Cross-lingual Lexical Semantic Change in Romance Languages

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

Contributions

We perform a comprehensive analysis of lexical semantic change in the five main Romance languages (Romanian, Italian, Spanish, French and Portuguese), based on related words including both cognate words and borrowings.

Research Questions:

- **RQ1**: Can synchronic semantic divergence reflect historical change?
- **RQ2**: To what extent have Romance languages diverged semantically in terms of lexical semantic shift in cognate and borrowing pairs?

Dataset & Corpora

- **RoBoCoP + ProtoRom** databases
  - **19,222** cognate sets
  - **46,490** borrowing pairs
  - **5** Romance languages (es, fr, it, pt, ro)
- **Corpora** :
  - Wikipedia (non-parallel)
  - Europarl (parallel, European Parliament proceedings)
  - RomCro 2.0 (parallel, literary texts)

Methodology

1. Word Meaning Representations

- **Multiling. Aligned Contextual Embeddings**
  - sBERT
  - computed on occurrences from corpora
- **Multiling. Aligned Static Embeddings**
  - FastText

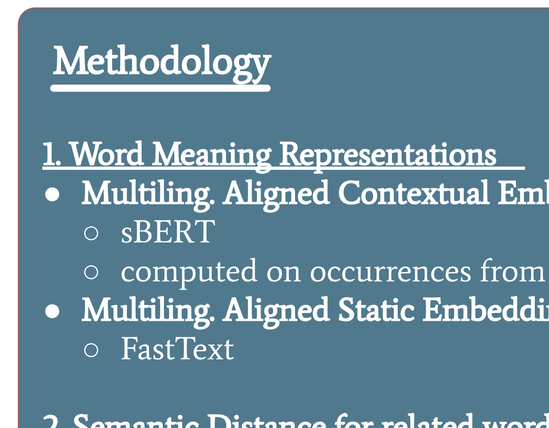
2. Semantic Distance for related word pairs

- **cosine** for static embeddings
- **avg. cosine, Jensen-Shannon Divergence, WiDiD** for contextual embeddings

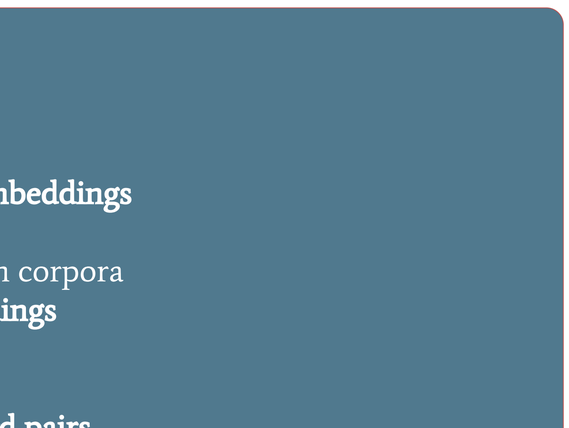
3. Language-level semantic divergence

- pair-level distances averaged across cognate/borrowing pairs to obtain a global semantic divergence measure for each language pair
- we separately measure semantic shifts for words with different parts of speech, according to Open multilingual WordNet.

Results



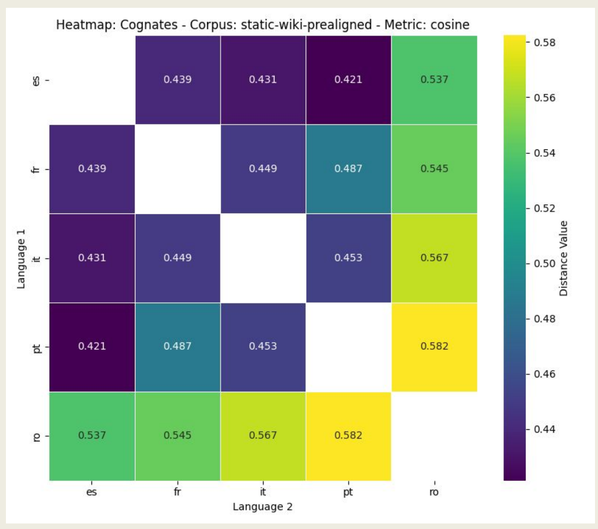
Cognates distribution (Wikipedia)

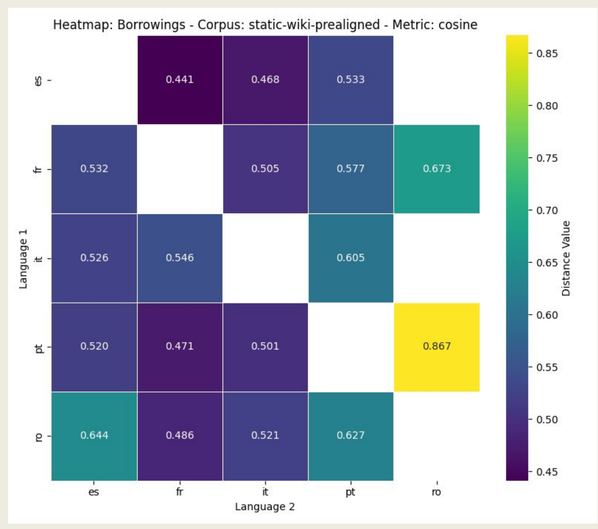


Borrowings distribution (Wikipedia)

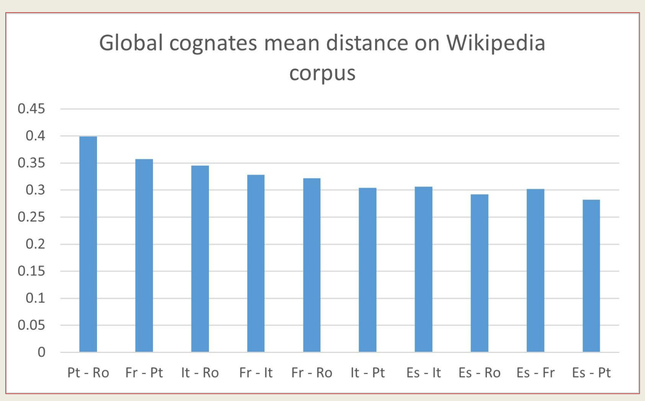
- There is a slight multimodality in the distribution of Spanish and Portuguese borrowings, with two main peaks in the distributions around distances of 0.3 and 0.5: for both the lower peak corresponds to borrowings from French, and the higher peak to distances with each other.
- Global distances vary by domain: 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 in literature (the RomCro corpus) shows a greater variety.

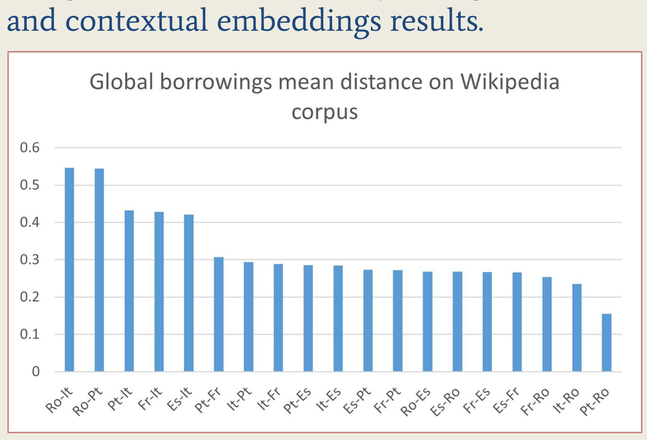
Results





- Consistently, the most divergent pair is Pt-Ro and the least divergent Es-Pt.
- Romanian is most divergent from all other Romance languages, which 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
- The Romanian-Portuguese divergence across corpora, 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.
- On average, the pairs between Spanish and any other language show the lowest degrees of semantic differentiation, while pairs containing French are moderately divergent.
- Few differences in results between the static and contextual embeddings results.





- For borrowings, divergence is more nuanced, and patterns applicable to all languages cannot be detected as readily.
- POS analysis shows for cognates verbs are most stable overall, for most languages except for Romanian. For borrowings, adjectives are most semantically stable across languages.

Manual analysis

Conceptual areas where low divergences occurred are generally generic words that designate scientific fields or general areas of activity ('mathematics', 'astronomy', 'agriculture', 'medicine', etc.), univocal verbs (that have not developed figurative meanings (to write, to kill), technical verbs ('to transport', 'to torture', 'to excommunicate'). \

High divergence scores occur for 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').

Future Work

- Additional spoken corpora might be a useful complement
- Diachronic corpora integration
- Polysemy-aware labeling
- Refining the contextual embedding representations by post-alignment of embedding spaces across languages
- Expanded language coverage

Acknowledgements

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