On the Intelligibility of Romance Language Varieties: Spanish and Portuguese in Europe and America

Liviu P. Dinu[♠,♡] **Ana Sabina Uban**[♠,♡]
Teodor-George Marchitan[♠,♡] **Ioan-Bogdan Iordache**[♠,♡] **Simona Georgescu**[♣,♡]
University of Bucharest, [♠] Faculty of Mathematics and Computer Science,
[♣]Faculty of Foreign Languages and Literatures, [♡]HLT Research Center
{ldinu, auban}@fmi.unibuc.ro,
{teodor.marchitan,ioan.iordache}@s.unibuc.ro,
simona.georgescu@lls.unibuc.ro

Abstract

Mutual intelligibility within language families presents a significant challenge for multilingual NLP, particularly due to the prevalence of dialectal variation and asymmetric comprehension. In this paper, we present a corpus-based computational analysis to quantify linguistic proximity across Romance language variants, with a focus on major Spanish (Argentine, Chilean and European) and Portuguese (Brazilian and European) varieties and the other main Romance languages (Italian, French, Romanian). We apply a computational metric of lexical intelligibility based on surface and semantic similarity of related words to measure mutual intelligibility for the five main Romance languages in relation to the Spanish and Portuguese varieties studied.

1 Introduction

The study of mutual intelligibility (the degree to which speakers of different, yet related, languages can understand one another, cf. Gooskens and Van Heuven (2021) is fundamental to both historical linguistics and practical applications in Natural Language Processing (NLP). This phenomenon is particularly pronounced within the Romance language family, a group united by common Latin ancestry but distinguished by centuries of independent evolution.

In this work, we present a systematic, corpus-based investigation into mutual intelligibility across major Romance language variants, aiming to quantify their linguistic proximity and understand the factors driving comprehension success or failure, from a lexical perspective. Specifically, we focus on different Spanish (Argentine, Chilean and European) and Portuguese (Brazilian and European) varieties in relation to the other languages in the main Romance language family (Spanish, Italian, French, Romanian and Portuguese).

We rely on a complete database of related words in the Romance languages (Dinu et al., 2023), as well as on a collection of corpora for each of these languages and varieties. By analyzing shared lexical features based on their vocabularies and corpus usage, we attempt to answer the following research questions:

1. Can computational metrics accurately model the known asymmetry of intelligibility (e.g. Portuguese speakers generally understanding Spanish better than the reverse)?

2. How do regional variations in Spanish and Portuguese impact intelligibility with non-Iberian Romance languages compared to their European variants?

We propose to explore the issue of language mutual intelligibility, taking as our basis the Romance lexicon present in the five main Romance languages in terms of number of speakers (Spanish, Portuguese, French, Italian, Romanian). We focus on vocabulary because it represents the surface of language, the first layer we come into contact with when we approach a new language (Gooskens and Van Heuven, 2021). Later, when we hear or read a language that is genetically close to our mother tongue or another language we know, before we comprehend the morphology and syntax, we understand the vocabulary. Several studies have indicated that, overall, morpho-syntactic differences contribute less to intelligibility than lexical and phonological differences (Hilton et al., 2013; Gooskens and Schneider, 2016). Words are the interface that either makes the language seem familiar and therefore intelligible, or raises a wall that is impossible to cross without prior training (Gooskens, 2024).

By addressing the issue of mutual intelligibility among Romance languages, we implicitly explore the issue of language similarity. Language

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 139–144
March 29, 2026. ©2026 Association for Computational Linguistics

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 (McMahon and McMahon, 1995), precisely because it is used by both academic scholars and the general public, giving rise to more or less justified intuitions about which languages are more similar to which others. In many situations, these intuitions are simply based on the very subjective opinions of either linguists or non-linguists[1].

When it comes to the degree of relatedness between languages, we believe that the best unit of measurement is the ability of a native speaker of one language to understand other languages in the same family, without having any prior knowledge of them (see also Gooskens (2013)). That is why our approach is based on measuring the degree of intelligibility between any two pairs of Romance languages.

We chose to address intelligibility at the level of linguistic varieties, not just standard languages, in order to track any asymmetry in the ability of speakers of other Romance languages to understand varieties in relation to the standard language. Personal experience has often allowed us to observe a higher level of understanding of South American Spanish by Italian or Romanian speakers, but a lower level when it comes to French or Portuguese speakers. Therefore, the study of the Chilean and Argentine varieties of Spanish tests this empirical observation using a scientific method. In addition, we add the Brazilian variety of Portuguese to provide a basis for comparison and to be able to draw more general conclusions about the relationship between the level of intelligibility within and beyond the borders of Europe.

Therefore, the findings of this paper offer a significant contribution to the field of computational sociolinguistics. The resulting quantitative models of cross-lingual and intra-lingual intelligibility can be directly applied to enhance several multilingual NLP tasks, including low-resource cross-lingual transfer learning, dialect-aware machine translation, and the design of more robust language processing tools for global Romance-speaking communities.

[1]An anecdote related by McMahon (1994) shows how subjective the perception on similarity is: when asked how close Hungarian and Finnish are, a connoisseur of the two languages answered "oh, very", but when asked whether "like Italian and Spanish", the answer was "not that close, more like English and Persian".

2 Methodology

In order to obtain intelligibility scores across the Romance languages and the varieties studied, we use a metric of lexical intelligibility based on usage of words with common etymology (cognates and loanwords), as they occur in language use, considering that intelligibility of a language for a non-speaker relies on related words which are similar in form and meaning.

2.1 Dictionaries and Corpora

Since we start from the premise that intercomprehension is based on the ability of the average speaker to identify words in the flow of speech, we ground our analysis on authentic acts of speech, not simply lexicographic works. Therefore, we analyze two public corpora for the five Romance languages, RomCro - containing literary texts in different languages, translated in Romance languages and Croatian (Mikelenić et al., 2024), and EuroParl - focusing on proceedings of the European Parliament (Koehn, 2005). We extract related word frequencies in corpora used in our metrics based on the two parallel corpora. We employ a highly diverse collection of five publicly available corpora, spanning both the Spanish and Portuguese languages across distinct regional varieties and modalities. The Spanish-language component consists of three corpora. We leveraged two written reference corpora from Latin America (Argentina Corpus (Marcos-Marín and Zumárraga, 1992) and Chile Corpus (Marcos-Marín and Evans Espiñeira, 1991) each with around $2,000,000$ words) to cover a broad range of formal written registers. These include substantial representations of Journalistic (Periodísticos), Humanistic, Scientific, Technical, Literary, Juridical and Scholastic (Escolares) texts, with detailed frequency breakdowns guiding the selection criteria. For Spanish oral data, we utilized the CORLEC (Corpus Oral de Referencia de la Lengua Española Contemporánea) (Marcos-Marín, 2000s), which provides transcribed spontaneous and formal speech across categories such as Conversation, Political, Educational, Scientific and various journalistic formats (e.g. debate, news, interviews). The Portuguese-language component includes both Brazilian and European varieties. The C-ORAL Brasil 2 (Raso et al., 2012–2015)

is a corpus of spontaneous Brazilian Portuguese speech, structured into three principal subcorpora: Formal in Natural Context (e.g., business, teaching, political speech), Media (e.g., interviews, talk shows), and Telephonic (public and private interactions). Lastly, the Corpus Português Fundamental (CLUL (Centro de Linguística da Universidade de Lisboa), 1984) provides data for European Portuguese, featuring a frequency corpus of spontaneous spoken language (approx. $700,000$ words) and an availability corpus of thematic vocabulary (approx. $481,800$ words) derived from domain-specific areas. This broad selection ensures robust coverage of stylistic, geographic and domain variation necessary for comprehensive analysis.

We perform our analysis on related word pairs extracted from the most comprehensive database of related words in cognate languages up to date, sourced from etymological dictionaries and manually curated, RoBoCoP (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 language pairs (Dinu et al., 2024b).

In order to identify occurrences in text for our pairs of related words (i.e. cognates and borrowings) for a given language pair, we process parallel sentence examples from the employed datasets as follows: we tokenize the sentences using spaCy (Honnibal et al., 2020), remove stop-words, and match each token with a corresponding instance from RoBoCoP, if possible. In order to account for inflections of the dictionary form from RoBoCoP perform normalization including accent removal and stemming, using a Snowball stemmer (Porter, 2001). For each sentence pair we count how many of the words from one sentence are words related to the other language. We also count, for a given related words pair, how many times that pair appears in aligned examples, in other words: how many times the related words pair corresponds to a proper translation. Table 1 shows these counts.

2.2 Surface and Semantic Similarity

The intelligibility computation is twofold: we combine the orthographic similarity of word pairs across languages with their semantic similarity.

For the former, we measure string similarity using the normalized Levenshtein distance (Levenshtein, 1966) on the orthographic (after removing accents) representation. Thus we obtain scores in the interval $[0, 1]$, where 1 means identical representations.

The latter (i.e. semantic similarity) is computed based on word representations trained on large corpora to represent meaning in context.

A metric of semantic similarity is computed using cosine similarity based on FastText word embedding vectors (Bojanowski et al., 2016). Since we are assessing similarity across languages, we need our embedding spaces to be aligned. For that reason, we use pretrained prealigned static embeddings, based on a large multilingual corpus (Wikipedia) (Joulin et al., 2018). Since even on a large corpus, such as the ones used to pretrain these embeddings, some of our related words extracted from RoBoCoP may not be represented, but an inflected form may be available, we find for each unrepresented word its closest match in form (via stemming and edit distance) that is present in the pretrained embeddings and we use that vector as our canonical representation.

2.3 Lexical Intelligibility Index

We rely on the D_{LI} metric for computing cross-lingual intelligibility automatically, introduced in Dinu et al. (2026). It is computed via the following formula:

$$D_{LI} = \frac{S_s S_L (2 - S_s - S_L)}{1 - S_s S_L} \quad (1)$$

where S_L is the formal lexical similarity between two words (i.e. 1 minus the normalized Levenshtein distance, computed on the orthographic representations), and S_S is the semantic similarity between two words (i.e. 1 minus the cosine distance between static word embeddings or the average cosine distance computed on contextual embedding clusters).

We further obtain aggregate intelligibility scores at language pair level. For each language pair (A, B), where A is the speaker language and B is the listener language, given one of our employed corpora, we pick each sentence in language A and compute an intelligibility score with respect to B as the weighted average of the intelligibility indices of all the words in the sentence that are related to language B (i.e. the sum of these indices divided by the total number of words in the sentence, excluding stop-words). We compute a corpus-level

Table 1: Corpus Statistics: total sentences, words, unique words.

Dataset	Sentences	Words	Unique	Unique (no stop)
C-ORAL Brasil 2	9,920	182,457	13,646	13,493
CORLEC Spanish (European)	83,512	1,005,158	36,881	36,590
CORLEC Spanish Chile	44,471	920,129	47,935	47,681
CORLEC Spanish Argentina	115,957	2,239,468	61,711	61,510
Português Fundamental	8,890	97,659	8,603	8,432
TOTAL	**262,750**	**4,444,871**	**168,776**	**167,706**

Table 2: Results for Spanish variants

Corpus	es-it	es-fr	es-pt	es-ro
Argentina	0.193	0.184	0.227	0.186
Chile	0.224	0.216	0.261	0.215
Spain (oral)	0.229	0.209	0.291	0.181

Table 3: Results for Portuguese variants

Corpus	pt-it	pt-es	pt-fr	pt-ro
Brasil (oral)	0.262	0.252	0.136	0.101
Portugal	0.280	0.297	0.170	0.121

aggregate scores by combining all of the sentence-level scores (equivalent to combining all of the sentences into a singular text). In this way, the final overall intelligibility score between two languages will be affected both by the usage of related words in corpora in context as well as by their mutual individual intelligibility.

3 Results Analysis

The results in Table 2 only partially confirm our initial assumption: the highest level of intelligibility for a Romanian listening to the three varieties of Spanish will be in relation to the Chilean variety, followed by the Argentine variety, with European Spanish coming in third place. In contrast, an Italian will understand peninsular Spanish better, followed closely by the Chilean variety and, lastly, the Argentine variety. For French speakers, the order will be the same as for Romanian speakers, only with a higher level of comprehension of peninsular Spanish. For Portuguese speakers, peninsular Spanish will be by far the most easily understood, which should come as no surprise given the geographical proximity and constant linguistic contact over the centuries (see also Gooskens (2024)). If we look at the results for Portuguese vs. Brazilian (table 3), we see in all language pairs the prevalence of peninsular Portuguese over the Brazilian variety.

There are two factors that can explain these differences in intelligibility. First, there is an inequality in the type of data: the corpus for peninsular Spanish is oral, while for the Chilean and Argentine varieties it consists of written texts. It is true that today diaphasic and diamesic differences are not as significant as they were a century ago, when access to formal education was limited; however, in oral communication there is a tendency to use familiar variants, with colloquial nuances, which are not found with the same frequency in writing. Therefore, a Romanian speaker who has no direct contact with Spanish and its evolution in everyday language will have greater difficulty understanding current speech. In addition, in a written text the style is more refined and the lexicon is not marked by sociolinguistic features, which makes it easier to achieve a much higher degree of intelligibility.

Secondly, one must take into account the specific characteristics of South American varieties, which are characterized by two trends that differ from those observed in European Spanish. On the one hand, the lexicon of the main corpus is somewhat more conservative than that of the peninsular variety, which is understandable from a historical perspective: the Spanish brought by colonists to South America is that of the 16th and 17th centuries, which, even though it has evolved, like any living language, retains features of classical Spanish that the language in Spain has lost; therefore, there is a greater chance of finding more Latinisms and more old meanings, which are also shared by the Romance language at the eastern end of Europe, Romanian. To give an example, in South American Spanish, the verb most commonly used for "to catch" is *prender*, a cognate of Romanian *prinde*, used with the same meaning (from Latin *prehendere* 'to catch'). In European Spanish, however,

a different verb, *pillar*, is increasingly preferred for this sense. It was borrowed from Italian and has no cognates in Romanian, making it potentially incomprehensible to a Romanian speaker.

The second trend is the much greater permeability to Anglicisms of South American varieties (also as a result of historical realities) compared to Peninsular Spanish, a permeability that makes them easier to understand for speakers of a language that adopts a large amount of vocabulary from English and who learn English from an early age, such as Romanians. For example, in South American varieties, speakers use the word *rancho* 'ranch' borrowed from En. *ranch*, which is no doubt more comprehensible for Romance speakers with a basic knowledge of English than *finca* or *hacienda*, its synonyms preferred in Peninsular Spanish.

In future work, including a similar analysis based on phonetic representations and taking into account regional pronunciations might be useful to confirm the differences in oral communication versus written text for intelligibility. It has been previously observed that identifying orthographic correspondences between languages can increase the intelligibility (Fischer et al., 2016), while pronunciation may pose a barrier to cross-linguistic intelligibility. We offer a concrete example: in standard European Spanish, the phonetic sequences /ce/ and /ci/, as well as /z/ in any position, are pronounced [θ], whereas in South America they are pronounced [s]. The [θ] sound is absent in the other Romance languages under consideration, which makes the South American pronunciation more comprehensible for speakers of these languages than European Spanish: for example, the word *zebra* pronounced [sébra] would be more intelligible for the average Romance language speaker than [θébra]. However, we can also encounter the opposite situation: a Romanian, Italian, French or Portuguese listener who hears the Spanish word *hacer* pronounced in a South American variety [asér] may interpret it as the sequence *a ser* (preposition *a* "to" + infinitive *ser* "be"), through a false analysis — provided they have at least minimal knowledge of Spanish. This South American pronunciation can be particularly misleading if the listener has previously been exposed to European Spanish, where the verb would be pronounced [aθér]. In the absence of any prior exposure, it is most likely that the word will not be understood when encountered in the flow of speech. However, when presented in written form, comprehension is more likely, even without prior knowledge. This may be partly due to speakers' ability to reconstruct meaning from context, even when elements are unintelligible in isolation — a capacity well documented in cloze tests (cf. Gooskens and van Heuven (2017); Gooskens (2024)). It may also result from the inherent orthographic similarity to the speaker's native-language verb, e.g. Ro. *face* / *facere*, whose Levenshtein distance from *hacer* is relatively small.

4 Conclusion

We proposed a study of the mutual intelligibility of Romance languages, with a focus on different varieties of Spanish and Portuguese. We apply a metric of lexical intelligibility based on surface and semantic similarity of related words to compute intelligibility between each of the five main Romance languages and European and American varieties of Spanish and Portuguese, based on a selection of corpora of Argentine, Chilean and European Spanish, and Brazilian and European Portuguese, respectively. We thus provide a corpus-grounded quantification of linguistic distance across these Romance variants, and open the way for further research into the study of language varieties from the perspective of lexical intelligibility.

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.

References

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

CLUL (Centro de Linguística da Universidade de Lisboa). 1984. Corpus Português Fundamental. Linguistic Resources, Centro de Linguística da Universidade de Lisboa.

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.

Liviu P Dinu, Ana Sabina Uban, Alina Maria Cristea, Anca Daniela Dinu, Ioan-Bogdan Iordache, Simona Georgescu, and Laurentiu Zoicas. 2023. Robocop: A comprehensive romance borrowing cognate package and benchmark for multilingual cognate identification. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Liviu P Dinu, Ana Sabina Uban, Bogdan Iordache, Anca Dinu, and Simona Georgescu. 2026. Measuring cross-language intelligibility between romance languages with computational tools. *Preprint*, arXiv:2602.07447.

Andrea Fischer, Klára Jágrová, Irina Stenger, Tania Avgustinova, Dietrich Klakow, and Roland Marti. 2016. Orthographic and morphological correspondences between related slavic languages as a base for modeling of mutual intelligibility. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4202–4209.

Charlotte Gooskens. 2013. Experimental methods for measuring intelligibility of closely related language varieties.

Charlotte Gooskens. 2024. *Mutual intelligibility between closely related languages*, volume 30. Walter de Gruyter GmbH & Co KG.

Charlotte Gooskens and Cindy Schneider. 2016. Testing mutual intelligibility between closely related languages in an oral society.

Charlotte Gooskens and Vincent J van Heuven. 2017. Measuring cross-linguistic intelligibility in the germanic, romance and slavic language groups. *Speech communication*, 89:25–36.

Charlotte Gooskens and Vincent J Van Heuven. 2021. Mutual intelligibility. *Similar languages, varieties, and dialects: A computational perspective*, pages 51–95.

Nanna Haug Hilton, Charlotte Gooskens, and Anja Schüppert. 2013. The influence of non-native morphosyntax on the intelligibility of a closely related language. *Lingua*, 137:1–18.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, and 1 others. 2020. spacy: Industrial-strength natural language processing in python.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

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

VI Levenshtein. 1966. Binary codes capable or correcting deletions, insertions, and reversals. In *Soviet physics-doklady*, volume 10.

Francisco A. Marcos-Marín. 2000s. Corpus Oral de Referencia de la Lengua Española Contemporánea (CORLEC). Web Archive of the Laboratorio de Lingüística Informática, Universidad Autónoma de Madrid.

Francisco A. Marcos-Marín and Ernesto Evans Espiñeira. 1991. Corpus Lingüístico de Referencia de la Lengua Española en Chile. Web Archive of the Laboratorio de Lingüística Informática, Universidad Autónoma de Madrid.

Francisco A. Marcos-Marín and Verónica Zumárraga. 1992. Corpus Lingüístico de Referencia de la Lengua Española en Argentina. Web Archive of the Laboratorio de Lingüística Informática, Universidad Autónoma de Madrid.

April MS McMahon. 1994. *Understanding language change*. Cambridge university press.

April MS McMahon and Robert McMahon. 1995. Linguistics, genetics and archaeology: internal and external evidence in the amerind controversy. *Transactions of the Philological Society*, 93(2):125–225.

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.

Martin F Porter. 2001. Snowball: A language for stemming algorithms.

Tommaso Raso, Heliana Mello, Maryualê Malvessi Mittmann, and Alessandro Panunzi. 2012–2015. C-ORAL Brasil 2. Corpus website of the Laboratory of Psycholinguistics, Federal University of Minas Gerais.