79 ± 5	73 ± 6
meaning (a) ↔ meaning (b)	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0	100 ± 0
cognate ↔ meaning	67 ± 5	66 ± 5	66 ± 6	65 ± 6	71 ± 7	71 ± 7
Cognates shifted in meaning	FRM-IT	FRM$_{ft}$-IT	FRM-LA	FRM$_{ft}$-LA	FRM-OSP	FRM$_{ft}$-OSP
cognate (a) ↔ cognate (b)	79 ± 8	75 ± 7	74 ± 8	69 ± 8	83 ± 6	77 ± 6
meaning (a) ↔ meaning (b)	81 ± 7	81 ± 7	79 ± 7	79 ± 7	89 ± 3	89 ± 3
cognate ↔ meaning	69 ± 5	68 ± 5	67 ± 6	66 ± 6	74 ± 5	73 ± 5
Shift measure	2	2	2	2	-4	-3

Table 8: Statistics when using mBERT embeddings, with OSP/FRM finetuning ($_{ft}$-) or without, for Old Spanish and Medieval French cognates. The 'shift measure' is the average difference between semantic item similarity, between non-shifted and shifted pairs.
The semantic shift between shifted and un-shifted items is slightly increased for fine-tuned OSP, and not at all for FRM, at the cost of an alignment drift with the meanings (line 3). We consider that this extremely small improvement is not worth the cost, and therefore only use vanilla embeddings. However, it would still be worth investigating how to improve fine-tuning.

Borrowings un-shifted in meanings	OSP-EN	OSP$_{ft}$-EN		
borrowing (a) ↔ borrowing (b)	75 ± 11	70 ± 12		
meaning (a) ↔ meaning (b)	100 ± 0	100 ± 0		
borrowing ↔ meaning	74 ± 6	69 ± 6		
Borrowings shifted in meaning	OSP-EN	OSP$_{ft}$-EN		
borrowing (a) ↔ borrowing (b)	76 ± 8	71 ± 7		
meaning (a) ↔ meaning (b)	84 ± 4	84 ± 4		
borrowing ↔ meaning	77 ± 3	72 ± 5		
Shift measure	-1	-2		
Borrowings un-shifted in meanings	FRM-ES	FRM$_{ft}$-ES	FRM-EN	FRM$_{ft}$-EN
borrowing (a) ↔ borrowing (b)	92 ± 1	86 ± 2	82 ± 8	78 ± 7
meaning (a) ↔ meaning (b)	100 ± 0	100 ± 0	100 ± 0	100 ± 0
borrowing ↔ meaning	67 ± 5	65 ± 6	71 ± 6	71 ± 7
Borrowings shifted in meaning	FRM-ES	FRM$_{ft}$-ES	FRM-EN	FRM$_{ft}$-EN
borrowing (a) ↔ borrowing (b)	83 ± 2	82 ± 3	75 ± 8	71 ± 7
meaning (a) ↔ meaning (b)	74 ± 5	74 ± 5	82 ± 6	82 ± 6
borrowing ↔ meaning	70 ± 5	70 ± 4	71 ± 6	70 ± 6
Shift measure	9	5	6	7

Table 9: Statistics when using mBERT embeddings, with OSP/FRM finetuning ($_{ft}$) or without, for Old Spanish and Medieval French borrowings with shifted and unshifted pairs. The 'shift measure' is the average difference between semantic item similarity, between non-shifted and shifted pairs.
We observe that the difference between shifted and non-shifted items decreases this time, when compared to cognates, for OSP-EN and FRM-ES, and increases for FRM-EN. We consider that variations are not consistent enough to draw conclusions.