Cross-lingual Lexical Semantic Change in Romance Languages

Ana Sabina Uban[♠,♡] **Liviu P. Dinu**[♠,♡]
Anca Dinu[♣,♡] **Simona Georgescu**[♣,♡]
University of Bucharest, [♠] Faculty of Mathematics and Computer Science,
[♣]Faculty of Foreign Languages and Literatures, [♡]HLT Research Center
{auban, ldinu}@fmi.unibuc.ro,
{anca.dinu, simona.georgescu}@lls.unibuc.ro

Abstract

We present a comprehensive analysis of lexical semantic change in the five main Romance languages (Romanian, Italian, Spanish, French and Portuguese), based on the most exhaustive database of related words in these languages. We include both cognate words and borrowings (for the first time, to our knowledge), and compute semantic shift measures using different static and contextual embedding models, as well as three different corpora. We publish[1] the obtained lists of semantic divergences across all related word pairs, compute global trends in language-level semantic divergence, and provide insights on particular study cases of highly stable and highly divergent words for different language pairs.

1 Introduction

Approximately 27% of the lexicon of Romance languages consists of words inherited from Latin, and another 40% consists of words borrowed from Latin (Reinheimer-Rîpeanu, 2001). However, the 27% represents the fundamental core of the vocabulary (consisting of commonly used concepts - e.g. family, body parts, natural elements, everyday actions, as well as pronouns, numerals, and prepositions) and has a frequency of 80% in everyday communication (Bîrlădeanu et al., 1988). We would therefore expect that their meaning has not changed – at least not significantly – from one Romance language to another. However, a quick glance at the REW (Romanisches Etymologisches Wörterbuch), a Romance languages etymological dictionary (Meyer-Lübke, 1911), shows us that faithful preservation of the Latin meaning is reserved only for a relatively small number of words that designate concrete realities, which have not changed over time and have not been subject to changes in perception (e.g. *manus* 'hand', *oculus* 'eye', *filius* 'son', dormire 'sleep', bibere 'drink', etc.). On the other hand, many words, including those from the fields mentioned above, have undergone divergent developments (e.g. Lat. *bucca* 'cheek' > It. *bocca*, Fr. *bouche*, Es. *boca*, Pt. *boca* 'mouth' vs. Rom. *bucă* 'butt cheek'; Lat. *salire* 'to jump' > It. *salire* 'climb up'/ Es. *salir* 'get out', etc.). This semantic divergence characterizes many pairs of cognates (e.g. Ro. *cugeta* 'to think'/ Es. *cuidar* 'to take care (of)', Ro. *vindeca* 'to cure'/ Pt. *vingar* 'to take revenge', Ro. *feri* 'to avoid' / Es. *herir* 'to hurt', etc.), whose divergent evolution has been insufficiently explained at a global level in specialized studies.

Similarly, words borrowed from Latin into Romance languages may undergo changes in meaning, albeit to a lesser extent, leading to obvious pairs of deceptive cognates (Uban et al., 2025): e.g. Es. *oficio* 'profession' / Ro. *oficiu* 'public service, office'; Ro. *transcendental* 'which is above the real world' / Es. *transcendental* 'of great importance', etc. They are still cognates, although they have entered the language through scholarly transmission.

Perhaps even more unexpectedly, there are also pairs of words borrowed from one Romance language to another that undergo a certain semantic shift, either by taking on a specialized meaning from the source language, or by subsequently changing meaning, either in the target language or in the source language (e.g. Fr. *hâbler* 'brag, boast', borrowed from Es. *hablar* 'speak'; Fr. *dame* 'lady' borrowed into Ro. *damă* 'prostitute'). These pairs of words are the ones that will be treated here as borrowings.

From the point of view of many linguists (Dworkin, 2006; Chauveau, 2016), semantic change seems to be the most difficult area to study because it is almost impossible to establish quantifiable parameters that can be analyzed according to scientific criteria in this field. While various formalization models have been attempted in synchronic

[1] https://nlp.unibuc.ro/resources.html#HistoricalLinguistics

The Proceedings for the 6th International Workshop on Computational Approaches to Language Change (LChange'26), pages 35–49

semantics, in diachronic semantics, i.e., the way in which the meaning of words evolves, attempts in this direction are concentrated in the field of computational linguistics and are still fairly recent, and mostly comprise of monolingual studies.

Therefore, we aim to automatically measure the semantic divergence between cognates in any two Romance languages from the main core (Romanian, Italian, French, Spanish, Portuguese) by representing them as contextual embeddings extracted from corpora (see section *Data*). Once we have obtained the distances between any two such representations of cognates, we can compute an aggregated global distance for any two languages, thus obtaining a corpus driven semantic divergence for the Romance language family. The interpretation of these results will show to what extent geographical distance is a source of proximity between languages, as proposed in Bartoli (1925), who showed that languages on the periphery of the former Roman Empire change less as a result of their isolation from the center) or, on the contrary, a source of semantic distance as a result of different interpretations of reality. We assume that a computational approach could provide a significant platform for a methodologically coherent analysis of semantic change, proposing quantifiable paradigms for historical semantics that can be explored using scientific tools, which can, of course, be improved or adapted depending on the particular task.

2 Related Work

Semantic change has become an increasingly central topic in computational historical linguistics over the past decade, driven by the availability of large multilingual corpora and significant advances in distributional and contextual semantic modeling. These research relied either on distributional similarity and static embeddings to capture cross-lingual meaning variation, or, more recently, on contextualized representations, clustering-based methods. Work on semantic change comprise both traditional diachronic approaches, usually monolingual, as well as synchronic, multi-lingual ones.

Montariol and Allauzen (2021) introduce a computational framework for tracking semantic divergence between translated word pairs across languages and time, showing that contextualized embeddings combined with clustering outperform static representations in capturing gradual meaning shifts.

Uban et al. (2021) analyze semantic change within cognate sets across English and Romance languages, highlighting systematic differences between cognate words and demonstrating that lexical properties such as frequency and polysemy correlate with semantic divergence.

Building on previous approaches, Kawasaki et al. (2022) explicitly draw on cross-lingual divergence measures to revisit classical statistical laws of semantic change in Romance cognates, refining the roles of frequency, polysemy, morphological complexity, and lexical age in explaining semantic stability and drift.

Most recently, Uban et al. (2025) extend this line of research through a large-scale computational investigation of semantic false friends across Romance languages, introducing new etymologically grounded resources and evaluation protocols that further clarify how shared origin can nevertheless result in substantial cross-linguistic semantic divergence.

This work draws from previous methodologies and data, integrating the most recent datasets and methods. Our contribution is twofold: on the one hand we perform all our experiments based on the exhaustive database of Romance related words RoBoCoP (Dinu et al., 2023), including cognates and borrowings, and, on the other, we employ different models including static and contextual embeddings for those related words, separately for cognates and borrowings, from three different multilingual and parallel large corpora, for multi-level comparison.

3 Data

3.1 Cognates and Borrowings Dataset

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

3.2 Word Embeddings Corpora

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

- *Wikipedia*[2]
- *Europarl*, a standard parallel corpus with aligned sentences including the Romance languages, based on proceedings of the European Parliament (Koehn, 2005),
- *RomCro2.0*, a recent parallel corpus including more general language sourced from literary works written in various original languages and translated in Romance languages and Croatian (Mikelenić et al., 2024).

4 Methodology

Our proposed algorithms rely on representing the related words in semantic space based on word embeddings and then measuring the semantic distances between them in the obtained multidimensional space, using two different embedding algorithms:

- contextual embeddings extracted from a BERT transformer pretrained on a multilingual sentence similarity task for optimizing sentence representations, based on a Sentence-BERT architecture (Reimers and Gurevych, 2019) [3], as well as the multilingual transformer `xlm-roberta-base` (Conneau et al., 2019) for a subset of the experiments,
- static FastText aligned embeddings (Bojanowski et al., 2016; Conneau et al., 2017), which have been previously used successfully for cognate semantic divergence measures (Uban et al., 2019; Uban and Dinu, 2020; Uban et al., 2021).

We use the publicly available pre-aligned multilingual static embedding spaces based on the Wikipedia corpus[4], obtained by training a linear transformation using the Procrustes alignment algorithm, as published in Conneau et al. (2017).

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

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

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

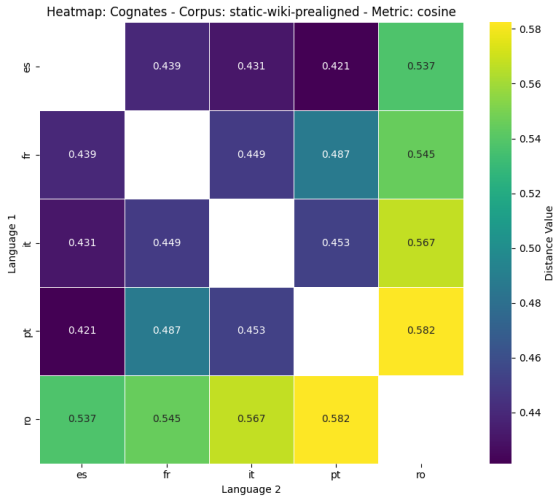


Figure 1: Semantic divergence between cognates in Romance languages based on cosine distance on static embeddings on the Wikipedia corpus

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

- mean distance: a simple dimension-wise average of the embeddings is computed to obtain unique representations per cognate, then cosine similarity is used to compute distances;
- JSD: embedding clusters for each cognate are generated using affinity propagation and cosine distance, the Janson-Shannon divergence is computed between the clusters as a distance metric between cognates;
- WiDiD (Periti et al., 2022): embedding clusters for each cognate are generated independently, and cluster centers are computed using simple averaging, then the distance between clusters is computed as the cosine distance between cluster centers.

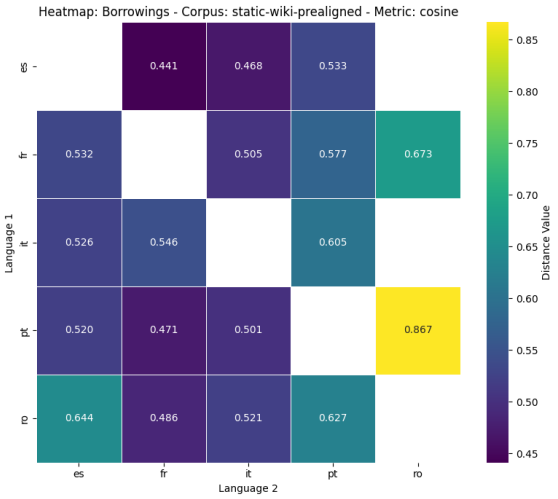
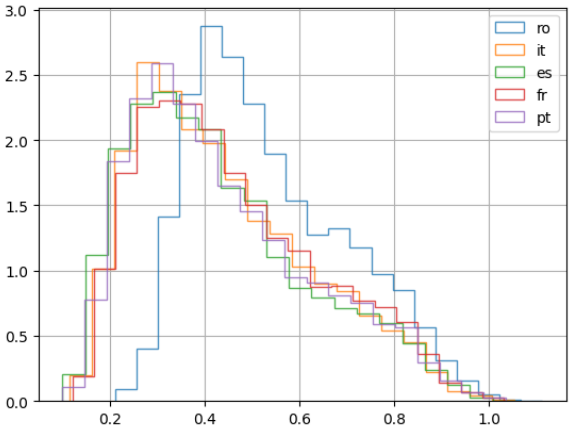
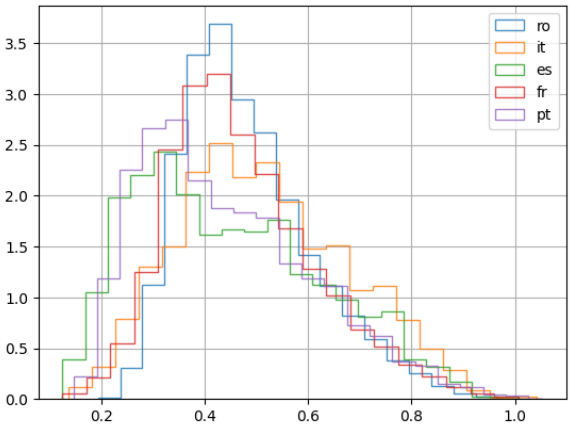


Figure 2: Semantic divergence between borrowings in Romance languages based on cosine distance on static embeddings on the Wikipedia corpus



(a) Distribution of cognates distances for each language based on Wikipedia static embeddings.



(b) Distribution of borrowings distances for each language based on Wikipedia static embeddings.

Figure 3: Distributions of cognates and borrowings distances for each language based on Wikipedia static embeddings.

5 Results

Global language divergence scores based on the Wikipedia corpus with static embeddings are shown in Figures 1 and 2 for cognates and borrowings, respectively. Figure 4 shows cognate semantic divergence computed with contextual embeddings on the three corpora. Similar heatmaps for borrowings using contextual embeddings, as well as results using additional distance metrics, are shown in the Appendix.

Figures 5 and 6 show semantic divergences based on contextual embeddings, as rankings of language pairs, from most distant to most similar. We notice few differences between the static and contextual embeddings results - the ranking of global language pair distances is generally maintained. Since the problem of evaluation is more difficult in the case of the present study, we rely on results in Uban et al. (2025) for choosing the models we choose to focus on primarily: here contextual embeddings based on Wikipedia using mean cosine distance, as well as static Wikipedia embeddings with cosine distance, are the most useful for detecting false friends.

The distribution of mean distances between related words in each language to any other language (where distances for each word in a given language to all related words, irrespective of their language, are averaged together), based on Wikipedia static embeddings, is shown in figures 3a for cognates and 3b for borrowings. We can observe that the curve for Romanian cognates is skewed to the right and more ample compared to all other Romance languages, meaning that Romanian cognates semantically diverged more than the Western cognates, most probably due to its geographical isolation in between non-Romance languages. The pattern is similar with the borrowings distribution, where Romanian borrowings changed the most, followed by French ones. Most distributions have a relatively normal shape, with a skew to the left. There is a slight multimodality in the distribution of Spanish and Portuguese borrowings, with two main peaks in the distributions around distances of 0.3 and 0.5: for both Spanish and Portuguese, the lower peak corresponds to borrowings from French (with an average distance of 0.44 for Fr-Es and 0.47 for Fr-Pt), and the higher peak to distances with each other (0.51 average distance for Es-Pt and 0.53 for Pt-Es). The individual distance distributions of borrowings from or to Spanish and Portuguese and

each of the other Romance languages are illustrated in the Appendix.

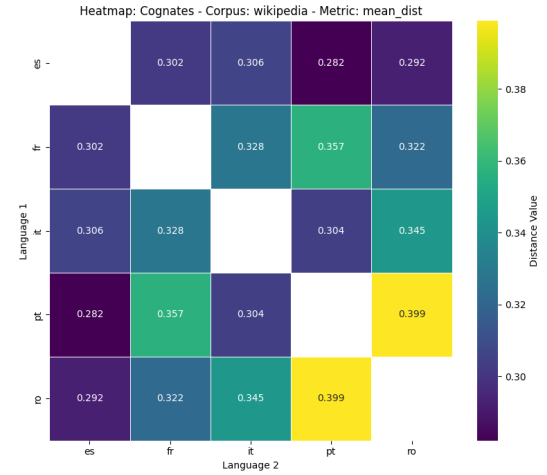
5.1 Global language divergence scores for different embedding models and corpora

For cognates, the mean distance between languages computed statically on Wikipedia and contextually on Wikipedia, Romcro, and Europarl corpora is represented in figures 5a, 5b, 5c, and 5d, respectively. The rankings of the semantic divergence between languages is dependent on the method and corpus, but the pattern is clear: the most divergent pair is Pt-Ro and the least divergent Es-Pt; on average, the pairs between Spanish and any other language show the lowest degrees of semantic differentiation, while pairs containing French are moderately divergent.

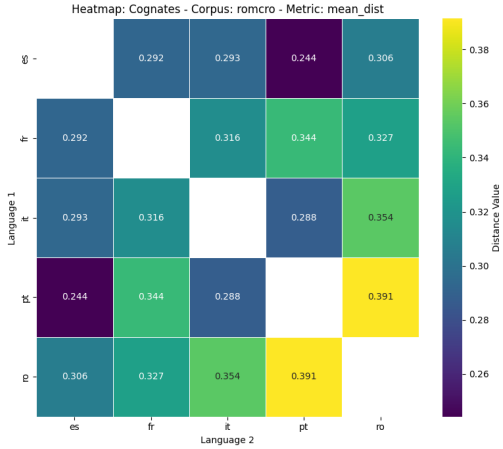
The proximity between Es-Pt in each corpus indicates not so much a greater degree of preservation in relation to Latin, but rather the fact that the two languages had a greater level of cohesion between them, a common evolution over a longer period of time and, at the same time, an evolution separate from other languages due to the geographical position of the Iberian Peninsula.

It is also noteworthy that in the static embeddings based on Wikipedia, all pairs containing Romanian show the highest level of divergence. This situation seems to reflect the effects of the isolation of the Romanian language, separated by a consistent Slavic fringe from the rest of the Romance languages, which predictably could lead to greater semantic divergence. At the same time, this distance, which appears in all corpora as the greatest, partially contradicts Bartoli's hypothesis, according to which lateral areas share more common features with each other than with the rest of the Romance languages. However, the distance between Ro and Es is significantly smaller than in other pairs, including those from central Romania, such as Fr-It.

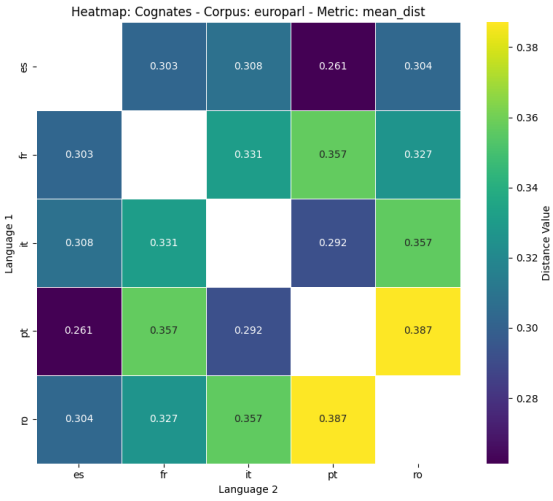
Following a comparative analysis of the corpora, it can be observed that the distances between Romanian and other languages (except for Portuguese) vary slightly, which required taking into account the specificity of the corpora. We were thus able to observe that in the corpus of parliamentary speeches, the distances are smaller because the language used is standard, specific to political and economic speeches, which leads to the use of a neological lexicon of Latin-Romance origin common to all Romance languages and, moreover, largely shared with English. In contrast, the language used



(a) Semantic divergence between cognates in Romance languages based on contextual embeddings trained on Wikipedia (using mean-dist)



(b) Semantic divergence between cognates in Romance languages based on contextual embeddings trained on RomCro (using mean-dist)



(c) Semantic divergence between cognates in Romance languages based on contextual embeddings trained on Europarl (using mean-dist)

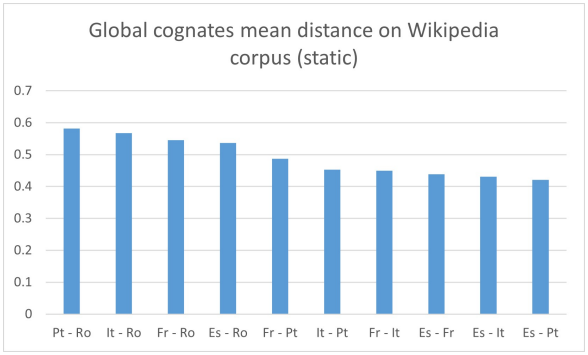
Figure 4: Heatmaps of semantic divergence scores between cognates in Romance languages based on contextual embeddings trained on three corpora (Wikipedia, RomCro, Europarl; using mean-dist).

in literature (the RomCro corpus) shows a greater variety, since the lexicon present in such texts is more diverse in terms of origin, while it does not give such a high weight to neologisms.

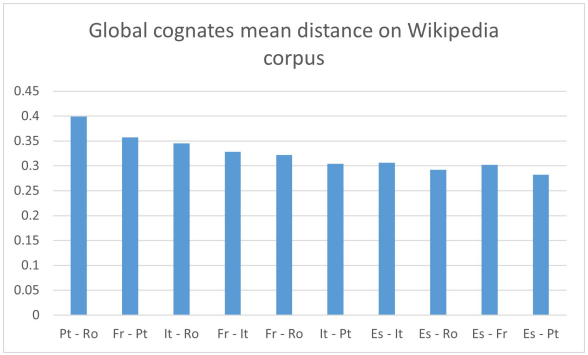
For borrowings, the number of language pairs is double, because the direction of borrowings matters. The mean semantic divergences between borrowings computed on static Wikipedia embeddings, as well as contextual embeddings on Wikipedia, Romcro, and Europarl are given in figures 6a, 6b, 6c, and 6d, respectively. When it comes to borrowing, things are much more nuanced, and patterns applicable to all languages cannot be detected, for several reasons. Firstly, the number of borrowings from one Romance language to another is incomparably smaller than cognates. Romanian will always produce unbalanced pairs because other languages have borrowed very few words from this language, which makes the data unreliable; Romanian, on the other hand, has borrowed heavily from French (approximately 9% of the Romanian vocabulary consists of French borrowings - although treated in lexicography as words of multiple etymology French/ Latin/ Italian -, a proportion that no other Romance language comes close to, cf. Reinheimer-Rîpeanu (2001)), but very few words from Spanish and almost none from Portuguese. Secondly, the conceptual domains from which borrowings were made are limited to certain elements that are perceived as specific to each culture, or that have penetrated a linguistic community along with the designated object (e.g. Romanian borrows from Spanish *telenovela*, *marijuana*, *cacao*; French from Italian *operetta*, *mosaico*, *miniatura*; Pt. from Es. *tabaco*, *guerrilla*, *coca*, etc.). The distances are small in the case of words that denote the same concept in the target language as they did in the source language (Fr. *chocolat*, *sangria*, It. *lama* < Es. *llama*), but large in situations where the borrowed word is taken with a single meaning selected from several that it had in the source language (Fr. *embargo* 'embargo' vs. Es. *embargo*, which is mostly used in the phrase *sin embargo* 'however', absent from Fr.).

5.2 Manual analysis of most changed and most stable related words

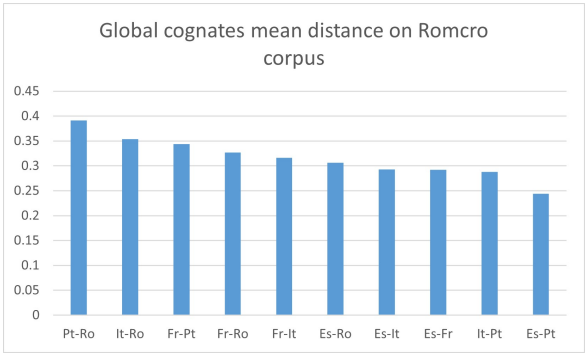
The results were manually analyzed in detail as follows. For each language pair, we sorted the cognates and borrowings according to the distance obtained and observed several patterns in terms of conceptual areas where low divergences occurred:



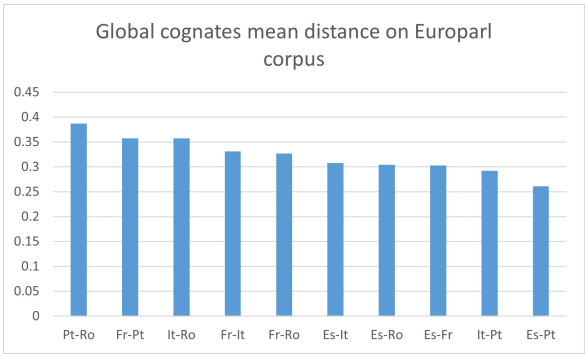
(a) Cognates semantic divergence between Romance languages based on static Wikipedia embeddings



(b) Cognates semantic divergence between Romance languages based on contextual embeddings trained on Wikipedia (using mean-dist)



(c) Cognates semantic divergence between Romance languages based on contextual embeddings trained on Romcro (using mean-dist)



(d) Cognates semantic divergence between Romance languages based on contextual embeddings trained on Europarl (using mean-dist)

Figure 5: Cognates semantic divergence between Romance languages based on different corpora (as language pair rankings).

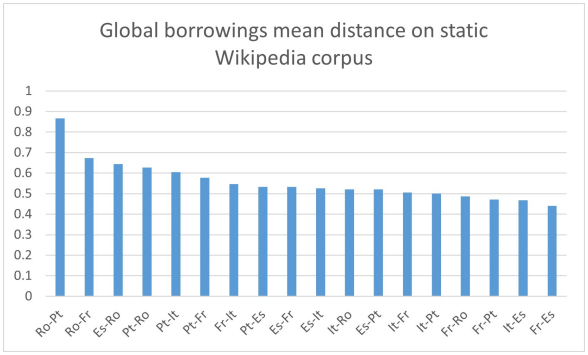
these are generally generic words that designate scientific fields or general areas of activity ('mathematics', 'astronomy', 'agriculture', 'medicine', etc.), univocal verbs, in other words, verbs that have not developed figurative meanings (to write, to kill), but also technical verbs ('to transport', 'to torture', 'to excommunicate'). At the other extreme we find terms that have either changed register (Ro. *muiere* is a regional and derogatory word for 'woman', whereas Es. *mujer* is the standard term for 'woman' and 'wife'), have restricted or expanded their area of application (e.g. Ro. *bucată* 'piece' - semantic expansion - vs Es. *bocado* 'bite' - from Lat. **buccata* 'mouthful'; Fr. *comprendre* 'to understand' - semantic narrowing - vs Ro. *cuprinde* 'to get hold of').

In most cases, the cause of the large semantic distance lies in the polysemic areas developed by cognates, which do not overlap, and therefore the terms appear in different contexts (e.g. Ro. *popor* 'people of a country' vs Es. *pueblo* 'people of a country' and 'village'). At the same time, cases of homonymy are misleading: while they can only be avoided through manual intervention, they result in the calculation of a large distance between cognates that would otherwise be semantically close: e.g. It. *aglio* 'garlic' etymologically corresponds to Ro. *ai* 'garlic' - a regional word -, which formally coincides with the the indicative 2nd pers. sg. of the verb *a avea* 'to have', incomparably more frequent: therefore, the divergence is 95% for this cognate pair; similarly, Es. *san* / *santo* 'saint' corresponds to Ro. *sân*, whose application is limited to contexts such as *Sân Nicolau*, *Sân Gheorghe*; otherwise, the adjective *sân* is homonymous with the noun *sân* 'breast'; in Pt-Es pair, Pt. *flama* 'flame' is the cognate of Sp. *llama* 'id.', which in its turn is homonymous with *llama* 'llama' (animal), and with the verb *llamar* 'to call', indicative 3rd pers. sg.).

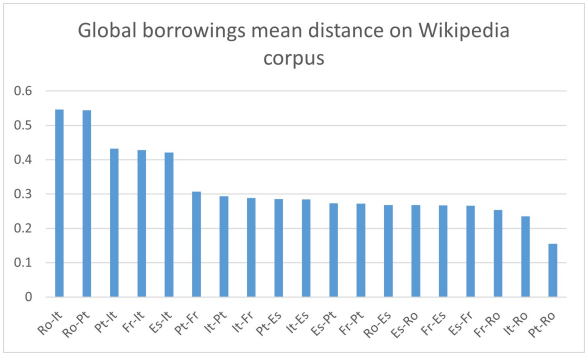
5.3 Part of speech distribution of words undergoing shifts

We separately measure semantic shifts for words with different parts of speech, according to Open multilingual WordNet. Some words can have multiple parts of speech according to WordNet - in these cases we consider them for both parts of speech. The coverage of words analyzed in WordNet is shown in the Appendix.

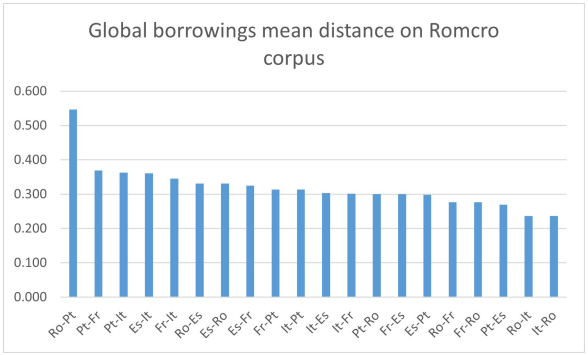
The mean cognate and borrowing distances, respectively, for each part of speech, based on



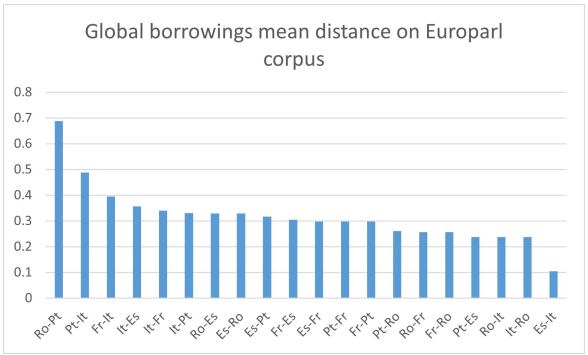
(a) Borrowings semantic divergence between Romance languages based on static Wikipedia embeddings.



(b) Borrowings semantic divergence between Romance languages based on contextual embeddings trained on Wikipedia (using mean-dist)



(c) Borrowings semantic divergence between Romance languages based on contextual embeddings trained on Romcro (using mean-dist)



(d) Borrowings semantic divergence between Romance languages based on contextual embeddings trained on Europarl (using mean-dist)

Figure 6: Borrowings semantic divergence between Romance languages across different corpora (as language pair rankings).

Wikipedia static embeddings are given in Tables 1 and 2 - shown for each language separately and overall for all words where we could extract the part of speech. The differences across POS are minor. Adjectives are most semantically stable across languages, particularly for borrowings, since borrowed adjectives are mostly relational adjectives, rather than describing a subjective quality (e.g. Fr-Es *biológico* 'biological', *informático* 'informational', Fr-Ro *casabil* 'breakable', *demonstrativ* 'demonstrative', *intestinal* 'intestinal'). For cognates, verbs are most stable overall, for most languages except for Romanian. Semantic shits in nouns vary by language: nouns are less stable in Italian, Spanish and Portuguese.

Table 1: Mean cognates distance for each part of speech, based on Wikipedia static embeddings.

Language	POS	#words	Avg. dist.
ro	noun	6200	0.5396
ro	verb	2379	0.5480
ro	adjective	777	0.5260
ro	adverb	1076	0.5275
it	noun	7352	0.4545
it	verb	3267	0.4148
it	adjective	1256	0.4343
it	adverb	137	0.5085
es	noun	5729	0.4432
es	verb	2736	0.3970
es	adjective	1100	0.4462
es	adverb	101	0.4781
fr	noun	5892	0.4616
fr	verb	3231	0.4378
fr	adjective	1002	0.4571
fr	adverb	282	0.5197
pt	noun	6701	0.4511
pt	verb	3464	0.4170
pt	adjective	937	0.4161
pt	adverb	163	0.5398
overall	noun	31874	0.470
overall	verb	15077	0.438
overall	adjective	5072	0.452
overall	adverb	1759	0.523

6 Conclusions and Future Work

We have presented a complete analysis of lexical semantic divergence in Romance languages based on different word embeddings models trained on different corpora, including the most exhaustive vocabulary of cognates as well as borrowings in Romance languages. We find the highest semantic proximity for related words in Spanish and Portuguese, both in the case of borrowings and cognate words. Romanian generally stands out with words

Table 2: Mean borrowings distance for each part of speech, based on Wikipedia static embeddings.

Language	POS	#words	Avg. dist.
ro	noun	8068	0.4844
ro	verb	1308	0.5093
ro	adjective	797	0.4570
ro	adverb	1136	0.4632
it	noun	2040	0.5109
it	verb	298	0.5045
it	adjective	296	0.5218
it	adverb	35	0.7112
es	noun	775	0.4527
es	verb	98	0.3975
es	adjective	52	0.4643
es	adverb	1	0.5784
fr	noun	9755	0.4737
fr	verb	2251	0.4890
fr	adjective	1737	0.4368
fr	adverb	103	0.5347
pt	noun	1340	0.4525
pt	verb	113	0.4022
pt	adjective	88	0.3714
pt	adverb	5	0.5318
overall	noun	21978	0.470
overall	verb	4068	0.492
overall	adjective	2970	0.449
overall	adverb	1280	0.476

diverging most from their cognates in other romance languages, with Romanian and Portuguese being the most distant language pair overall, in contradiction to Bartoli's lateral areas hypothesis. We find some differences in the rankings of language pairs based on average related word divergence, due to differences in vocabulary as well as divergence tendencies across language registers, confirming that including additional spoken corpora might be a useful complement to our results.

In the future, refining the contextual embedding representations by post-alignment of embedding spaces across languages could improve the precision of the resulted distance measures. Handling polysemy and treating each word sense separately could offer additional interesting insights.

Limitations

While we rely on three different multilingual corpora in different domains to obtain a complete perspective on the usage of the words analysed in different languages and contexts, complementing these with spoken language or social media corpora might be useful for capturing more subtle colloquial or metaphorical meanings.

The manual analysis of word pairs with high

semantic distance has revealed a small number of dictionary errors which might introduce some noise in the results.

Ethical Statement

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

Acknowledgements

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

References

Matteo G. Bartoli. 1925. *Introduzione alla neolinguistica: Principi, scopi, metodi*. Olschki, Firenze.

Mihaela Bîrlădeanu, M. Iliescu, Liliana Macarie, Ioana Nichita, Mariana Ploae-Hănganu, Marius Sala, Maria Theban, and Ioana Vintilă-Rădulescu. 1988. *Vocabularul reprezentativ al limbilor romanice*. Editura Ştiinţifică şi Enciclopedică, Bucureşti.

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

Jean-Paul Chauveau. 2016. Reconstruction comparative et histoire sémantique. In Éva Buchi and Wolfgang Schweickard, editors, *Dictionnaire Étymologique Roman (DÉRom) 2. Pratique lexicographique et réflexions théoriques*, pages 53–65. De Gruyter, Berlin/Boston.

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

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

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

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

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

Steven N. Dworkin. 2006. Recent developments in spanish (and romance) historical semantics. In *Selected Proceedings of the 8th Hispanic Linguistics Symposium*, pages 50–57, Somerville. Cascadilla Proceedings Project.

Yoshifumi Kawasaki, Maëlys Salingre, Marzena Karpinska, Hiroya Takamura, and Ryo Nagata. 2022. Revisiting statistical laws of semantic shift in romance cognates. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)*, pages 141–151, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

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

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

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.

Francesco Periti, Alfio Ferrara, Stefano Montanelli, and Martin Ruskov. 2022. What is done is done: an incremental approach to semantic shift detection. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 33–43, Dublin, Ireland. Association for Computational Linguistics.

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

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

Sanda Reinheimer-Rîpeanu. 2001. *Lingvistica romanică. Lexic – morfologie – fonetică*. All, Bucarest.

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

Ana Sabina Uban, 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*, pages 64–74, Online. Association for Computational Linguistics.

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.

Ana Sabina Uban, Liviu P. Dinu, Bogdan Iordache, Simona Georgescu, and Claudia Vlad. 2025. Friend or foe? a computational investigation of semantic false friends across romance languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15310–15324, Online. Association for Computational Linguistics.

A Appendix

A.1 Infrastructure and libraries

The experiments were performed on an RTX 2080 Ti GPU and a Ryzen 5 3600X CPU. Libraries used for embedding extraction, cognate and corpora preprocessing (extracting stems), synonym extraction based on WordNet, and distance metrics computation:

- keras==3.8.0
- keras-hub==0.18.1
- keras-nlp==0.18.1
- nltk==3.9.1
- scikit-learn==1.6.1
- scipy==1.13.1
- sentence-transformers==3.4.1
- spacy==3.7.5

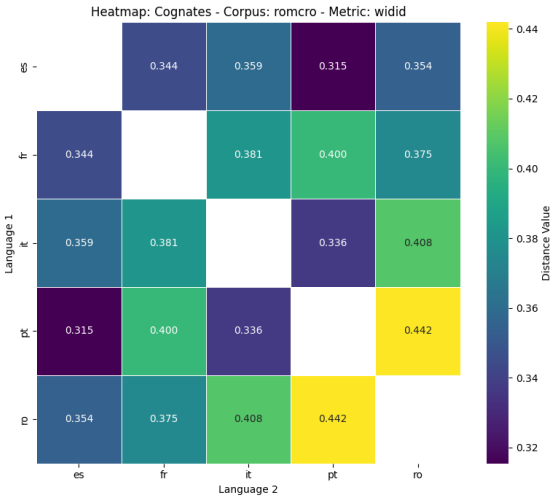


Figure 7: Semantic divergence between cognates in Romance languages based on contextual embeddings trained on RomCro (using Widid)

- tensorflow==2.18.0
- tensorflow-datasets==4.9.7
- transformers==4.48.3
- and fasttext vector support based on `https://github.com/babylonhealth/fastText_multilingual/`.

Transformer models used:

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

Hyperparameters:

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

A.2 Additional Results

- IT-ES: total: 3666, not in WordNet (WN): 1923
- IT-FR: total: 2172, not in WN: 918

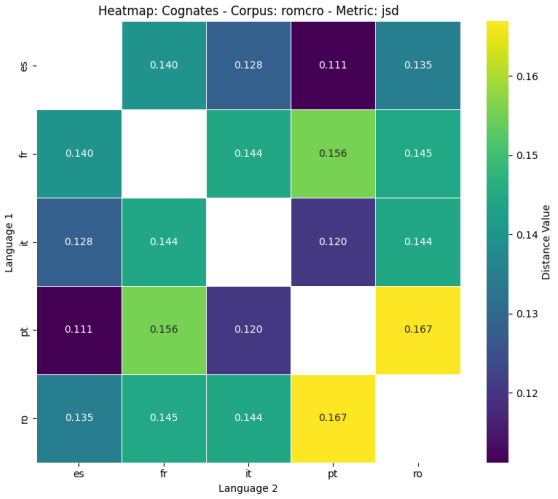


Figure 8: Semantic divergence between cognates in Romance languages based on contextual embeddings trained on RomCro (using JSD)

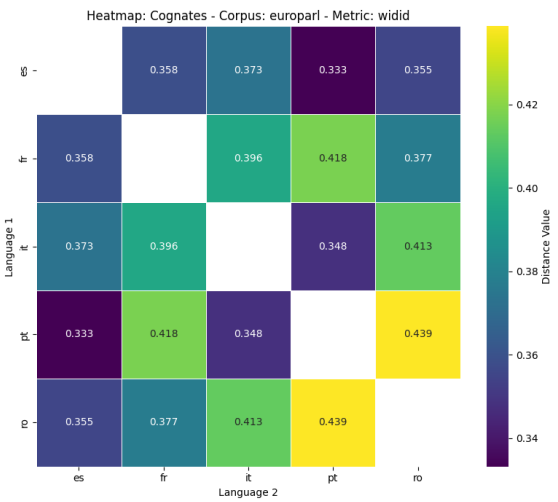


Figure 9: Semantic divergence between cognates in Romance languages based on contextual embeddings trained on Europarl (using widid)

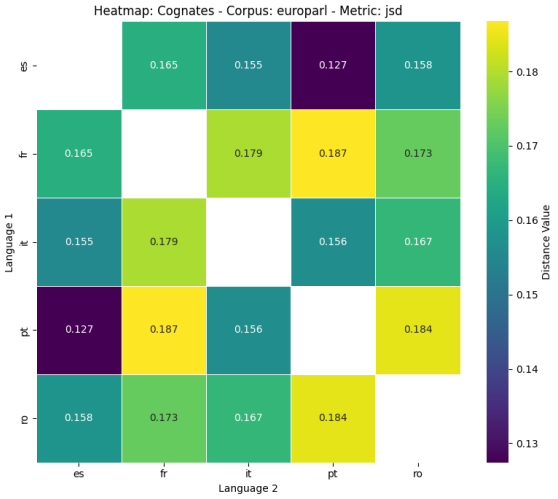


Figure 10: Semantic divergence between cognates in Romance languages based on contextual embeddings trained on Europarl (using JSD)

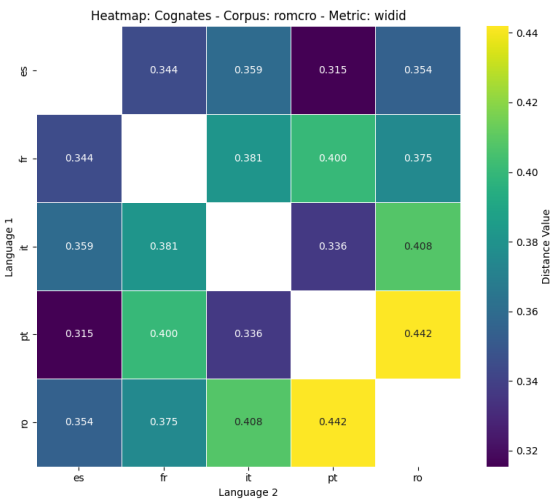


Figure 11: Semantic divergence between cognates in Romance languages based on contextual embeddings trained on RomCro (using Widid)

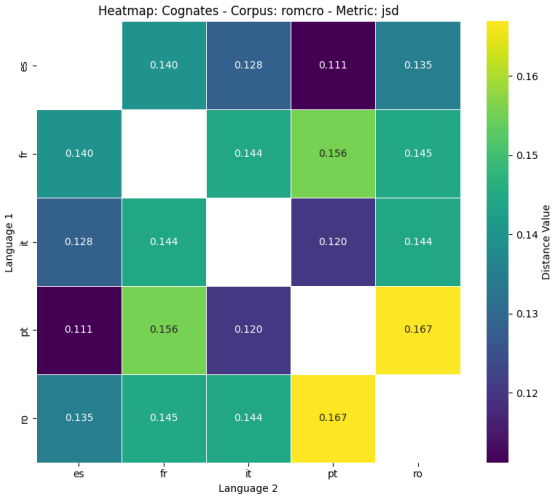


Figure 12: Semantic divergence between cognates in Romance languages based on contextual embeddings trained on RomCro (using JSD)

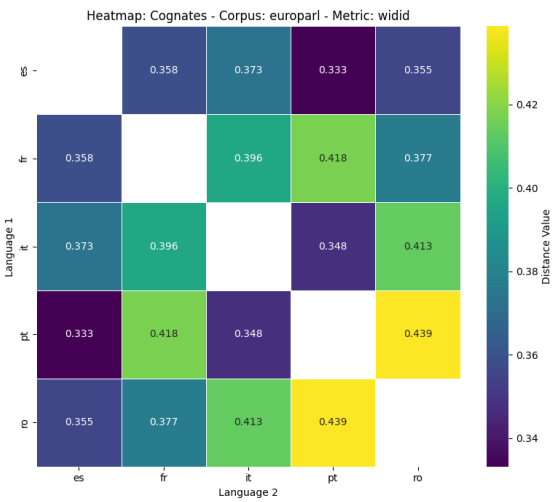


Figure 13: Semantic divergence between cognates in Romance languages based on contextual embeddings trained on Europarl (using widid)

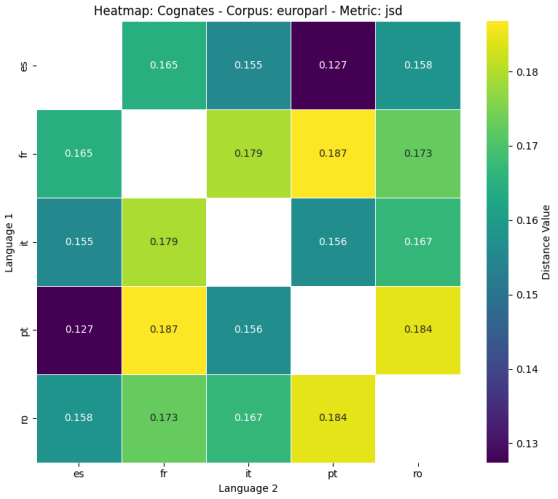


Figure 14: Semantic divergence between cognates in Romance languages based on contextual embeddings trained on Europarl (using JSD)

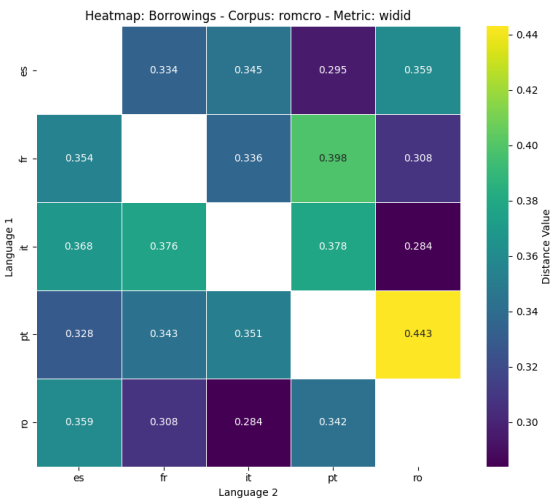


Figure 15: Semantic divergence between borrowings in Romance languages based on contextual embeddings trained on RomCro (using Widid)

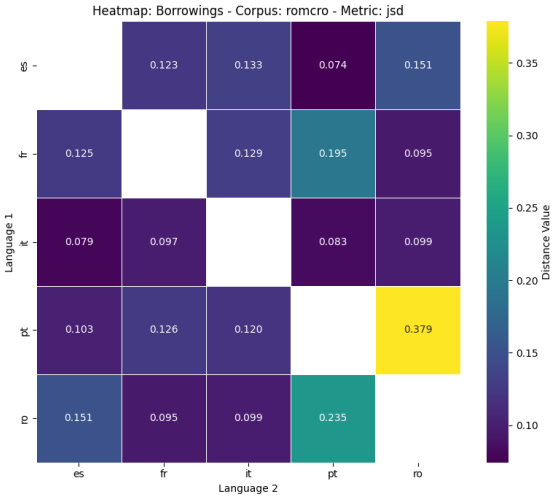


Figure 16: Semantic divergence between borrowings in Romance languages based on contextual embeddings trained on RomCro (using JSD)

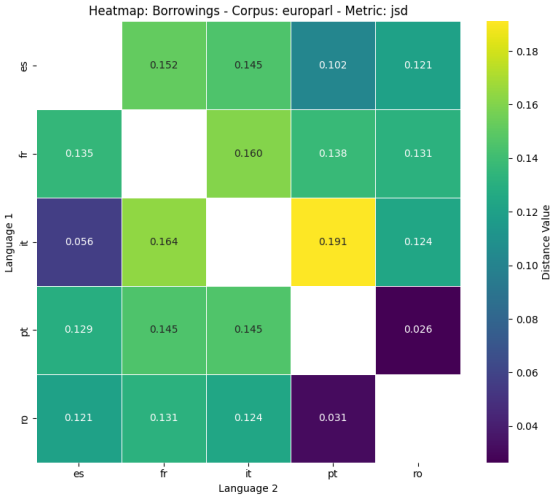


Figure 18: Semantic divergence between borrowings in Romance languages based on contextual embeddings trained on Europarl (using JSD)

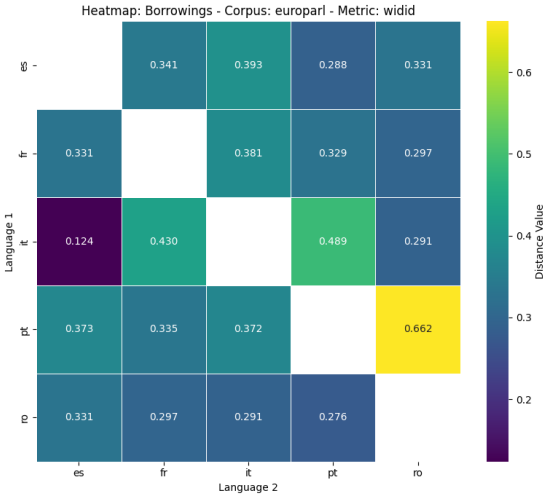


Figure 17: Semantic divergence between borrowings in Romance languages based on contextual embeddings trained on Europarl (using widid)

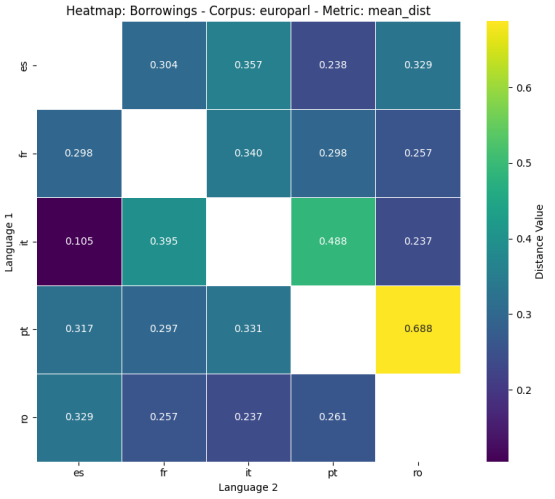


Figure 19: Semantic divergence between borrowings in Romance languages based on contextual embeddings trained on Europarl (using mean-dist)

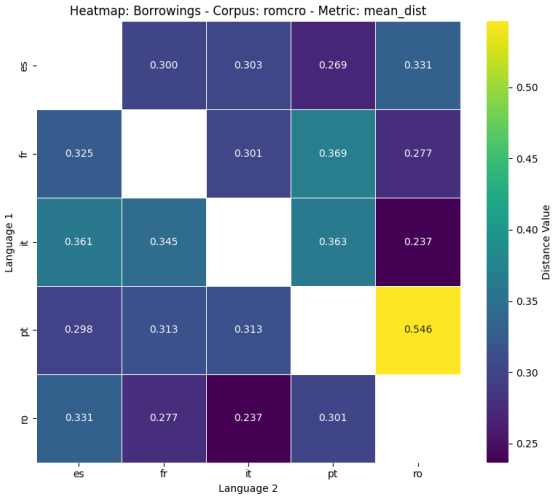


Figure 20: Semantic divergence between borrowings in Romance languages based on contextual embeddings trained on RomCro (using mean-dist)

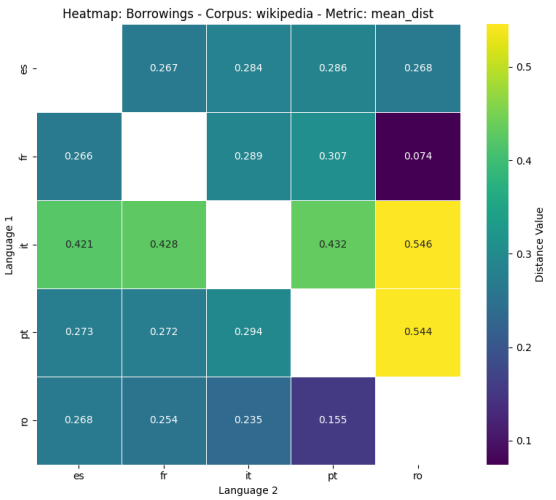


Figure 21: Semantic divergence between borrowings in Romance languages based on contextual embeddings trained on Wikipedia (using mean-dist)

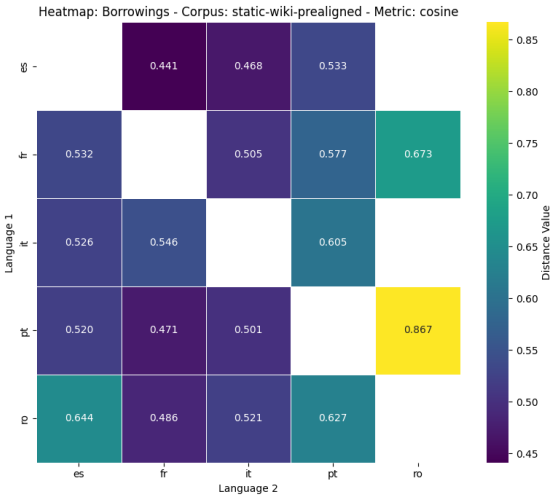


Figure 22: Semantic divergence between borrowings in Romance languages based on static embeddings trained on Wikipedia

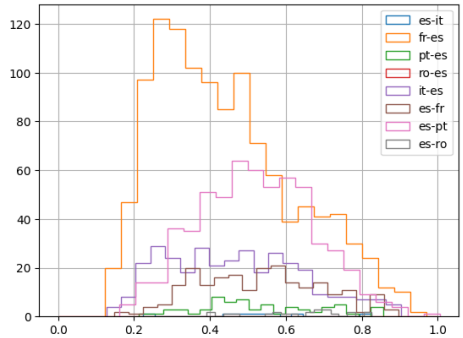


Figure 23: Borrowings distances distribution for Spanish.

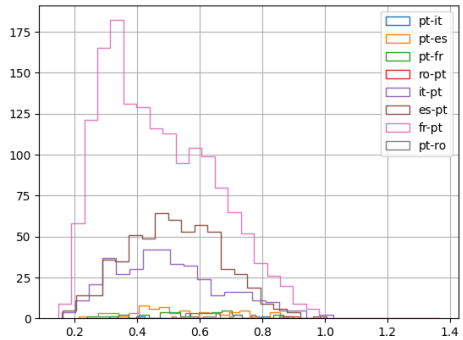


Figure 24: Borrowings distances distribution for Portuguese.

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

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

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