

Computational Tools for Measuring and Analyzing Similarities in Romance Languages (SiRoLa)

Liviu P. Dinu^{1,3}, Ana Uban^{1,3}, Simona Georgescu^{2,3}, Anca Dinu^{2,3}, Laurentiu Zoicas^{2,3} and Claudia Vlad^{2,3}

¹University of Bucharest, Faculty of Mathematics and Computer Science, Panduri 90, Bucharest, Romania

²University of Bucharest, Faculty of Foreign Languages and Literature, Panduri 90, Bucharest, Romania

³University of Bucharest, Research Center for Human Language Technologies, Panduri 90, Bucharest, Romania

Abstract

The SiRoLa project aims to shed light on the degree of similarity between languages through quantitative and computational methods, focusing on main Romance languages: Romanian, Italian, French, Spanish, Portuguese. In order to provide an in-depth examination of the different similarity parameters, we will consider four distinct levels of analysis – phonetic, lexical, semantic and syntactic –, with the following objectives: (1) compute and compare the phonological similarity of any pair of Romance languages, based on the syllables' distributions and stress; (2) measure the lexical similarity as string distances between words originated from the same etymon or borrowed from one language to another; (3) compute global semantic similarity between any pair of Romance languages, as weighted averages of distances between related word pairs, along with detecting false friends and measuring their semantic divergence; (4) compute string and tree similarity measure for any pair of languages.

Keywords

Languages similarities, Romance languages, degree of similarity, intelligibility.

1. Introduction

There are two fundamental challenges in historical linguistics: (1) how do languages change across time and space; (2) how are languages related to each other. To answer the first question, we proposed the CoToHiLi project (Computational Tools for Historical Linguistics, PCE 108/2021, <https://nlp.unibuc.ro/projects/cotohili.html>, 2021-2023), To find an answer to the second question, we propose the project SiRoLa (Computational Tools for Measuring and Analyzing Similarities in Romance Languages), which comes as a natural extension of the team's interests in historical linguistics [1, 2, 3].

The SiRoLa project aims to shed light on the degree of similarity between languages through quantitative and computational methods, focusing on pairs of languages from the Romance family, namely Romanian, Italian, French, Spanish, and Portuguese. Language similarity is one of the most challenging research problems addressed in historical and comparative linguistics, because of the difficulty of finding a unitary scientific method to quantify the degree of closeness between languages. In most cases, the similarity of natural languages is treated as a fairly vague notion [4], precisely because it is open to both academic scholars and the general public, giving rise to subjective intuitions about which languages are more similar to which others. While linguists mostly agree regarding the grouping of languages into families, the degree of similarity between two languages on a scale from 1 to 100 has yet to be established (cf. [5]), even when it comes to languages from the same family. In this context, we aim at analyzing language similarity at four distinct levels: phonetic, lexical, semantic and syntactic. As such, our project can provide scientific benefits both for Romance linguistics and for general and historical linguistics.

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.

✉ lldinu@fmi.unibuc.ro (L. P. Dinu); auban@fmi.unibuc.ro (A. Uban); simona.georgescu@lls.unibuc.ro (S. Georgescu); ancaddinu@gmail.com (A. Dinu); lzoicas@yahoo.fr (L. Zoicas); claudia.vlad@lls.unibuc.ro (C. Vlad)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

The reliable data provided by Romance languages ensure a strong empirical basis for our investigation. Thus, we aim: (1) to study the various types of relationships and degrees of connection within each pair of languages from the Romance core, and (2) to provide computational tools that could be successfully used in the study of other linguistic families. Firstly, although the Romance languages are closely related, the degrees of similarity between any two Romance languages vary from one author to another and their classification is also controversial. For example, [6] report two different results for the classification of Romanian within the Romance family, either marginal or more integrated within the group. Our approach would provide evidence either sustaining or invalidating the theory of lateral areas [7], enabling the distinction between inherited affinity and gained similarity, due to language contact.

Secondly, the computational tools that we create for the specific analysis of the Romance languages could play an important role in the very first research question addressed by historical linguistics, namely the grouping of languages in families. Even if today we have an extended perspective on language classification, with over 2000 language varieties easily accessible through multiple online resources [8], we still lack a complete diagram of the degree of relatedness between them, as not all phylogenetic networks have been accurately mapped [9, 10, 11].

Even in a well-researched linguistic area like the Indo-European family, linguists still argue over the degree of closeness between languages [12, 13], the variation in opinions affecting the chronologic and geographical features that are attributed to the archaic communities. Given the assumption that the higher the similarity degree between two languages, the closer they are, the measure of similarity proves to have direct implications in paleolinguistics, history and sociology, by recovering the structures of language family trees and by inferring information concerning the dispersals of given families as they diversified through prehistory [12].

Another field of linguistics that can scientifically benefit from the results of this project, based on the impact of language similarities, is that of language acquisition, as previously acknowledged [14]. The SiRoLa project will additionally allow to observe which parameters have a greater weight in facilitating language acquisition. From a technological point of view, our proposal could help improve the machine translation software, by establishing the correct lexical and semantic correspondences, mainly through discrimination between real similarities and false friends (cf. [15]). From a socio-economic point of view, our proposal can provide also real benefits in the field of intercomprehension, in the context of the increasingly close collaboration at a global level.

The issue of determining the degree of similarity between languages using computational methods has rarely been addressed by researchers until now. Most approaches lack a detailed comparative examination at the lexical and/or syntactic level, generally reflecting the application of mathematical formula on textual data, leading to confusing results, mainly because the notion of language similarity can be confounded with the notion of text similarity. [16] present four types of language similarity (typological, morphological, syntactic and lexical), insisting on the misleading direction of research that is limited to the surface similarity of words. [17] analyze several phonetic and lexical predictors, concluding that lexical similarity can be seen as a predictor of language intelligibility. Consequently, as an innovative element, we propose the delimitation of four levels of research:

1. the surface level, namely the phonetic physiognomy (written and oral);
2. the lexical correspondence between languages (words inherited from Latin or borrowed from Latin or another Romance language);
3. the degree of semantic proximity (the most difficult aspect to capture in a quantitative measurement, to the best of our knowledge never approached before);
4. the in-depth level of syntactic similarity.

For the phonetic similarity, rarely considered in the literature, we will start intriguing syllable behavior, according to which a small number of syllables cover a large surface of the language. The issue was addressed at the syllable level within the Romance languages through rank distance [18], confirming previous linguistic theories, while also adding a quantitative and rigorous approach, based

on the representative vocabularies of Latin, Romanian, Italian, Spanish, Catalan, French and Portuguese. We will start from the premises assessed here to order the syllables of the Romance languages, both graphically and phonologically, in rankings, which we will then compare using Rank distance (a metric introduced by the project leader) and other appropriate metrics, while also considering stressed syllables, for a more detailed analysis.

Regarding the lexical similarity, the present project will put together both borrowings and cognate pairs, exploiting the results obtained through the CoToHiLi project, namely RoBoCoP, the database of cognates and borrowings between any two Romance languages [19]. At the same time, given that not all pairs of words are alike (e.g. Ro. om ‘man’ is formally more distant from Es. hombre than Ro. lună ‘moon’ is from Es. luna), we will apply a measurement method to quantify the similarity of word forms, and we will quantify these results.

Moving forward, we will extend similarity to the semantic level, given that words naturally tend to change their meaning over time [20], at various degrees: e.g., Ro. Țară ‘country’ has slightly diverged from its Es. cognate tierra ‘land’, while between Ro. lume ‘world’ and Es. lumbre ‘light, fire’ we notice a strong divergence, making them unrecognizable to the average speaker. To measure semantic similarity, we will determine the degree of meaning divergence in all cognate pairs, for all pairs of Romance languages, and, through an appropriate mathematical formula, we will combine this divergence with the degree of closeness of the forms, determined previously.

Syntactic similarity was investigated in previous research by using a computational method on a large dataset of parallel sentences in 16 languages from 5 distinct families, resulting in a map of syntactic similarities between the European natural languages [21]. Syntactic similarity has been analyzed especially at the level of SVO order, few other studies going beyond this aspect (cf. [22]). We propose two approaches: (1) automatically labeling parallel corpora of Romance languages with POS tags, and, via appropriate metrics, measuring the similarities between strings tagged with POS and (2) obtaining the associated parse trees, and then measuring their similarities. Another novelty consists in the data we will use, previously limited to the online dictionaries (mostly Wiktionary).

In this project the quality and reliability of the data will be significantly increased due to the RoBoCoP database, carried out during the project CoToHiLi [19]. By providing all the Romance cognate pairs identified between any two of the five main Romance languages we are dealing with, RoBoCoP enables an exhaustive perspective on the lexical relationship between the Romance languages. Moreover, we will include in our analysis several public corpora for the five Romance languages, going beyond the graphic level and focusing on the “language in use” perspective, based on the distribution of authentic language data. The project aims to address in-depth themes that would provide a significant contribution to comparative linguistics, and whose results would represent a real breakthrough both in the field of computational linguistics and in that of historical linguistics in general.

2. Objectives and methodology

Our main goal is to establish the degree of similarity between any two languages from the Romance core (Romanian, Italian, French, Spanish, Portuguese) on four different linguistic levels: phonologic, lexical, semantic and syntactic, corresponding to four major objectives. Below, we explain each.

Objective 1: Measuring and analyzing the phonological similarity of the Romance languages

The first objective of the project is to analyze and measure the phonological similarity of the Romance languages. Our method for measuring phonological similarity relies on the distributions of syllables in each language and the distances between them, considering both their graphic form and their phonetic transcriptions. To meet this objective, our strategy involves using text corpora in each of the considered languages, from which we will first extract the individual tokens, then syllabify each token, both at graphic and at phonetic level. In a separate analysis, we will further include in our approach the position of the stressed syllable, quantifying the differences between the Romance languages in terms of stress patterns. The main motivation for this strategy based on syllable rankings is the unusual distributional behavior of the syllables, which results in a relatively small number of distinct syllables covering a

large part of the corpus (the first 50 syllables cover around 50% of the number of all the syllables of the used vocabulary, the first 200 syllables cover around 80%, after this number, the percentage increases slowly). Rank Distance [18] is well suited for this task, since it accounts for the fact that the items on the first positions are more important than the ones on the last positions of a ranking.

Objective 2: Measuring and analyzing the lexical similarity of the Romance languages Our second objective is to analyze and measure the lexical similarity between Romance languages. Our strategy thus involves identifying related words (cognates and borrowings) for each language pair [23], measuring the lexical similarity as string distances between any pair of related words, and finally adding and normalizing this measure.

Objective 3: Measuring and analyzing the semantic similarity of the Romance languages The third objective is to measure the semantic similarity between Romance languages. The main motivation is that some related words (cognates or borrowings) might change their meaning to a certain extent (e.g., Ro. *stânga* ‘left’ has semantically diverged from its It. cognate *stanco* ‘tired’) [24]. For each pair of related words from, we will use vectorial representations obtained automatically by training deep learning models on large corpora (word embeddings) to determine whether the two words still have the same meaning in the two languages, and to measure how much the meanings have diverged. Based on vectorial distance metrics in the produced multilingual embedding spaces, we will design an algorithm for determining whether two related words still share the same meaning or not. We will complement our previous work in which we used static word embeddings trained on comparable corpora in each of the languages with contextual embeddings. We intend to use contextual embeddings based on pre-trained large language models such as multilingual-BERT2 to compare word meanings in context. Contextual embeddings are obtained by training large pre-trained language models based on the transformer architecture, which have revolutionized the field of natural language processing. In comparison to static embeddings, contextual embeddings have the advantage of also being able to model polysemy (where each context has a separate vectorial representation, and a unique representation of a given word can be obtained by averaging the representations of the contexts they occur in), embeddings of words in any languages supported by the model are directly comparable. Measures of distance in these vectorial spaces can be directly used as measures of semantic similarity between related words.

Objective 4: Measuring and analyzing the syntactic similarity of the Romance languages Our fourth objective is to assess the syntactic similarity between languages. We will preprocess the corpora and extract the parallel sentences. As a result, we will obtain a list of parallel sentences for all five languages. Then, we will use a state-of-the-art POS tagger, apply them to the parallel obtained sentences, and output lists of parallel POS-tagged sentences for all investigated languages. Then, for any two POS-tagged strings obtained, we will compute a string similarity measure that reflects the syntactic similarity and normalize the obtained values. We will add all these values for all parallel sentences in each language pair, to obtain a measure of syntactic similarity of any two Romance languages. We will identify state-of-the-art dependency and constituency parsers for the five languages and use them on the lists of parallel sentences obtained. The result will consist of lists of parallel parse trees for all Romance language pairs. For any pair of parse trees we will compute a normalized tree similarity measure that reflects the syntactic similarity at tree level and add all these values for all parallel trees in each language pair.

3. Preliminary results

So far, we have focused on the first three objectives. Using two parallel corpora (Europarl and RomCro) and one non-parallel corpus (Wikipedia), we managed to identify and extract databases containing all the syllables in the Romance languages under investigation, ordered by frequency, while also extracting the lists of stressed syllables per each language. Based on these, we compared the five languages via rank distance, obtaining the degrees of similarity at the phonetic level [25].

For the second objective, we extracted the lists of related words (cognates and borrowings) from the three corpora, based on the RoBoCoP database [19], and applied various similarity distances to

determine lexical and semantic similarity. Thus, we were able to determine semantic similarity between any pair of Romance languages, and additionally proposed [20] an algorithm for detecting and correcting false friends and show it obtains good results for the Romance languages studied.

Since some words change their meaning, for the third objective we introduced a new intelligibility index, DLI, which considers both lexical distance and semantic shift between two related words, and extended it at the language level to obtain cross-lingual intelligibility scores for all Romance languages [26]. It is worth mentioning that this approach allowed us to identify some asymmetric similarities, revealing, among other things, that Romanian is the least intelligible Romance language when compared to the others; however, all the other Romance languages are considerably more intelligible to Romanian speakers than vice versa. We also looked more closely at the intelligibility of Spanish and Portuguese varieties in Europe compared to South America [27].

4. Project's details

Project title: Computational Tools for Measuring and Analyzing Similarities in Romance Languages (SiRoLa), 2025-2027 (<https://nlp.unibuc.ro/projects/sirola.html>)

Author's name: Liviu P. Dinu (PI), Ana Uban, Simona Georgescu, Anca Dinu, Laurentiu Zoicas, Claudia Vlad

Affiliation: University of Bucharest, Faculty of Mathematics and Computer Science (LD& AU) & Faculty of Foreign Languages and Literature (SG, AD, LZ&CV)

Contact information: Liviu P Dinu, liviu.p.dinu@gmail.com, ldinu@fmi.unibuc.ro Financial institutions: Ministry of Research, Innovation and Digitization, CNCS-UEFISCDI, project SIROLA, number PN-IV-P1- PCE-2023-1701, within PNCDI IV, Exploratory Research Projects, Bucharest, Romania

Acknowledgements

This research was supported by the Ministry of Education and Research, CNCS-UEFISCDI, project SIROLA, number PN-IV-P1- PCE-20231701, within PNCDI IV, and by InstRead: Research Instruments for the Text Complexity, Simplification and Readability Assessment CNCS - UEFISCDI project number PN-IVP2-2.1-TE-2023-2007.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

- [1] A. M. Ciobanu, L. P. Dinu, An etymological approach to cross-language orthographic similarity. application on romanian, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1047–1058.
- [2] A. M. Ciobanu, L. P. Dinu, Automatic identification and production of related words for historical linguistics, Computational Linguistics 45 (2020) 667–704.
- [3] L. P. Dinu, A. S. Uban, A. M. Cristea, I.-B. Iordache, T.-G. Marchitan, S. Georgescu, L. Zoicas, 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, 2024, pp. 6314–6326.
- [4] A. M. McMahon, R. McMahon, Linguistics, genetics and archaeology: internal and external evidence in the amerind controversy, Transactions of the Philological Society 93 (1995) 125–225.
- [5] L. Lebart, M. Rajman, Computing similarity, Handbook of Natural Language Processing (2000) 477–505.

- [6] A. McMahon, R. McMahon, Finding families: Quantitative methods in language classification, *Transactions of the Philological Society* 101 (2003) 7–55.
- [7] M. Bartoli, *Introduzione alla neolinguistica:(principi, scopi, metodi)* (1925).
- [8] J.-M. List, R. Forkel, S. J. Greenhill, C. Rzymiski, J. Englisch, R. D. Gray, Lexibank, a public repository of standardized wordlists with computed phonological and lexical features, *Scientific data* 9 (2022) 316.
- [9] P. Heggarty, Beyond lexicostatistics: How to get more out of ‘word list’ comparisons, *Diachronica* 27 (2010) 301–324.
- [10] P. Heggarty, Prehistory by bayesian phylogenetics? the state of the art on indo-european origins, *Antiquity* 88 (2014) 566–577.
- [11] J.-M. List, S. J. Greenhill, R. D. Gray, The potential of automatic word comparison for historical linguistics, *PloS one* 12 (2017) e0170046.
- [12] P. Heggarty, Cognacy databases and phylogenetic research on indo-european, *Annual Review of Linguistics* 7 (2021) 371–394.
- [13] J. Mallory, D. Adams, *The Oxford Introduction to Proto-Indo-European and the Proto-Indo-European World* (Oxford Linguistics), Oxford University Press Oxford, 2006.
- [14] B. VanPatten, A. G. Benati, *Key terms in second language acquisition* (2015).
- [15] A. Uban, A. M. Ciobanu, L. P. Dinu, Studying laws of semantic divergence across languages using cognate sets, in: *Proceedings of the 1st international workshop on computational approaches to historical language change*, 2019, pp. 161–166.
- [16] P. Homola, V. Kubon, A structural similarity measure, in: *Proceedings of the Workshop on Linguistic Distances*, 2006, pp. 91–99.
- [17] C. Gooskens, W. Heeringa, K. Beijering, Phonetic and lexical predictors of intelligibility, *International journal of humanities and arts computing* 2 (2008) 63–81.
- [18] A. Dinu, L. P. Dinu, On the syllabic similarities of romance languages, in: *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, 2005, pp. 785–788.
- [19] L. P. Dinu, A. Uban, A. Cristea, A. Dinu, I.-B. Iordache, S. Georgescu, L. Zoicas, 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*, 2023, pp. 7610–7629.
- [20] A. S. Uban, L. P. Dinu, I.-B. Iordache, S. Georgescu, C. Vlad, 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*, 2025, pp. 15309–15323.
- [21] A. M. Ciobanu, L. P. Dinu, A. Sgarro, Towards a map of the syntactic similarity of languages, in: *International Conference on Computational Linguistics and Intelligent Text Processing*, Springer, 2017, pp. 576–590.
- [22] G. Longobardi, S. Ghirotto, C. Guardiano, F. Tassi, A. Benazzo, A. Ceolin, G. Barbujani, Across language families: genome diversity mirrors linguistic variation within europe, *American journal of physical anthropology* 157 (2015) 630–640.
- [23] L. P. Dinu, A. S. Uban, I.-B. Iordache, A. M. Cristea, S. Georgescu, L. Zoicas, Pater incertus? there is a solution: Automatic discrimination between cognates and borrowings for romance languages, in: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 12657–12667.
- [24] A.-S. Uban, A. Cristea, L. P. Dinu, A computational approach to measuring the semantic divergence of cognates, in: *International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2019)*, Springer, 2019, pp. 96–108.
- [25] L. P. Dinu, A. S. Uban, I.-B. Iordache, S. Georgescu, C. Vlad, It runs in the family: Syllabic similarity in the romance languages, 2026. Submitted.
- [26] L. P. Dinu, A. S. Uban, B. Iordache, A. Dinu, S. Georgescu, Measuring cross-language intelligibility between romance languages with computational tools, *arXiv preprint arXiv:2602.07447* (2026).
- [27] L. P. Dinu, A. S. Uban, T.-G. Marchitan, I.-B. Iordache, S. Georgescu, On the intelligibility of romance language varieties: Spanish and portuguese in europe and america, in: *Proceedings of*

the 13th Workshop on NLP for Similar Languages, Varieties and Dialects, 2026, pp. 139–144.