

What Makes Blood Thicker than Water: Related Word Discrimination in Romance Languages from a Multi-dimensional Perspective

Ana Sabina Uban^{♠,♡} Teodor-George Marchitan^{♠,♡} Ioan-Bogdan Iordache^{♠,♡}
Simona Georgescu^{♣,♡} Liviu P. Dinu^{♠,♡}

University of Bucharest, [♠] Faculty of Mathematics and Computer Science,

[♣] Faculty of Foreign Languages and Literatures, [♡] HLT Research Center

{auban,ldinu}@fmi.unibuc.ro,

{teodor.marchitan,ioan.iordache}@s.unibuc.ro,

simona.georgescu@lls.unibuc.ro

Abstract

We provide in this study a multi-faceted analysis of related word detection with automatic methods for Romance languages (including the five main Romance languages: Spanish, French, Romanian, Italian, Portuguese), by combining complementary orthographic, phonetic and semantic features (including both cognates and borrowings in all five languages and a series of classification tasks), with a focus on assessing the impact of the semantic component in specific patterns for Romance related words. We perform quantitative explorations as well as classification experiments for every language pair and different types of etymological relations separately: cognate detection, cognate versus borrowing discrimination and borrowing direction detection.

1 Introduction and related work

Languages rarely exist in isolation. Across language families, words preserve traces of shared origins, historical contact, but are also subject to inevitable linguistic change. On the one hand, we find obvious similarities between related languages, that are also transparent to non-expert speakers: it is, in the first place, the case of cognates (words that derive from the same etymon, e.g. Ro. *casa*, Es. *casa* 'house' are cognates because they both come from La. *casa* 'hut'), or, in fewer cases, borrowings (words adopted in language A - target language - from language B - source language, e.g. Es. *valija* 'luggage' is a borrowing from It. *valigia* 'id.'). On the other hand, a complex network of phonetic change, semantic shift, contact interference, and divergence from the common ancestor may obscure the genetic relationships between words, often rendering them unrecognizable to the non-specialist speaker (e.g. It. *casa* n. 'house' – Fr. *chez* prep. 'at someone's home'), and sometimes even to historical linguists (e.g. Es. *hola* and Ro. *aoleu*, whose

common origin in Ar. *Allah* remained unnoticed until very recently (PAVEL and Georgescu, 2025; Georgescu and Pavel, 2024).

In recent years, advances in computational linguistics and natural language processing have opened new possibilities for exploring these relationships automatically, allowing researchers to identify lexical correspondences on a scale that would be difficult to achieve through manual analysis alone. In historical and comparative linguistics, automatic cognate detection can support the reconstruction of relationships between languages (Malory and Adams, 2006; Dunn, 2015; Alekseyenko et al., 2012), the identification of common ancestral forms (Dinu et al., 2024a), and the analysis of patterns of language change (Uban et al., 2021). Computational approaches are particularly valuable in this context because they allow researchers to process large lexical datasets that would be difficult and time-consuming to analyze manually.

Beyond historical research, the identification of lexical similarities across languages also has important implications for language learning. Learners often rely on similarities between known and target languages when acquiring vocabulary, and automatic systems can assist this process by identifying transparent cognates (Ciobanu and Dinu, 2014; Rama, 2016; Jäger et al., 2017; Rama et al., 2018; Fourier and Sagot, 2022; Dinu et al., 2023b), highlighting potentially misleading false friends (Uban and Dinu, 2020; Inkpen et al., 2005; Nakov et al., 2009; Chen and Skiena, 2016; St Arnaud et al., 2017), and supporting more efficient vocabulary acquisition strategies.

In the field of natural language processing, cognate detection contributes to the improvement of machine translation systems, multilingual embeddings, or cross-lingual information retrieval. These applications are especially significant for low-resource languages, where the scarcity of annotated data makes the transfer of linguistic information

from related languages particularly useful.

In the case of endangered or underdocumented languages, automatic comparison techniques may help uncover relationships with better-described languages, thereby contributing to language documentation and revitalization efforts. Similarly, the detection of related words can provide insights into language contact phenomena by helping distinguish inherited vocabulary from borrowings and by revealing patterns of historical and cultural interaction between linguistic communities.

Previous work has shown that it is possible to detect related words automatically with high accuracy, based on their surface form alone (Dinu et al., 2023a, 2024c,b; Ciobanu and Dinu, 2014; Rama, 2016; Jäger et al., 2017; Rama et al., 2018; Fourrier and Sagot, 2022; Dinu et al., 2023b). Nevertheless, in cases where word forms have suffered considerable changes compared to the original etymon, related words might remain undistinguishable based on orthography alone. Semantic clues in combination with orthography and phonetics might provide a more complete picture of the phenomenon of language change in related words. There are almost no studies that report using semantic features in combination with orthographic and phonetic features for the automatic detection of related words including cognates and borrowings, to our knowledge. Labat and Lefever (2019) perform false friend and true cognate discrimination based on manually annotated pairs of words in English and Dutch, using a combination of orthographic and semantic features. Other studies focusing on Romance languages report experiments for automatic discrimination of true and deceptive cognates based on purely semantic representations (Uban and Dinu, 2020; Uban et al., 2025).

In this study, we aim to provide a complete analysis of related word detection with automatic methods for the five main Romance languages, from different perspectives: we compare a set of multi-dimensional complementary features including orthographic, phonetic and semantic similarities and drifts in Romance related words, we include both cognates and borrowings in all five languages (for the first time, to our knowledge), based on an exhaustive database of cognates and borrowings in Romance languages sourced from etymological dictionaries (Dinu et al., 2023a), with a focus on assessing the understudied impact of the semantic component in specific patterns for Romance related words. We perform quantitative explorations

as well as classification experiments for every language pair in the five main Romance languages (Spanish, Italian, Portuguese, French, Romanian), for different types of etymological relations separately: cognate detection, cognate versus borrowing discrimination and borrowing direction detection. We aim to address the following research questions:

RQ1. How well can we automatically identify related word pairs in Romance languages (cognates and borrowings) using complementary information on word orthography, phonetics and semantics, derived from etymological dictionaries, phonetical transcriptions, and word usage in corpora?

RQ2. What can we say about orthographic, phonetic, and semantic drifts that related word pairs undergo across time and how do they compare?

RQ3. What is the impact of the semantic component particularly on related word detection and classification?

2 Related words dictionaries and corpora

Cognates and borrowings dataset We perform our analyses on related word pairs extracted from the most comprehensive database of related words in Romance languages up to date, sourced from etymological dictionaries and manually curated, RoBoCoP (Dinu et al., 2023a), comprising a total of 125,598 words across all languages. For cognate detection, we focus on the words with Latin etymons. As such, we additionally 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), containing only words of Latin origin. We extract borrowings from the original RoBoCoP database, totaling 46,490 borrowing pairs (Dinu et al., 2024b).

Word embeddings corpora For our computational experiments, we rely on word embeddings as models of meaning representation. We include three different corpora to extract contextual embeddings:

- Wikipedia¹
- *Europarl*, a standard parallel corpus with aligned sentences including the Romance languages, based on proceedings of the European Parliament (Koehn, 2005),

¹<https://huggingface.co/datasets/wikimedia/wikipedia>

- *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),

3 Methodology

3.1 Preprocessing

Word matching. All experiments operate on a single union vocabulary per language, assembled from the related-word resources described in Section 2. Surface forms are harvested with their original orthography and diacritics, which is the form expected by the multilingual encoders. A second, ASCII-normalised copy of every word is produced via `unidecode` and is the one fed to the alignment-based n-gram extractor used in previous work (Dinu et al., 2023a, 2024a), so that dictionary lookups and form-based features remain character-for-character comparable to the baselines.

For static embeddings (static LABSE) every vocabulary word receives a vector, so coverage is by construction 100%. For contextual embeddings (LABSE, XL-LEXEME), each vocabulary word must additionally be located in the parallel corpora before it can be encoded. One difficulty encountered when obtaining contextual representations was related to the incomplete coverage in corpora of the words in our dictionaries. In order to mitigate this problem, we rely on strategies meant to improve coverage by relaxing the matching criteria between the dictionary words and their occurrences in corpora. Occurrence matching is performed on stems rather than surface forms: each vocabulary word is stemmed with the language-specific Snowball stemmer and each corpus document is tokenised with `spaCy`; a stem match counts as a hit. Italian stems are computed on the `unidecode`-normalised token, mirroring the pre-existing phonetic pipeline. When two vocabulary words collide on the same stem, the shorter surface form is kept as the canonical key, so no entry is silently dropped because it shares a stem with another entry. The final coverage after this matching strategy is shown in (Figure 18 in the Appendix. A brief evaluation of the precision of this matching method and its faults is also presented in the Appendix. Moreover, the minority of words for which an in-corpus context is still not found are not removed from the dataset: they are backfilled with their corresponding static variant, depending on the embedding model. In this

way, representations of words that are not covered in the corpora are kept in the same representation space, and the vocabularies used in all experiments are kept identical.

Phonetic transcription. IPA transcriptions for the five Romance languages are obtained off-line with `eSpeak NG`². Each surface form is first filtered to keep only alphabetic characters plus hyphens, apostrophes and intra-word spaces, then passed to the `g2p` engine with the IPA flag enabled. The resulting phoneme strings are post-processed by two language-agnostic rules: optional segments emitted in parentheses are dropped and runs of underscore separators are contracted to a single underscore so that phoneme tokenisation by splitting on `_` produces one token per phoneme.

3.2 Word meaning representations

Our proposed algorithms rely on word embeddings to measure semantic distances. Our task requires comparing meanings (distances) between words across languages. While for static embeddings such as FastText, solutions such as post-alignment algorithms have been established as reasonable solutions for comparing embedding spaces across languages (Bojanowski et al., 2016), this is a rarer application of most popular contextual embedding models. To our knowledge, there are no studies in published literature reporting a systematic comparison of existing multilingual encoders for cross-lingual lexical semantics tasks such as the ones we are interested in our study. Most evaluations are performed on either monolingual lexical tasks such as lexical semantic change detection (where word similarity is measured only within language and not across languages), or sentence-level cross-lingual tasks, such as bi-text retrieval. For our experiments, we chose to select two semantic representation models which were reported to obtain state-of-the-art performance for each of these respective tasks, and so were more promising than other popular alternatives such as mBERT or sBERT. In future work, such alternatives could be useful to explore.

LaBSE. As a general multilingual contextual word embedding model, we use LaBSE (Feng et al., 2022), which was trained for cross-lingual sentence similarity. LaBSE is an extension of sBERT shown to achieve state-of-the-art performance for bi-text retrieval for low-resource languages (Feng et al.,

²<https://github.com/espeak-ng/espeak-ng>

2022); in 2025 it obtained second best performance for the bi-text retrieval task in the Massive multilingual text embedding benchmark (Enevoldsen et al., 2025) and the best overall performance for low resource languages on the Multilingual semantic textual similarity benchmark (Ranasinghe et al., 2025).

XL-LEXEME. We additionally experiment with a model designed specifically for the task of multilingual lexical semantic change detection and multilingual word disambiguation (Cassotti et al., 2023) that was shown to perform well in monolingual lexical semantic change tasks, settings which are more similar to our task than the previously mentioned benchmarks, even though their cross-lingual capabilities have not yet been rigorously evaluated in the literature so far.

We experiment with these representations in two settings: one "static" approach where we extract meanings of individual tokens out of context (we input the single tokens into the model directly, preprocessed with special character normalization such as accents and diacritics), and a regular contextual approach, where we extract for every word one embedding for each occurrence in the corpus, for all three corpora considered.

Contextual embeddings aggregation. For every word in the union vocabulary we harvest up to 100 in-corpus occurrences, drawn deterministically (seed 42) and length-capped at 128 LaBSE subword tokens per sentence. Contexts are merged across the three corpora in a round-robin schedule (EuroParl → Wikipedia → RomCro), so high-frequency Wikipedia matches do not starve the smaller corpora.

Each occurrence is encoded independently: for LABSE we run the underlying transformer with offset mappings, build a boolean mask over subwords whose character span overlaps the target span, and mean-pool the last-layer hidden states across that mask, falling back to the [CLS] vector when the target is truncated; for XL-LEXEME the target span is wrapped with the model’s `<t>/</t>` markers and the sentence is passed through the SentenceTransformer wrapper, which returns a single pooled vector per occurrence. The per-word representation is the unweighted mean of all per-occurrence vectors (768-dim for LABSE, 1024-dim for XL-LEXEME), persisted alongside the static caches under an identical .npz schema so that downstream classifiers consume both regimes

through the same lookup.

3.3 Classification experiments

We perform classification experiments for three main classification tasks. We follow the classification setups and model types from previous work on related word classification using the same data but only graphic and phonetic features (Dinu et al., 2024c, 2023a, 2024b), and we enrich the feature space with semantic information. We eventually compare our results with the baseline results reported in (Dinu et al., 2023a) to assess the importance of the added semantic criterion.

I. Cognate detection. We frame cognate detection as a binary task of classifying word pairs as cognates or non-cognates, with separate training per language pair.

II. Cognate versus borrowing classification. We discriminate cognate from borrowing pairs and include two setups for the experiments: cognates versus borrowings, and cognates versus borrowings versus unrelated pairs.

III. Borrowing direction detection. We train models to detect, for a given pair of borrowings (w_1, w_2) from languages (l_1, l_2) whether w_1 was borrowed from l_2 or w_2 was borrowed from l_1 .

Negative examples selection. In order to ensure a less trivial distinction, we choose to experiment with two different setups for the negative class: random pairs of unrelated words generated from our word lists, as well as more difficult pairs of unrelated but graphically similar words, by selecting only words with Levenshtein distance lower than the average value for related words in our dataset. We use these negative examples for the task of cognate detection, as well as for the second task of cognate-borrowing classification, in a three-way setup including unrelated words as a third class.

For all experiments, we use a 90% : 10% stratified split to generate train and test sets, which are initially shuffled. The reported metrics were validated and averaged over multiple training runs, with different random seeds for the models’ parameters. These runs displayed very little to no variation in the scores.

3.3.1 Features

We use three types of features: graphic, phonetic and semantic. As far as we are aware, the inclusion

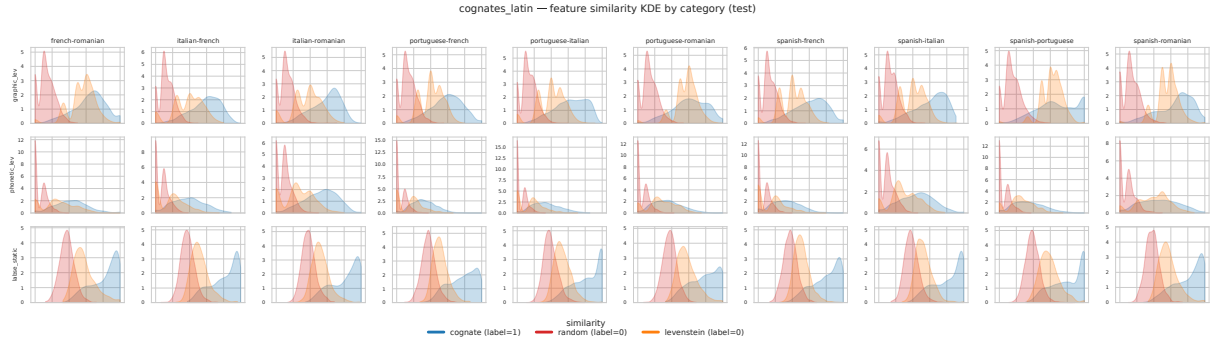


Figure 1: Distribution of graphic, phonetic and semantic similarity for cognate pairs, based on Levenshtein distance for graphic and phonetic features and cosine similarity for semantic features (LaBSE static - for other semantic representations see Appendix).

of a semantic feature for related word classification task (including cognates and borrowings) for Romance languages is introduced here for the first time. Orthographic and phonetic features are encoded in different ways according to the model.

To integrate semantic representations, we follow established NLP pair-classification methodologies (Conneau et al., 2018) by computing a composite feature block from the embedding pairs. Given two word embeddings e_1 and e_2 , we extract three complementary geometric metrics:

Cosine similarity: A global, scale-invariant scalar that effectively captures broad conceptual alignment, serving as a baseline to separate related pairs from unrelated pairs.

Element-wise absolute difference ($|e_1 - e_2|$): A symmetric vector capturing per-dimension divergence. This allows the linear classifiers to learn and weight the specific latent dimensions that differentiate evolutionary semantic shifts (cognates) from transfer patterns (borrowings).

Element-wise product ($e_1 \odot e_2$): A signed vector representing per-dimension co-activation. This provides the classifier with a granular, uncollapsed view of the interactions that make up the cosine similarity, highlighting areas of strong mutual activation.

Notably, because these features are inherently symmetric, they are highly optimized for bidirectional classification tasks (e.g., cognates versus borrowings). Consequently, their order-invariance limits their direct predictive power in asymmetric tasks, which explains the variable impact of semantic features in borrowing direction detection.

Ablation experiments are built into our methodology, since our approach is to extend methods

previously proposed in other related word classification studies (Dinu et al., 2024c,b, 2023a) to include semantic features and compare results to assess the impact of semantics on classification performance. We additionally experiment with semantic features in combination with graphic-only or phonetic-only features.

3.3.2 Models

Classic machine learning model ensembles.

Our first experimental approach relies on ensemble models. It consists of applying a set of machine learning algorithms (Support Vector Machine, Naive Bayes, and other linear classifiers trained using stochastic gradient descent) on the features extracted from the alignments and then building a stacking ensemble classifier from the top 5 performing models (Table 1). The individual models are evaluated and selected via 3-fold cross validation on the training split, and we pick the best 5, for each language pair separately. The final ensemble model is trained using all of the training examples and its performance is ultimately assessed on the test split.

The ensemble pipeline used in previous work is built on top of Needleman–Wunsch alignment of the two surface forms (orthographic and / or phonetic) followed by extraction of character n-gram differences and a stack of linear classifiers (linear SVMs and stochastic gradient descent classifiers with various losses, over $n \in 1, 2, 3$ and word / phoneme configurations) combined by a stacking meta-classifier. Specifically, two words are aligned, then we look at indices of misalignments (insertions/deletions/replacements) and build character n-grams around those indices from both words. One feature is an n-gram misalignment pair (e.g.

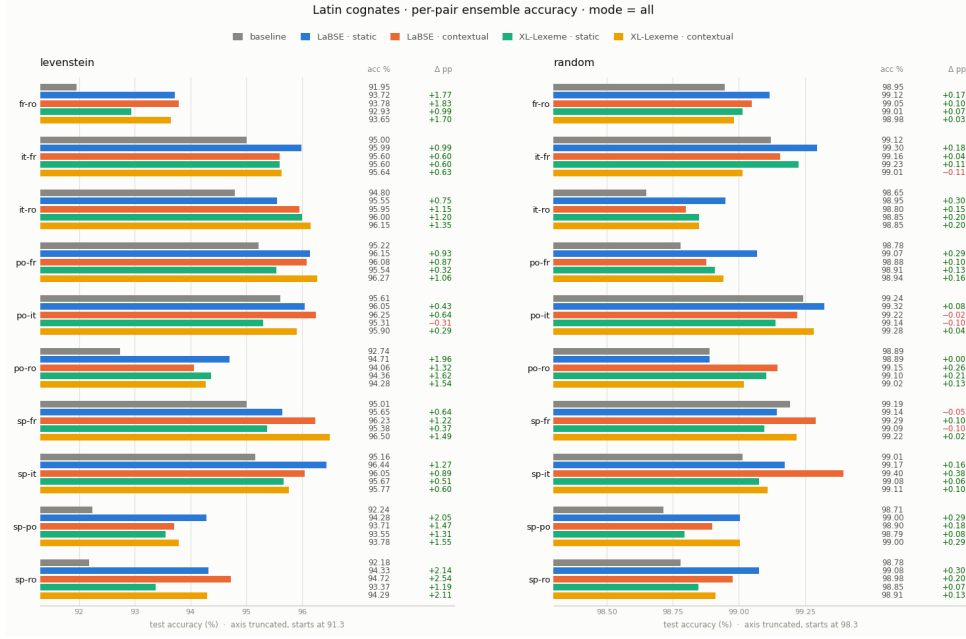


Figure 2: Cognate detection results (accuracy) using the best ensemble models and all features, with LaBSE static representations, in comparison to baseline results without semantic features.

$AAA - BBB$, where AAA is a 3-gram from the first word and BBB is the corresponding 3-gram from the second one, at the same indices). These features are then collected in a bag-of-words manner. In the current study, we maintain the same encoding for the graphic and phonetic features, to which semantic features are added as a parallel feature branch. Semantic features are concatenated and the alignment ones, following an L2-normalization.

Character transformer. Our second approach uses transformer models. We train such models either on the graphic, or on the phonemic form of the words. The "tokenization" is performed by splitting the representations into letters or into phonemes, without any other normalization. For a given pair of words, we prepend the first sequence of tokens with a special [CLS] token, and insert a [SEP] token between the two sequences. The resulted list of tokens is then positionally embedded and fed to a multi-layered transformer encoder. The embedding returned by the last layer of the model for the [CLS] token is used for classification via a feed-forward layer, that reduces it to a size 2 vector. For training we use the same 90% train split as for the ensembles. The transformer models are finally evaluated on the 10% test split in order to allow fair comparisons with the ensemble approaches. Graphic and phonetic encodings of words are fed

as simple character sequences to the transformer model. Semantic embeddings are L2-normalized prior to being fed to the model.

4 Results

4.1 Exploratory analysis

The distribution of similarities at different levels (graphic, phonetic, semantic) for cognates is illustrated in Figures 1. We notice a higher variability in the case of graphic similarity, while phonetic similarities are more sharply skewed towards small values. Semantic similarities are more widely and more normally distributed, but generally clearly separable across groups, already indicating that semantic information can be a useful clue for distinguishing cognates from unrelated words. The same analysis for borrowings can be found in the Appendix.

4.2 Cognate detection

Figure 2 shows the best results for the main cognate detection task using the ensemble classifier with all features. Results are shown for both negative sampling settings, in comparison with the baseline results. Results per language pair using only graphic or phonetic features, in combination with semantic features, are shown in the Appendix, as is the comparison with the results obtained with the transformer models, which are generally slightly lower.

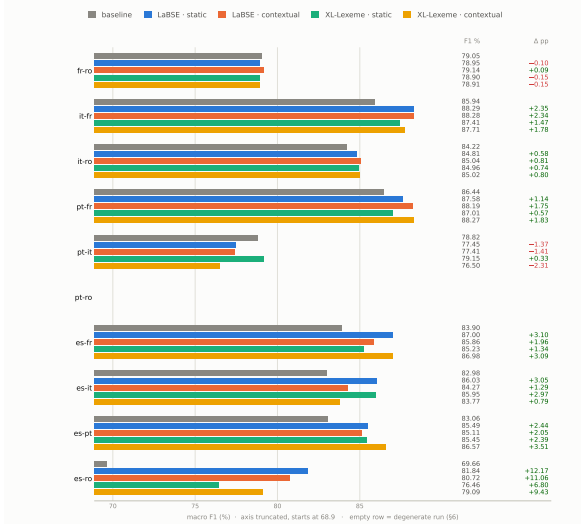


Figure 3: Classification results (macro-F1) for the borrowings-cognates-negatives 3-way classification, Levenshtein-based negative selection, using the ensemble with all features.

We notice that for all language pairs results are improved when adding the semantic criterion, with most pairs reaching accuracies up to 99% in the random negative sampling setting and up to 96% in the more difficult setting using Levenshtein-similar words as negative examples. We suggest that this result can constitute a new benchmark for automatic cognate detection for Romance languages and conclude that cognate words are automatically detectable with very high accuracy even in difficult settings, especially when all dimensions of word similarity are considered. The best detection performance is obtained for Spanish-Italian and Spanish-French, and the lowest for French-Romanian. The highest gains compared to the baseline are generally obtained precisely for the language pairs where the baselines were lowest, such as Spanish-Romanian and Spanish-Portuguese, showing that the addition of the semantic feature is able to compensate partially for the performance gaps between language pairs.

Among embedding models, LaBSE static embeddings generally outperform other variants.

4.3 Cognate-borrowing discrimination

Results for borrowing discrimination against cognates and negative cases using the best overall model (the ensemble with all features) are shown in Figure 3, with additional results in the Appendix.

For cognate-borrowing discrimination, adding

semantic features still improves upon classification performance for most language pairs, although it does not universally guarantee improved performance. Performances for ternary classification in the more difficult negative sampling setting reach .87 F1. While the gains provided by semantics are generally comparable to the ones seen for cognate detection, here semantics especially provide distinct advantages for specific language pairs, such as Spanish-Romanian, which shows a .12 improvement in ternary classification with ensemble models and all features, and Portuguese-Romanian especially when excluding phonetic features (+.25). Similarly, results for transformer models are significantly improved for Spanish-Portuguese related words by adding semantics in transformer models when using only graphical features in the baseline.

For the binary setting, detection performance using all features reaches at most .87 macro-F1 for Spanish-Portuguese and Italian-French (with the exception of the trivial Italian-Romanian test set obtaining .99), with the lowest performance using all features sitting at .72 for French-Romanian. Borrowings are generally more distinguishable from cognates when including the semantic criterion across models and language pairs. As in the ternary case, Spanish-Romanian obtains a distinct advantage in the case of ensemble models with all features (+.14), and an even higher one when using only phonetic features (+.20).

4.4 Borrowing direction detection

For borrowing direction classification, the impact of adding semantic features is highly variable across language pairs and heavily dependent on the chosen model architecture and input representation (Figure 4 shows ensemble results, Figure 13 shows transformer results, additional results in the Appendix, including scores per language pair and direction can be seen in Figure 14). When using all features, semantics provide a gain on average, reaching an average F1 of .89 across language pairs. In these experiments, contextual embeddings seem to generally outperform their static variants. It might be the case that contextual embeddings are especially useful in settings with fewer training data, such as is the borrowing direction task, as well as some of the previous tasks including Romanian, for which contextual embeddings sometimes obtain best results. Overall, transformer models tend to outperform ensemble models in many language pairs. The highest average performance across lan-

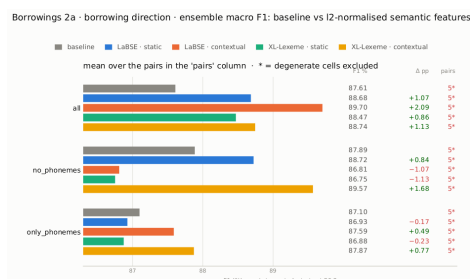


Figure 4: Classification results (macro-F1) for the borrowing direction detection, using the ensemble with all feature combinations, averaged across language pair.

guage pairs is .89 F1 for the ensemble using all features. The most significant gains are obtained for Italian-Spanish and Spanish-Italian borrowings.

5 Error analysis and discussion

Features comparison. The impact of semantics.

We compare detection results using only graphic and phonetic features (in different combinations and with different models) with the same models including semantic features. Most experiments are generally benefited by the inclusion of semantics, with the most conclusive results in terms of the impact of semantics are in the case of cognate detection: in all settings and for all language pairs, adding semantic features improves detection performance, even for tasks where accuracy was already very high (>95%). This proves that semantic representations encode fine-grained relational information that remains useful beyond surface-form similarity.

Generally, including all features provides best results, and graphic features only in combination with semantics are more useful than phonetics and semantics alone. In the case of borrowings, the high variability in the volume of borrowing data compared to cognate data makes patterns less stable even in baseline experiments.

Feature-based error analysis. We show the errors that are recovered by including semantic features, and the errors that are introduced by adding semantics, in Figure 5. All points are to the left of the main axis, showing that for all experiments, including semantic features fixes more errors than it introduces. The gains are higher when excluding either graphic or phonetic features. For certain language pairs for which the errors fixed by introducing semantics are most dominant: Spanish-Romanian, Spanish-French.

Models comparison. The classic ensemble-based models generally outperform the character-transformer models, which might prove the importance of alignment features, which are in accordance with the comparative method in linguistics. LaBSE embeddings generally outperform XL-LEXEME ones, and the contextualized embeddings do not seem to improve upon the simple "static" version of LaBSE embeddings. Contextual embeddings have the advantage for some lower resource settings and language pairs, such as in the case of cognate detection for language pairs including Romanian, and for some of the borrowing direction detection experiments.

Example-based error analysis. Below are some examples of cognate pairs that could not be correctly identified on the basis of external criteria alone (orthography/phonetics), but whose relationship was recovered through the inclusion of the semantic component:

- cases involving highly divergent phonetic forms: Fr. *aider* – Ro. *ajuta* 'help'; Es. *objeto* – It. *oggetto* 'object';
- cases involving relatively similar phonetic forms: Fr. *titre* – Ro. *titlu* 'title'; Es. *cero* – Ro. *zero* 'zero';
- cases involving identical phonetic forms: Fr. *marin* – Ro. *marin* 'marine'; Es. *mandarín* – Ro. *mandarin* 'Mandarin' / 'mandarin tree'.

There are also cases in which words displaying a highly similar phonetic form were initially classified as cognates, while the semantic component added in the second stage helped differentiate them: e.g Fr. *petite* 'small (fem.)' – Ro. *pepită* 'nugget'; Es. *ascenso* 'promotion, ascent' – Ro. *ascuns* 'hidden'.

If we take a closer look specifically at which types of words tend to be misclassified based solely on the graphical form but are recovered by the classifier as true cognates when including semantics, we notice: Several cases of related words that were missed by graphic and phonetic models are very short 3 or 4 letter words (Pt. *bom* / It. *buono* 'good', Pt. *caso* / Fr. *cas* 'case', etc.). Interestingly, there are also many examples of longer cognates which are almost identical orthographically between the two languages (e.g. Es. *mandarín* / Ro. *mandarin*, Es. *serenadal* / Ro. *serenadă*), which were misclassified as unrelated by the graphic and phonetic

models, and correctly identified as cognates only when including semantic features. This might be explained by the limitations of purely form-based representations, which may fail to capture broader lexical correspondences when orthographic or phonetic similarity alone is insufficiently weighted by the model. In such cases, semantic features appear to provide an additional layer of confirmation by encoding the fact that the words occur in highly similar contextual and conceptual environments across languages.

Some examples of word pairs that are difficult to distinguish solely based on form, but become easier to classify as related words when semantics is included are the following: Pt. *bispo* – It. *vescovo* ‘bishop’ from Latin *episcopus*, where both words underwent several irregular phonetic developments: they both display apheresis (*episcopus* > **biscopu*), the Portuguese form additionally underwent syncope (**biscopu* > **bispo* > *bispo*), while the lenition found in the Italian form is also unexpected (*p* > *v*). Likewise, in the pair Sp. *gritar* – Fr. *crier* ‘shout’ from Lat. *quiritare*, the two Romance outcomes diverge from the regular path of phonetic evolution: there are no other Spanish words beginning with *qui-* that developed into *g-*, and, similarly, the syncope in both cases (*quiri-* > *gri-/cri-*) constitutes an irregular phonetic development. As a result, the two forms are difficult to identify as cognates in the absence of semantic information.

It is worth noting that, with the help of the semantic criterion, the algorithm also identified as cognates certain word pairs initially classified as unrelated, which, unlike the previously discussed cases, also underwent semantic changes that further distanced them from one another, e.g. It. *sella* ‘saddle’ – Ro. *șale* ‘hips’: in this case, the system effectively reconstructed the missing conceptual link between the two meanings by identifying that Ro. *șale* refers to the part (of the horse) on which the saddle is placed.

On the other hand, some related words are correctly detected by the original models and missed after including semantics. This can be explained in cases where semantic shifts were more severe or atypical, leading to lower semantic similarity despite the existence of a clear historical relationship between the forms. In such cases, the semantic representations may reflect the contemporary usage of the words rather than their etymological origin, causing genuinely related items to appear seman-

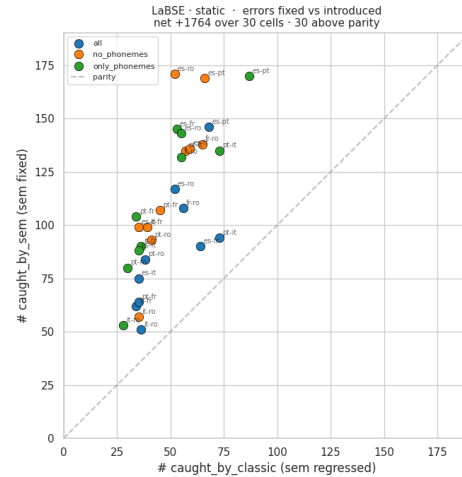


Figure 5: Error comparison with and without semantic features for cognate detection using ensemble model with LaBSE embeddings.

tically distant (e.g. Es. *luego* ‘afterwards’ and Ro. *loc* ‘place’ both originate in Lat. *locus* ‘place’, but the Spanish descendant underwent a metonymic shift from space to time). As a result, the semantic features can introduce noise or conflicting evidence that outweighs the strong graphic and phonetic correspondence captured by the original models. This highlights an important limitation of semantic information in cognate detection: while it is generally beneficial, its effectiveness depends on the degree of semantic preservation between related words across languages.

6 Conclusions

We have provided an analysis of related words and automatic methods for detecting such pairs in the five main Romance languages, and studied the specific impact of the semantic information on detection performance, for different types of etymological relations. For most language pairs, cognate detection is achievable with 96% accuracy, while borrowing detection, as well as borrowing direction detection performance reach up to .87 F1 scores in different setups, with more variations across language pairs. We show that semantics is universally relevant for cognate detection across language pairs, and generally useful in the case of borrowing detection as well, while borrowing direction detection scenarios remain more complex. Future work experimenting with additional semantic representations optimized for cross-lingual similarity and a more careful consideration of polysemy could further improve results.

Limitations

Some refinements of the data in the RoBoCoP database, such as excluding words with multiple etymologies, which are not currently handled in the available version of the database, could be useful for a more accurate model of the linguistic phenomenon. Homonyms also pose an issue when matching dictionary words with corpus occurrences, which points to a necessity to handle polysemy in semantic representations, but especially to be able to distinguish homonyms in the source dictionary. Graphic issues in the dictionary can also lead to errors in detecting cognates. In future work, improvements to the word matching algorithm, as well as to the way missing words are handled in the models (currently replaced with zero-valued embeddings) could be useful.

The underperformance of contextual embeddings also suggests room for potential improvement in handling polysemy, for example by experimenting with cluster-based similarity measures such as Jensen-Shannon divergence. Limitations in meaning representation might be more severe in the case of lower-resource languages, such as Romanian, where the benefits of contextual representations seem to be higher in some scenarios, possibly by compensating with the corpus contexts for the overall poor representation of Romanian words during model pretraining. Moreover, the current lack of dedicated pretrained models optimized for cross-lingual semantic representations might limit the final performance of the solutions including semantic features.

Additional setups for related word detection, such as unsupervised learning, or language transfer, remain unexplored in this study.

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 upon request.

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, by "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-

2027, MySMIS no. 351416, and by the UEFISCDI Project "Centru de Excelenta pentru Schimbări Climatice si Sociale (CESCS)", project number PNIV-P6-6.1-CoEx-2024-0042.

References

- Alexander V. Alekseyenko, Quentin D. Atkinson, Remco Bouckaert, Alexei J. Drummond, Michael Dunn, Russell D. Gray, Simon J. Greenhill, Philippe Lemey, and Marc A. Suchard. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. *Science*, 337:957–960.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.
- Pierluigi Cassotti, Lucia Siciliani, Marco DeGemma, Giovanni Semeraro, and Pierpaolo Basile. 2023. X1lexeme: Wic pretrained model for cross-lingual lexical semantic change. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585.
- Yanqing Chen and Steven Skiena. 2016. False-friend detection and entity matching via unsupervised transliteration. *arXiv preprint arXiv:1611.06722*.
- Alina Maria Ciobanu and Liviu P. Dinu. 2014. [Automatic detection of cognates using orthographic alignment](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 99–105. The Association for Computer Linguistics.
- Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2018. [Supervised learning of universal sentence representations from natural language inference data](#). *Preprint*, arXiv:1705.02364.
- Liviu P Dinu, Ana Uban, Alina Cristea, Anca Dinu, Ioan-Bogdan Iordache, Simona Georgescu, and Laurentiu Zoicas. 2023a. 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.
- Liviu P. Dinu, Ana Sabina Uban, Alina Maria Cristea, Anca Dinu, Ioan-Bogdan Iordache, Simona Georgescu, and Laurentiu Zoicas. 2023b. 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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7610–7629. Association for Computational Linguistics.
- Liviu P. Dinu, Ana Sabina Uban, Ioan-Bogdan Iordache, Alina Maria Cristea, Simona Georgescu, and Laurentiu Zoicas. 2024c. [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)*, pages 12657–12667, Torino, Italia. ELRA and ICCL.
- Michael Dunn. 2015. Language phylogenies. *The Routledge handbook of historical linguistics*, pages 190–211.
- Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Winata, et al. 2025. Mmteb: Massive multilingual text embedding benchmark. In *International Conference on Learning Representations*, volume 2025, pages 101715–101771.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 878–891.
- Clémentine Fourrier and Benoît Sagot. 2022. [Probing multilingual cognate prediction models](#). In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3786–3801. Association for Computational Linguistics.
- Simona Georgescu and Cătălin Pavel. 2024. Sp. hola'hello'revisited: Etymological and historical insights. *Philologica Jassyensia*, 20(2).
- Diana Inkpen, Oana Frunza, and Grzegorz Kondrak. 2005. Automatic identification of cognates and false friends in french and english. In *Proceedings of RANLP 2005*, volume 9, pages 251–257.
- Gerhard Jäger, Johann-Mattis List, and Pavel Sofroniev. 2017. [Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1205–1216, Valencia, Spain. Association for Computational Linguistics.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.
- Sofie Labat and Els Lefever. 2019. A classification-based approach to cognate detection combining orthographic and semantic similarity information. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 602–610.
- James P Mallory and Douglas Q Adams. 2006. *The Oxford introduction to proto-Indo-European and the proto-Indo-European world*. Oxford University Press on Demand.
- Bojana Mikelenić, Antoni Oliver, and Marko Tadić. 2024. Expansion of the romero corpus with texts in catalan. In *CLARIN Annual Conference Proceedings 2024*, pages 135–139. Barcelona: CLARIN.
- Svetlin Nakov, Preslav Nakov, and Elena Paskaleva. 2009. Unsupervised extraction of false friends from parallel bi-texts using the web as a corpus. In *Proceedings of RANLP 2009*, pages 292–298.
- CĂTĂLIN PAVEL and Simona Georgescu. 2025. Historical and linguistic notes on the laicization of foreign sacred words in romance languages: Spanish hala 'come on' and romanian aoleu, 'oh, woe' from arabic allah. *REVUE ROUMAINE DE LINGUISTIQUE- ROMANIAN REVIEW OF LINGUISTICS*, 70(1-2):239–260.
- Taraka Rama. 2016. Siamese convolutional networks for cognate identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1018–1027.
- Taraka Rama, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger. 2018. [Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics?](#) In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 393–400, New Orleans, Louisiana. Association for Computational Linguistics.
- Tharindu Ranasinghe, Hansi Hettiarachchi, Constantin Orasan, and Ruslan Mitkov. 2025. Musts: Multilingual semantic textual similarity benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 331–353.
- Adam St Arnaud, David Beck, and Grzegorz Kondrak. 2017. Identifying Cognate Sets Across Dictionaries of Related Languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2519–2528.

Ana Sabina Uban, Alina Ciobanu, and Liviu P Dinu. 2021. Cross-lingual Laws of Semantic Change. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors, *Computational Approaches to Semantic Change*. Berlin: Language Science Press.

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, Ioan-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 15309–15323.

A Appendix

A.1 Infrastructure and libraries

The experiments were performed on an RTX 2080 Ti GPU and a Ryzen 5 3600X CPU for a total of 72 hours.

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
- 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/.

A.2 Configurations for transformer-based experiments

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.
- embedding_size=200
- hidden_state_size=200
- num_attention_heads=8
- num_layers=4
- dropout=0.2
- trainable parameters=980, 202
- 50 epochs
- batch_size=64
- loss=cross entropy loss
- optimizer=Adam
- initial_lr=0.001
- lr_decay=0.6
- patience=5 epochs

As an implementation detail, we evaluate the cross-entropy loss function on a fraction of the training examples after each epoch, in order to stop the training process early, before divergence occurs. This fraction of the examples is not used for backpropagation, so we validate each epoch on an unseen subset of pairs.

A.3 Configurations for classical machine learning experiments

Models used in ensembles: SVC (kernel=rbf, C=0.1/1/10), LinearSVC (C=0.1/1/10), Multinomial Naive Bayes, XGBoost classifier, Linear classifiers trained using Stochastic Gradient Descent (SGDClassifier, loss=hinge/log_loss/modified_hubersquared_hinge/perceptron, class_weight=balanced, max_iter=10,000)

A.4 Additional Results

A.5 Dictionary-corpus matching and corpora coverage

The stem-based matching used for retrieving corpus occurrences of dictionary words is designed to maximize coverage, but it stands to precision errors. We perform a brief analysis of the precision of the matching algorithm for each language, by manually annotating a sample of the overall most frequent 100 words in each language and their top 10 most frequent matches in the corpus based on the Rom-Cro corpus. For each language, one native speaker among the authors annotated each matched word-occurrence pair as correct or incorrect match ("Is this dictionary word correctly identified as the word in this corpus occurrence?"), and the precision was computed as a mean of these annotations weighted by the frequency of each occurrence in the corpus, which should reflect the impact of the match on the final contextual representation used in training our models. Results show modest precision, pointing to potential improvements in the matching algorithm in future work: 58% for Romanian, 59% for Spanish, 71% for Portuguese, 58% for Italian. Aside from some errors introduced by the aggressive normalization, we also find and penalize as errors many potentially faulty matches occurring due to homonymy, in cases where homonyms have different etymologies, which are impossible to avoid based on the available dictionary and data. This points to a potential improvement in the RoBoCoP database to support separate entries for homonyms, for example by including part-of-speech information.

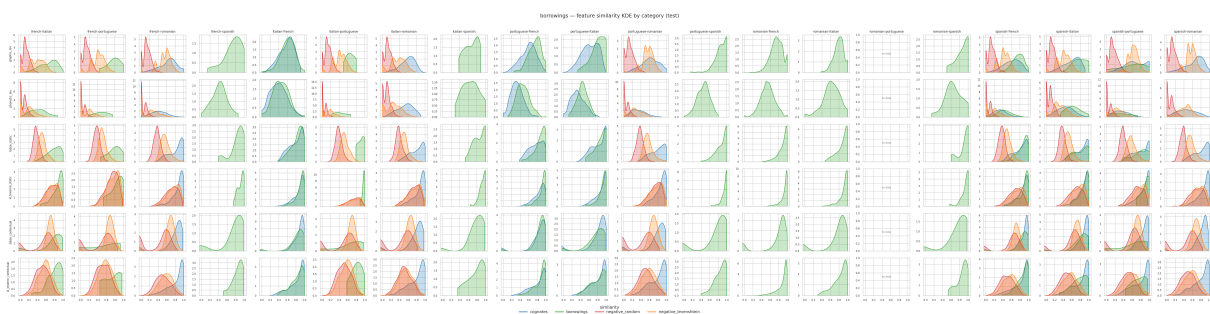
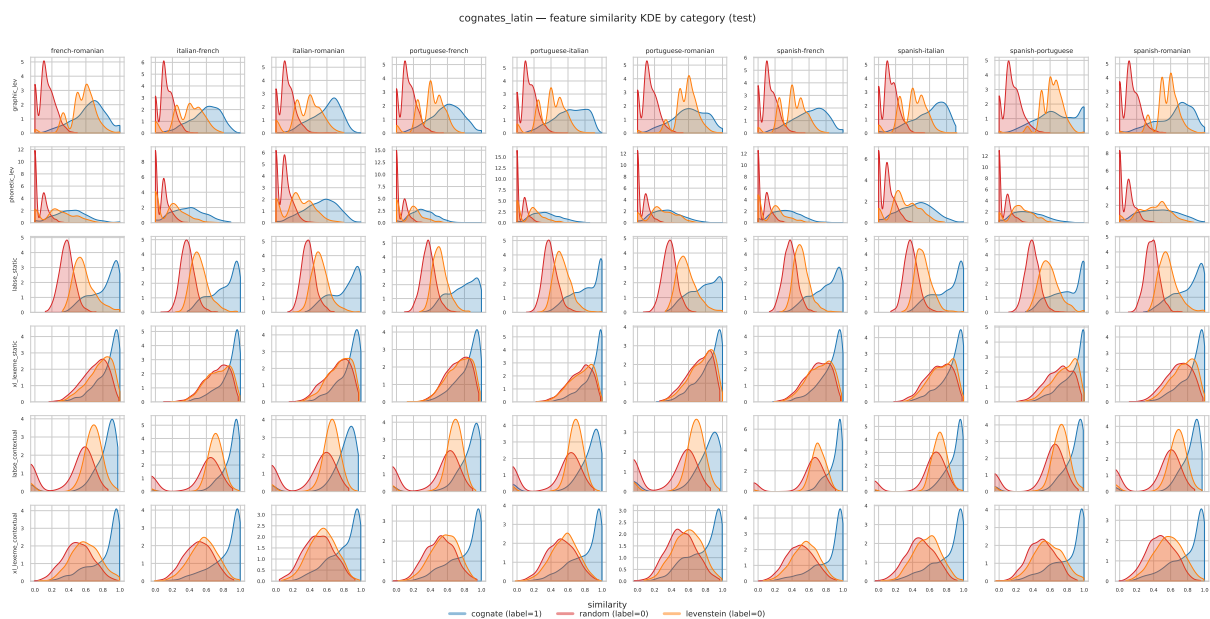




Figure 8: Comparison of ensemble and transformer results for cognate detection.



Figure 9: Comparison of ensemble and transformer results for borrowing versus cognate binary classification.



Figure 10: Comparison of ensemble and transformer results for borrowing versus cognate ternary classification with Levenshtein-based negative sampling.



Figure 11: Comparison of ensemble and transformer results for borrowing versus cognate ternary classification with random negative sampling.



Figure 12: Comparison of ensemble and transformer results for borrowing direction detection, for all models and features.

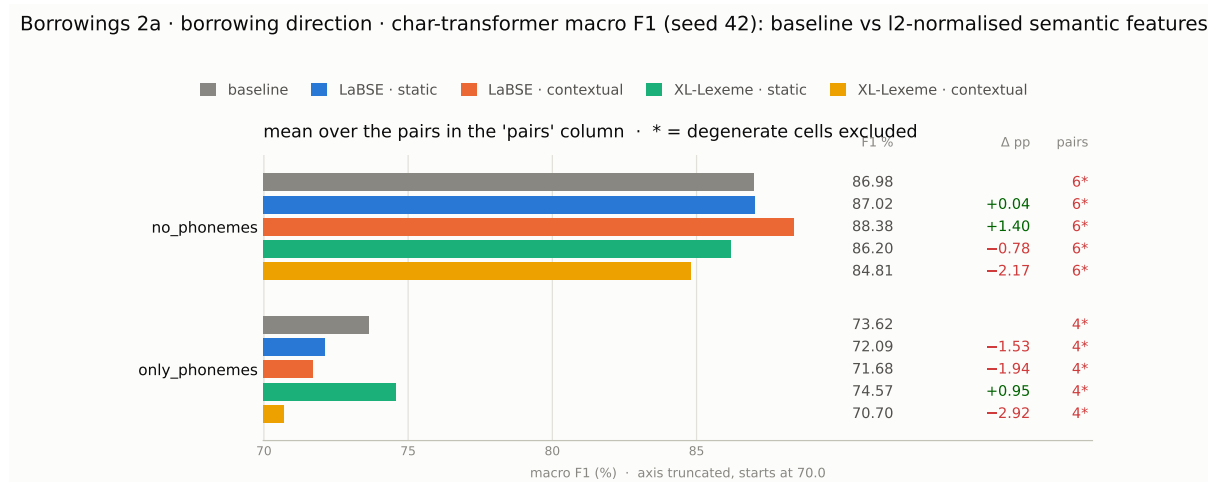


Figure 13: Transformer results for borrowing direction detection, average across language pairs.



Figure 14: **Per-direction F1 on borrowing direction.** One panel per model configuration; twelve groups per panel (six language pairs $\times$ two directions, $\ell_1 \leftarrow \ell_2$ meaning words ℓ_1 borrowed from ℓ_2); five bars per group, baseline and four semantic variations. Hatching marks runs that collapsed to a single direction.

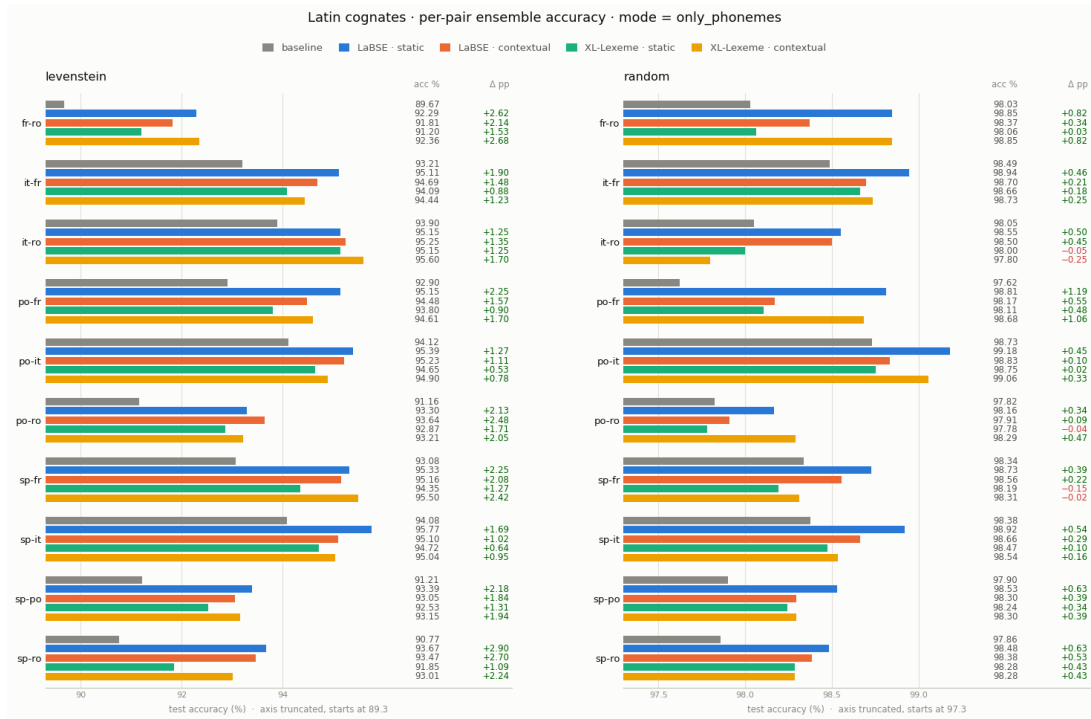
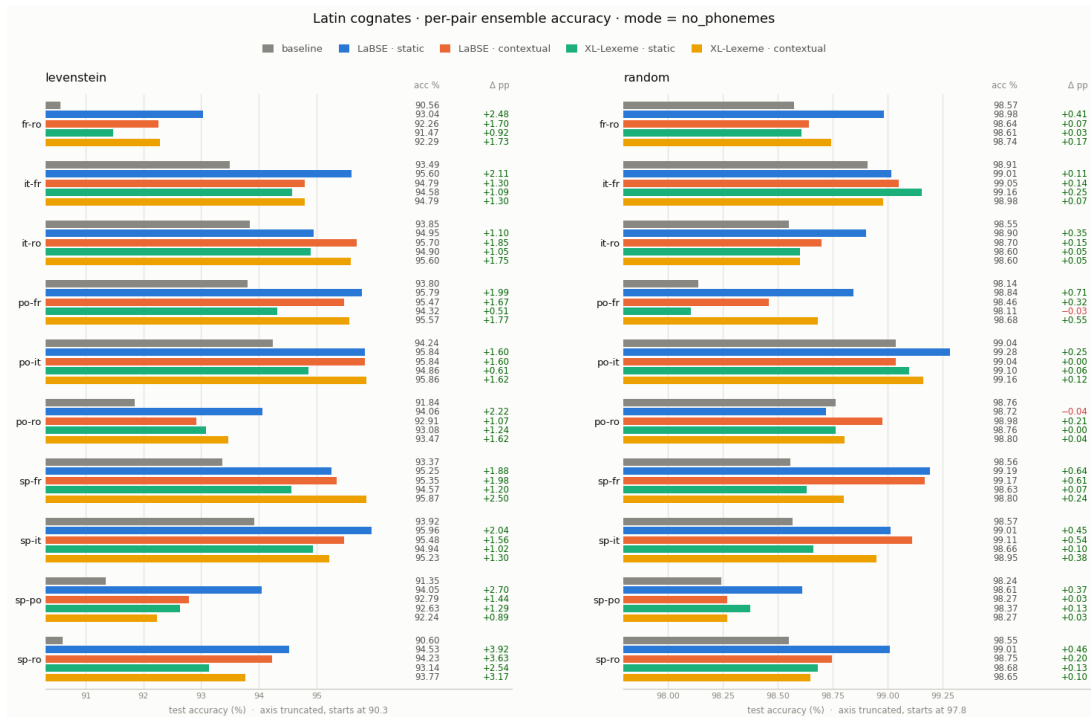


Figure 15: Cognate detection results using the best ensemble models and phonetic and semantic features, in comparison to baseline results without semantic features. (mode = *only_phonemes*)



Language Pair	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
French–Romanian	svm-rbf_C=1_n=2 (93.39)	svm-linear_ C=0.1_n=3 (93.34)	svm-linear_ C=0.1_n=2 (93.16)	svm-rbf_C=1_n=1 (93.07)	svm-rbf_C=1_n=3 (92.83)
Italian–French	svm-linear_ C=0.1_n=3 (94.92)	svm-rbf_C=10_ n=3 (94.77)	svm-rbf_C=10_ n=2 (94.70)	svm-linear_C=1_ n=3 (94.61)	svm-linear_ C=0.1_n=2 (94.48)
Italian–Romanian	svm-linear_ C=0.1_n=3 (94.44)	svm-linear_C=1_ n=3 (94.29)	svm-rbf_C=10_ n=2 (94.27)	svm-linear_ C=0.1_n=2 (94.11)	svm-linear_ C=10_n=3 (94.10)
Portuguese–French	svm-linear_ C=0.1_n=3 (94.64)	svm-linear_ C=0.1_n=2 (94.31)	svm-rbf_C=1_n=2 (94.29)	svm-rbf_C=10_ n=3 (94.28)	svm-rbf_C=10_ n=2 (94.20)
Portuguese–Italian	svm-linear_ C=0.1_n=3 (95.37)	svm-rbf_C=10_ n=3 (95.26)	svm-rbf_C=1_n=2 (95.18)	svm-linear_ C=0.1_n=2 (95.13)	svm-rbf_C=10_ n=2 (95.12)
Portuguese–Romanian	svm-linear_ C=0.1_n=3 (93.62)	svm-linear_ C=0.1_n=2 (93.56)	svm-rbf_C=10_ n=2 (93.44)	svm-rbf_C=1_n=2 (93.29)	svm-rbf_C=10_ n=3 (93.18)
Spanish–French	svm-linear_ C=0.1_n=3 (94.54)	svm-rbf_C=10_ n=3 (94.38)	svm-linear_ C=0.1_n=2 (94.28)	svm-rbf_C=1_n=2 (94.26)	svm-rbf_C=10_ n=2 (94.26)
Spanish–Italian	svm-linear_ C=0.1_n=3 (94.86)	svm-rbf_C=10_ n=3 (94.63)	svm-rbf_C=10_ n=2 (94.52)	svm-linear_ C=0.1_n=2 (94.49)	svm-linear_C=1_ n=3 (94.36)
Spanish–Portuguese	svm-rbf_C=1_n=2 (93.32)	svm-rbf_C=10_ n=2 (93.22)	svm-linear_ C=0.1_n=3 (93.22)	svm-rbf_C=1_n=1 (93.20)	svm-linear_ C=0.1_n=2 (93.13)
Spanish–Romanian	svm-linear_ C=0.1_n=3 (93.81)	svm-linear_ C=0.1_n=2 (93.73)	svm-rbf_C=1_n=2 (93.73)	svm-rbf_C=10_ n=2 (93.57)	svm-rbf_C=1_n=3 (93.42)

Table 1: Top 5 base classifiers per Romance language pair for the LaBSE static + ℓ_2 -normalised semantic, on the Levenshtein negative-selection split. Configurations are ranked by 3-fold cross-validated accuracy (%) on the training split – the criterion used to select the ensemble’s members. Grapheme-only (no_phonemes) configurations; the reported ensemble stacks the five listed for each pair.

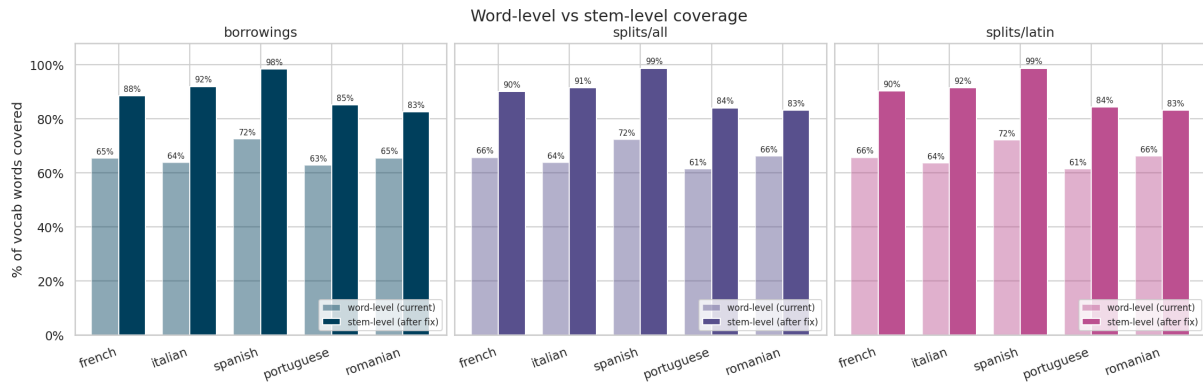


Figure 17: Contextual embeddings coverage for each dataset

		random						levenstein					
		LaBSE			XL-Lexeme			LaBSE			XL-Lexeme		
Pair	Track	base	stat	ctx	stat	ctx	base	stat	ctx	stat	ctx		
fr-ro	Ens.	98.6	99.0	98.6	98.6	98.7	90.6	93.0	92.3	91.5	92.3		
	Trf.	94.4	95.3	95.2	94.9	96.1	87.5	89.0	89.5	88.8	90.2		
it-fr	Ens.	98.9	99.0	99.0	99.2	99.0	93.5	95.6	94.8	94.6	94.8		
	Trf.	94.7	95.8	95.2	95.4	96.1	90.2	90.5	91.3	90.1	92.0		
it-ro	Ens.	98.6	98.9	98.7	98.6	98.6	93.8	95.0	95.7	94.9	95.6		
	Trf.	94.1	94.7	95.1	94.5	95.3	90.0	90.3	90.9	90.8	91.5		
po-fr	Ens.	98.1	98.8	98.5	98.1	98.7	93.8	95.8	95.5	94.3	95.6		
	Trf.	92.7	94.4	93.5	93.9	94.7	89.5	91.3	90.2	90.2	91.5		
po-it	Ens.	99.0	99.3	99.0	99.1	99.2	94.2	95.8	95.8	94.9	95.9		
	Trf.	95.2	95.8	95.7	95.5	96.2	89.7	89.7	91.1	89.9	91.1		
po-ro	Ens.	98.8	98.7	99.0	98.8	98.8	91.8	94.1	92.9	93.1	93.5		
	Trf.	93.5	95.2	94.5	94.0	94.8	87.2	89.1	88.9	88.8	90.4		
sp-fr	Ens.	98.6	99.2	99.2	98.6	98.8	93.4	95.3	95.4	94.6	95.9		
	Trf.	93.8	94.8	94.0	94.3	95.0	88.5	91.1	90.9	89.7	91.8		
sp-it	Ens.	98.6	99.0	99.1	98.7	98.9	93.9	96.0	95.5	94.9	95.2		
	Trf.	94.3	95.0	95.1	94.7	95.7	89.4	89.8	90.6	89.7	91.6		
sp-po	Ens.	98.2	98.6	98.3	98.4	98.3	91.3	94.0	92.8	92.6	92.2		
	Trf.	92.9	94.8	93.6	93.3	94.2	84.7	88.3	88.4	87.4	88.6		
sp-ro	Ens.	98.5	99.0	98.7	98.7	98.6	90.6	94.5	94.2	93.1	93.8		
	Trf.	93.2	94.9	94.6	93.5	95.2	87.8	89.6	90.5	88.9	90.7		

Table 2: Cognate classification test accuracy (%), no_phonemes setting. *base* = no semantic features; *stat* / *ctx* = static / contextual embeddings. Every language pair is split into its ensemble (Ens.) and transformer (Trf.) row. Best model per language pair and selection in bold.

		random						levenstein					
		LaBSE			XL-Lexeme			LaBSE			XL-Lexeme		
Pair	Track	base	stat	ctx	stat	ctx	base	stat	ctx	stat	ctx		
fr-ro	Ens.	98.0	98.8	98.4	98.1	98.8	89.7	92.3	91.8	91.2	92.4		
	Trf.	93.3	96.5	94.4	94.9	95.0	86.6	88.4	88.8	88.1	89.4		
it-fr	Ens.	98.5	98.9	98.7	98.7	98.7	93.2	95.1	94.7	94.1	94.4		
	Trf.	94.3	97.1	94.8	95.4	95.5	87.4	90.3	89.3	89.2	90.5		
it-ro	Ens.	98.0	98.6	98.5	98.0	97.8	93.9	95.2	95.2	95.2	95.6		
	Trf.	93.1	95.9	93.9	92.5	94.5	89.2	91.1	90.4	91.2	91.3		
po-fr	Ens.	97.6	98.8	98.2	98.1	98.7	92.9	95.2	94.5	93.8	94.6		
	Trf.	92.9	96.1	93.2	93.6	94.4	88.8	90.4	90.0	90.5	91.1		
po-it	Ens.	98.7	99.2	98.8	98.8	99.1	94.1	95.4	95.2	94.7	94.9		
	Trf.	94.4	97.4	94.4	95.5	95.6	88.9	91.8	90.4	90.6	91.4		
po-ro	Ens.	97.8	98.2	97.9	97.8	98.3	91.2	93.3	93.6	92.9	93.2		
	Trf.	92.5	95.8	92.6	93.3	94.3	87.2	89.2	89.2	89.3	90.6		
sp-fr	Ens.	98.3	98.7	98.6	98.2	98.3	93.1	95.3	95.2	94.3	95.5		
	Trf.	91.9	96.1	93.2	91.9	93.4	88.4	90.0	90.2	88.8	91.2		
sp-it	Ens.	98.4	98.9	98.7	98.5	98.5	94.1	95.8	95.1	94.7	95.0		
	Trf.	93.3	96.6	94.4	95.0	94.9	89.1	91.6	90.2	90.9	91.2		
sp-po	Ens.	97.9	98.5	98.3	98.2	98.3	91.2	93.4	93.1	92.5	93.2		
	Trf.	91.4	95.7	92.7	92.2	93.3	84.2	88.5	87.0	87.7	88.1		
sp-ro	Ens.	97.9	98.5	98.4	98.3	98.3	90.8	93.7	93.5	91.9	93.0		
	Trf.	92.0	95.9	93.8	93.6	94.0	87.1	89.4	89.9	88.7	90.5		

Table 3: Cognate classification test accuracy (%), only_phonemes setting. *base* = no semantic features; *stat* / *ctx* = static / contextual embeddings, both ℓ_2 -normalised. Every language pair is split into its ensemble (Ens.) and transformer (Trf.) row. Best model per language pair and selection in bold.

Model	Representation	Language pair										Avg.
		it-fr	it-ro	pt-fr	pt-it	pt-ro	es-fr	es-it	es-pt	es-ro	fr-ro	
Ens. baseline	Graphic+Phonetic	84.9	80.8	86.5	69.2	78.4	83.3	80.3	84.0	69.4	72.7	79.0
	Graphic	82.1	79.5	82.3	57.8	49.8 [†]	75.4	74.8	72.8	49.2 [†]	71.6	69.5
	Phonetic	84.3	80.4	84.0	48.8 [†]	87.4	81.7	77.3	83.0	66.0	71.2	76.4
Trf. baseline	Graphic	82.2	99.2	76.7	96.4	49.8 [†]	70.6	87.2	73.8	49.2 [†]	45.7 [†]	73.1
	Phonetic	78.5	74.2	73.0	48.8 [†]	49.8 [†]	68.8	74.2	71.4	49.2 [†]	53.3	64.1
Ens. +labse-all	Graphic+Phonetic	86.4	81.5	43.2 [†]	69.4	49.8 [†]	46.2 [†]	84.6	86.2	83.8	73.3	70.4
	Graphic	84.2	81.7	83.6	63.7	49.8 [†]	84.2	79.5	85.8	81.1	72.5	76.6
	Phonetic	85.4	80.7	85.2	67.5	49.8 [†]	86.4	83.1	84.7	84.6	70.1	77.8
Trf. +labse-all	Graphic	85.0	99.4	73.7	95.6	49.8 [†]	65.2	86.7	78.0	57.6	56.7	74.8
	Phonetic	79.3	76.1	80.1	48.8 [†]	49.8 [†]	72.4	73.8	76.9	53.6	46.6	65.7
Ens. +labse-ctx	Graphic+Phonetic	86.2	80.6	86.6	48.8 [†]	49.8 [†]	84.2	80.6	85.2	80.1	72.7	75.5
	Graphic	80.6	80.0	83.9	48.8 [†]	49.8 [†]	46.2 [†]	47.7 [†]	82.6	49.2 [†]	72.7	64.1
	Phonetic	85.4	80.8	85.4	48.8 [†]	49.8 [†]	84.3	47.7 [†]	84.4	81.9	70.6	71.9
Trf. +labse-ctx	Graphic	84.4	99.2	71.1	95.1	49.8 [†]	70.7	85.7	78.8	68.7	59.7	76.3
	Phonetic	78.9	75.2	77.4	48.8 [†]	49.8 [†]	71.2	74.2	79.4	49.2 [†]	54.3	65.8
Ens. +xl-lexeme-all	Graphic+Phonetic	85.5	82.6	86.1	48.8 [†]	49.8 [†]	84.6	81.1	86.2	49.2 [†]	72.9	72.7
	Graphic	84.2	80.5	83.5	48.8 [†]	66.5	83.1	47.7 [†]	80.0	49.2 [†]	72.2	69.6
	Phonetic	84.3	80.8	85.6	69.3	49.8 [†]	46.2 [†]	47.7 [†]	85.6	49.2 [†]	70.5	66.9
Trf. +xl-lexeme-all	Graphic	84.1	99.2	77.2	95.1	49.8 [†]	72.9	86.1	78.2	67.2	58.0	76.8
	Phonetic	79.7	74.6	75.4	48.8 [†]	49.8 [†]	56.1	77.1	78.2	59.8	56.4	65.6
Ens. +xl-lexeme-ctx	Graphic+Phonetic	86.0	80.7	87.5	71.2	78.4	86.8	47.7 [†]	86.4	49.2 [†]	73.1	74.7
	Graphic	86.0	80.8	84.8	48.8 [†]	74.8	80.3	47.7 [†]	83.9	78.4	71.8	73.7
	Phonetic	85.6	80.5	86.4	70.3	49.8 [†]	46.2 [†]	81.9	86.3	86.5	70.3	74.4
Trf. +xl-lexeme-ctx	Graphic	83.4	99.2	77.3	95.6	49.8 [†]	76.4	88.8	75.4	64.1	53.0	76.3
	Phonetic	77.4	77.0	78.1	48.8 [†]	49.8 [†]	64.4	77.1	73.8	70.1	61.5	67.8

Table 4: Cognate-borrowing classification: test macro F_1 (%) across all language pairs. **Ens.** = ensemble, **Trf.** = transformer; representation is the input signal used (Graphic = orthographic, Phonetic = phonetic, Graphic+Phonetic = both). [†] marks a run that collapsed to predicting a single class; those scores are shown as they are and still count towards Avg. The best score in each column is shown in **bold**.

Model	Representation	Language pair						Avg.
		it-fr	pt-fr	pt-it	es-fr	es-it	es-pt	
Ens. baseline	Graphic+Phonetic	70.5	49.3	47.5	77.4	67.5	7.7	53.3
	Graphic	70.9	49.3	62.1	64.6	68.6	77.2	65.5
	Phonetic	71.9	49.3	47.5	75.7	63.6	7.7	52.6
Trf. baseline	Graphic	97.6	61.9	90.8	71.6	100.0	47.8	78.3
	Phonetic	61.5	69.4	72.3	75.8	63.6	47.8	65.1
Ens. +labse-all	Graphic+Phonetic	70.7	49.3	47.5	68.0	72.7	7.7	52.6
	Graphic	72.3	49.3	47.5	44.8	71.3	7.7	48.8
	Phonetic	68.5	49.3	47.5	72.6	66.2	7.7	51.9
Trf. +labse-all	Graphic	97.6	61.9	90.8	71.8	100.0	47.8	78.3
	Phonetic	58.9	61.9	47.5	74.1	67.6	47.8	59.6
Ens. +labse-ctx	Graphic+Phonetic	72.0	49.3	47.5	69.1	76.5	7.7	53.7
	Graphic	70.5	49.3	47.5	71.3	63.5	7.7	51.6
	Phonetic	73.0	49.3	73.3	68.0	64.9	7.7	56.0
Trf. +labse-ctx	Graphic	98.4	71.7	90.8	69.4	100.0	47.8	79.7
	Phonetic	58.6	61.9	70.0	71.8	66.2	47.8	62.7
Ens. +xl-lexeme-all	Graphic+Phonetic	72.3	49.3	47.5	44.8	70.0	57.1	56.8
	Graphic	68.8	49.3	47.0	59.0	64.9	7.7	49.5
	Phonetic	72.1	49.2	47.5	44.8	62.3	7.7	47.2
Trf. +xl-lexeme-all	Graphic	98.8	60.4	87.1	70.9	100.0	47.8	77.5
	Phonetic	60.3	71.7	47.5	63.8	66.2	47.8	59.6
Ens. +xl-lexeme-ctx	Graphic+Phonetic	72.4	49.3	47.5	77.3	71.3	7.7	54.2
	Graphic	73.8	49.3	73.3	63.4	74.0	7.7	56.9
	Phonetic	74.5	49.3	47.5	75.8	64.8	7.7	53.3
Trf. +xl-lexeme-ctx	Graphic	98.0	61.9	87.1	63.1	98.7	47.8	76.1
	Phonetic	62.6	61.9	63.1	44.8	58.3	47.8	56.4

Table 5: Borrowing direction classification: test macro F_1 (%) across all language pairs. **Ens.** = ensemble, **Trf.** = transformer; representation is the input signal used (Graphic = orthographic, Phonetic = phonetic, Graphic+Phonetic = both). The best score in each column is shown in **bold**.

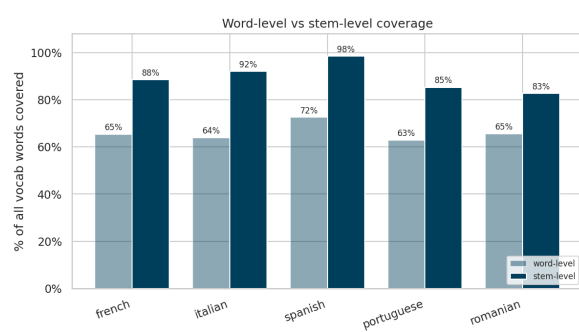


Figure 18: Contextual embeddings coverage - all datasets