It takes two to borrow: a donor and a recipient. Who's who?

Liviu P. Dinu[♠,♡] **Ana Sabina Uban**[♠,♡] **Anca Dinu**[♣,♡]
Bogdan Iordache[♡] **Simona Georgescu**[♣,♡] **Laurențiu Zoicaș**[♣,♡]

University of Bucharest, [♠] Faculty of Mathematics and Computer Science,
[♣]Faculty of Foreign Languages and Literatures, [♡]HLT Research Center
{ldinu, auban,}@fmi.unibuc.ro, iordache.bogdan1998@gmail.com,
{anca.dinu, simona.georgescu, laurentiu.zoicas}@lls.unibuc.ro

Abstract

We address the open problem of automatically identifying the direction of lexical borrowing, given word pairs in the donor and recipient languages. We propose strong benchmarks for this task, by applying a set of machine learning models. We extract and publicly release a comprehensive borrowings dataset from the recent RoBoCoP cognates and borrowings database (Dinu et al., 2023) for five Romance languages. We experiment on this dataset with both graphic and phonemic representations and with different features, models and architectures. We interpret the results, in terms of F1 score, commenting on the influence of features and model choice, of the imbalanced data and of the inherent difficulty of the task for particular language pairs. We show that automatically determining the direction of borrowing is a feasible task, and propose additional directions for future work.

1 Introduction and Related Work

Lexical borrowing is an essential witness to socio-cognitive and cultural evolution of mankind. Beside its immediate implications in linguistic phylogeny (Mallory and Adams, 2006; Dunn, 2015; Alekseyenko et al., 2012), lexical borrowing, as primary evidence for contact between different communities, can be used as a source on history (Heggarty, 2015), anthropology (Pakendorf, 2015), or sociology (Epps, 2014), providing important clues to linguistic changes in the past (Campbell, 2003). While offering significant proofs of migration or geographic expansion (Mallory, 1992; Campbell, 2013), the lexical data attesting linguistic contact corroborate the archaeological inventory (cf. (Heggarty, 2015)) and provide the basis for 'linguistic paleontology' or 'socio-cultural reconstruction' (Epps, 2014).

While there is extensive literature (Ciobanu and Dinu, 2014; Rama, 2016; Jäger et al., 2017; Rama et al., 2018; Fourrier and Sagot, 2022; Dinu et al., 2023) on the related problem of cognate identification both in classical and computational linguistics, automatic borrowing detection is much less studied (Jäger, 2019). In our present study we approach the new task of automatic borrowing direction detection.

Determining the direction of lexical borrowing - still one of the open questions in historical linguistics (List, 2019) - is crucial for an accurate inference of information for the above-mentioned domains: given a pair of languages, L_1 and L_2, and a word that is present in a very similar form in both languages, without it being inherited in one or both, and with no traces of a third language from which it could have entered in both L_1 and L_2, the question that arises in not a few cases is which one of the two languages is the donor. The correct identification of the borrowing direction helps us decide which community had primacy over a certain object or concept (setting thus the basis of an archaeology of concepts), just as it enables us to trace back the type of social interaction among groups, by revealing the dominant language (cf. (Miller and List, 2023)) as donor of core borrowings (Haspelmath, 2009). Pointing out the donor language of a certain loanword is also vital for a correct etymological analysis, narrowing the track for the search of the word's origin.

Lexical borrowing has been studied from several points of view: leaving aside the traditional approaches which focused mostly on the phonemic or morphological adaptation of loanwords (Deroy, 1956; Haugen, 1950), more recent – still classical – approaches either aim at detecting the semantic areas which are more permeable to borrowings (Tadmor, 2009), or propose an interpretation of the socio-economical relations between languages based on the lexical reflection of their linguistic contact (Epps, 2014). The new research methods brought in by computational linguistics have shed a new light on the issue (Jäger, 2019), while the

Findings of the Association for Computational Linguistics: ACL 2024, pages 6023–6035

search for an automatic procedure to be used in the analysis of borrowings has led to breakthrough works that address difficult problems like the distinction between borrowings and virtual cognates (Ciobanu and Dinu, 2015; Tsvetkov et al., 2015; Dinu et al., 2024), determination of monolexical borrowings (Miller et al., 2020), computational etymology of borrowings (Wu et al., 2021) or discrimination between inherited and borrowed words from the same source language (Cristea et al., 2021b). However, to the best of our knowledge, the possibility of determining the direction of borrowing has not yet been assessed.

This problem has various facets and implications. It is not infrequent the case in which a word that exists simultaneously in L_1 and L_2 is registered in the etymological dictionaries of L_1 as coming from L_2 and in the dictionaries of L_2 as a borrowing from L_1. The difficulty of identifying the direction of borrowing is not limited to ancient, non-documented languages, or to under-studied linguistic families. Even in extensively researched linguistic domains, like the Romance languages family, we may still encounter lexicographical controversies regarding the donor-recipient relationship. For instance, a long-standing etymological controversy concerns the origin of Sp. *tacaño* / It. *taccagno* 'stingy', attributed by the Italian dictionaries to Spanish (DELI2, Mauro[1]) and vice-versa (DLE); the same situation emerges for Sp. *tozo* 'dwarf', interpreted as a borrowing of It. *tozzo* 'thick and short' (DLE), while It. *tozzo* is labeled as borrowed from Sp. *tozo* (DELI2).[2]

The traditional approach to such uncertainties (cf. (Campbell, 2013)) could only make use of two criteria: identifying in a word unusual sounds or infrequent phonological structures for the language in question, implying that the word may be borrowed, and looking for clues in the phonological history of the supposed donor and recipient languages (sound changes undergone by one language and absent from the other may be an indicator of the relationship between them). However, these criteria prove to be insufficient in certain cases (see above examples). The debate does not only take place at a linguistic level, but there are also cases where two languages dispute their primacy over a certain concept, be it part of the social, cultural, gastronomic, artistic or military domain. Where the traditional methods could not provide a certain answer, a computational approach could help solve long-standing disputes.

The approach we propose is only a first step forward, while trying to meet the "challenge" posed by List (2019): "Whether it will be possible to identify even the direction of borrowings, when developing these methods further, is an open question."

Starting with these remarks, our main contributions are:

- We investigate whether the direction of borrowing for a pair of borrowings can be automatically identified based on their graphic and phonemic forms. More specifically, our task is as follows: given a pair of borrowings (w_1, w_2) in two different Romance languages (L_1, L_2) respectively, we want to determine whether word w_1 in language L_1 was borrowed from word w_2 in L_2 or word w_2 in language L_2 was borrowed from word w_1 in L_1. We run several experiments, and we propose strong benchmarks for this task, by applying a set of machine learning models (using various feature sets and architectures) on any two pairs of Romance languages.

- We explore what kind of features and machine learning models are more effective for accurately detecting the direction of borrowing. We also investigate whether graphic or phonemic similarities between words are more relevant.

- We discern for which languages in the Romance family it is more challenging to detect the direction of borrowings.

The rest of the paper is organized as follows. In Section 2 we present the database that we use, and

[1]Italian: *Il dizionario della lingua italiana De Mauro*, dizionario.internazionale.it.
Spanish: *Diccionario de la lengua española* published by *Real Academia Española*, lema.rae.es/drae.
Portuguese: *Dicionário infopédia da Língua Portuguesa*, published by Porto Editora, www.infopedia.pt/lingua-portuguesa.
French: *Trésor de la Langue Française informatisé* published by *Centre National de Ressources Textuelles et Lexicales*, www.cnrtl.fr

[2]Similar dissensions are to be found between French and Italian (e.g. Fr. *pantalon* 'pants' < It. *pantalone*, cf. TLFi / It. *pantalone* < Fr. *pantalon*, cf. DELI2; the same for Fr. *bastion* / It. *bastione* 'bastion'), Spanish and Portuguese (e.g. Pt. *caramelo* < Sp. *caramelo*, cf. Infopedia/ Sp. *caramelo* 'candy' < Pt. *caramelo*, cf. DLE; the same for Pt. *carambola* / Sp. *carambola* 'star fruit'), or Italian and Portuguese (It. *caravela* 'caravel' < Pt. *caravela*, cf. DELI / Pt. *caravela* < It. *caravela*, cf. Infopedia). The same uncertainty floats over a term as culturally significant as *Baroque*, explained in the French lexicography as borrowed from Portuguese, while the Portuguese dictionaries see it as a loanword from French.

offer details about the processing steps involved, while in Section 3 we introduce our approach for the automatic detection of borrowing direction and present the experiments. Extensive results and error analyses are presented in Section 4. The last section is dedicated to final remarks and future works.

2 Data

To obtain a dataset of borrowings for the five Romance languages (Romanian, Italian, Portuguese, French, and Spanish), we used the RoBoCoP cognate and borrowing database (Dinu et al., 2023). Our choice is motivated by its quality and high-coverage. Since the sources of this database are dictionaries of five Romance languages, RoBoCoP has a wide coverage, including all the words currently in use for five languages. It is one of the most comprehensive, inclusive, and complex databases of Romance related words. Moreover, because it was obtained in a computer-assisted manner and manually checked, it minimizes the noise and is reliable, in contrast to other resources created from Wiktionary (Meloni et al., 2021) or from automated translations (Dinu and Ciobanu, 2014). Based on the etymologies available in RoBoCoP, we extracted lists of borrowed words across five Romance languages, for each language pair (the database will be available for research purposes upon request).

Formally, for any triplet in the RoBoCoP database <*u, e, L_1*> where *u* is a word in Romance language L_1 and *e* is its etymon, for which *e* is a word in a Romance language L_2, we add the tuple <*u, e*> to the list of borrowing pairs from language L_2 to language L_1). We distinguish between words borrowed by language L_1 from language L_2 and the words borrowed by language L_2 from language L_1. We thus obtain twenty lists of borrowed words, for each ordered language pair <L_1, L_2>, where L_1 and L_2 are distinct languages in the set of the five Romance languages considered.

To the best of our knowledge, these are the most comprehensive lists of borrowings for the five Romance languages publicly available in digital format. Table 1 depicts the total number of borrowings pairs for each ordered pair of languages in the extracted dataset.

As a direct result of using the RoBoCoP database, the borrowings in our dataset are cleaned of potential noise resulted from the data collection process, but they preserve the accents, diacritics and any other characters that are part of the orthography of the words and etymons. The only cleaning operation we perform on the graphic form of the words is to remove the accents that mark the stress of the word, but are not part of the orthography of the language and, therefore, do not represent relevant information for our task. The phonemic representation is left as it is[3].

	Ro	It	Es	Pt	Fr
Ro		3,135	209	102	33,311
It	4		376	62	1,981
Es	0	394		104	1,366
Pt	1	558	1,097		2,369
Fr	1	915	324	81	

Table 1: Dataset statistics: number of borrowing pairs per pair of languages, in both directions. The number in a cell represents the number of borrowings from the language on the same column to the language on the same line.

3 Experiments and Methodology

In this section we report experiments on our dataset containing the five Romance languages, by training several models to automatically recognize the direction of borrowing, given a pair of words in two languages where one was borrowed from the other. Given a language pair (L_1, L_2) and a set of word pairs

$$WP = \{(w_1, w_2) \mid w_1 \in L_1, w_2 \in L_2, \text{ and } w_1 \text{ borrowed from } L_2 \text{ or } w_2 \text{ borrowed from } L_1\},$$

the task is to determine the actual direction of borrowing for any given pair (w_1, w_2).

We use two main approaches: ensemble of machine learning classifiers and transformer based models. In all scenarios, we train bilingual models that learn to identify the borrowing direction for a given pair of languages. We experiment with both the graphic and phonemic representations of the borrowings and with different types of features, which are described in detail below. Note that, the order in which the words from a pair are fed to these models is constant across training and evaluation (e.g. for Italian-French, the Italian word always precedes the French word in the example

[3]The phonemic representation was obtained in RoBoCoP by using the eSpeak library https://github.com/espeak-ng/espeak-ng

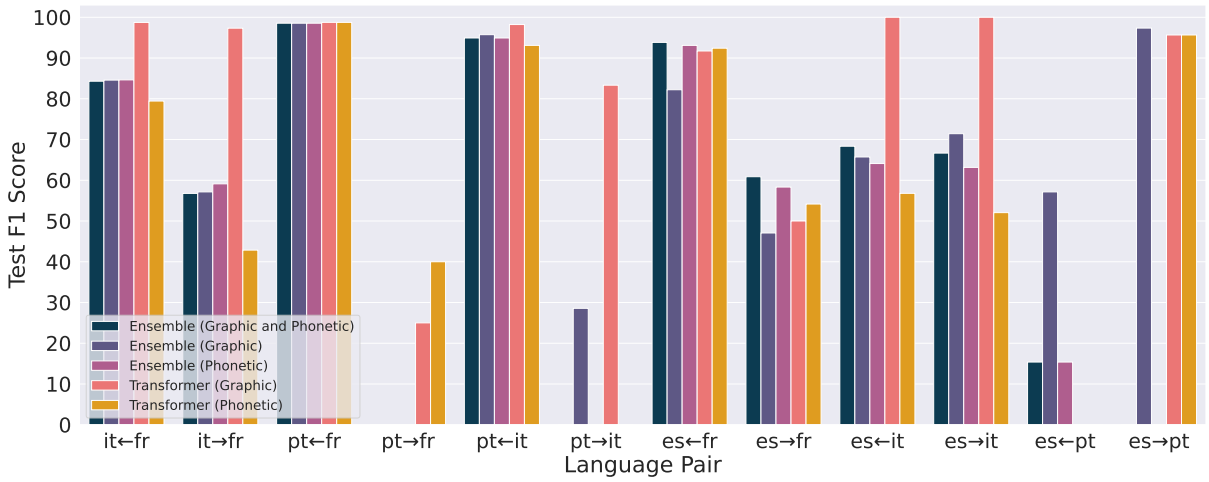

Figure 1: F1-scores for all models and language pairs considered, separately for each borrowing direction. The direction of the arrow represents the borrowing direction considered as the positive class in computing the F1-score. Missing bars from the plot correspond to null scores.

pairs). These orders were chosen arbitrarily, but because of the symmetry of the proposed approaches, they should not affect the results in any way.

3.1 Experimental settings

We use a $90\% : 10\%$ stratified split to generate train and test sets, which are initially shuffled. For a language pair, the ratio between the number of examples in one borrowing direction vs. the other is the same for both the training and testing splits. The unusually high percent of training data is motivated by the fact that the data was heavily imbalanced due to asymmetric direction of borrowings for some language pairs. For instance, the number of borrowings from Portuguese to Spanish is 104, while from Spanish to Portuguese is 1097. An extreme case of such asymmetry is Romanian, which borrowed massively from other Romance languages (especially from French and Italian), but barely loaned any words (0 borrowings to Spanish, 1 to French and Portuguese and 4 to Italian). Consequently, we had to exclude pairs including Romanian from the experiments.

3.2 Metrics

We compute accuracy as well as F1-score to measure the performances of each model and feature set for every language pair. Since many of the language pairs (even excluding Romanian) are still imbalanced in terms of the two directions of borrowing, for a more in-depth understanding of the performance of the models, we include performance metrics (precision, recall and F1-score) for both classes corresponding to the two directions of borrowings.

3.3 Features

Some of our models rely on handcrafted features extracted from the graphic and phonemic forms, while others are deep models trained directly on the raw representations. For the former category, feature extraction is performed by computing the alignments returned via the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). Previous attempts from the literature employed successfully these features for discriminating between cognate and non-cognate pairs (Ciobanu and Dinu, 2019), although it is worth noting that, unlike the previous studies, we also extract features from the alignment of the phonemic representation, as opposed to the mere graphic one. The process is similar (i.e. we align the phonemes, instead of the letters). After computing the alignment, we select n-grams around mismatches (i.e. insertions, deletions, substitutions). More precisely, for a given n we extract all i-grams with the length $i \leq n$. We exemplify the process for the graphic forms of the Portuguese-Italian pair `(empostar, impostare)`. The computed alignment is

```
$empostar-$
$impostare$
```

where `$` marks the start and the end of the alignments and `-` represents an insertion/deletion; for $n = 2$, the i-gram misalignment features are: `-$>e$`, `em>im`, `e>i`, `r->re`, `->e`, `$e>$i`. In order to vectorize these features, we use a binary

bag of words.

3.4 Models

3.4.1 Ensemble Models

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. The individual models are evaluated and selected via 3-fold cross validation on the training split, but note that the final ensemble model is trained using all of the training examples and its performance is ultimately assessed on the test split. The size of the misalignment n-grams was one of the hyper-parameters used for selecting the best models ($1 \leq n \leq 3$).

For all the models, we employ the implementation provided by the *scikit-learn* library (Pedregosa et al., 2011). The models are trained using the graphic features, the phonemic features, and both, to assess if any category of features outperforms the other, or if their combination is more favorable.

3.4.2 Transformer Models

Our second approach uses **transformer models** (Vaswani et al., 2017). 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. 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.

4 Results and Error Analysis

In this section we present the main results obtained and we perform an in-depth analysis of the results and errors.

4.1 Results

We report the results in two complementary ways, for each language pair: by taking into consideration the borrowing direction (Figure 1, Table 4), and also by computing macro-average results across the two borrowing directions (Figure 2, Table 3). In other words, after training the models we computed three types of metrics: two where we consider either of the directions as the positive class, and one that averages over the former two, the final objective being to assess what direction is harder to predict.

In Figure 1 we represent the F1 scores for each direction of borrowing separately, by considering each of the two directions as the positive class in turn. The plots illustrate results for each language pair and direction of borrowings obtained with both types of models and features. The difference in model type is doubled by the difference in features used: the ensemble models use alignment features, whereas transformers use sequence representations - in both settings with graphic and phonemic variants.

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.

The highest F1-score is obtained for detecting the direction of borrowing from Spanish to Italian and from Italian to Spanish. The graphical transformer generates in these cases perfect predictions on the test set, obtaining an F1-score of 1. A plausible explanation for this perfect result might be that the two languages have the most training data and they are balanced with regard to the borrowing direction (see Table 1). In order to validate this high performance, and check that the perfect score was not a fluke generated by some lucky split, we re-run the training multiple times on different dataset splits. We were able to see that in 3 out of 5 split scenarios, the model was able to correctly predict the borrowing direction for all 77 training pairs, whereas in the other 2 was misclassifying exactly one pair.

In the remarkable case of Portuguese borrowings from French, all the models obtain F1 scores over

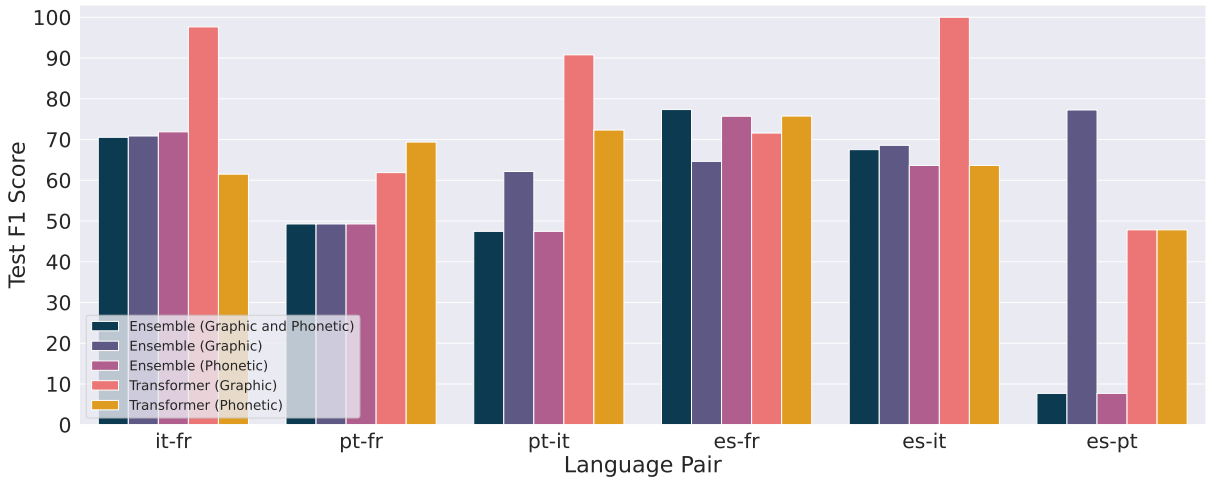

Figure 2: F1-scores for all models and language pairs considered, macro-averaged across the two borrowing directions for each language pair.

	Correct	Incorrect
It←Fr	avviso ← avis	chic → chic
It→Fr	futurismo → futurisme	melica ← mélique
Pt←Fr	aviation ← aviação	-
Pt→Fr	banean → banian	selva ← selve
Pt←It	alegro ← allegro	bora → bora
Pt→It	calau → calào	lusiadas ← lusiade
Es←Fr	bufanda ← bouffante	-
Es→Fr	yucca → yucca	castañetas ← castagnette
Es←It	cartucho ← cartoccio	-
Es→It	cumpleaños → compleànno	-
Es←Pt	ollar ← ollo	mandarín → mandarim
Es→Pt	caramelo → caramelo	-

Table 2: Examples of model predictions for the best performing model, for each language pair. The "→" sign represents the borrowing direction from left to right, while the "←" sign represents the borrowing direction from right to left. The "-" sign in some cells from the Incorrect column indicates that there were no cases of misclassification.

0.97, which indicates the existence of consistent features that discriminate between the two directions of borrowing. We have also noticed that over 85% of the incorrect predictions of all the models and settings are the same for this pair of languages. Moreover, for all pairs of languages, most of the errors of all models overlap considerably. This suggests that our approach is quite robust, since the same errors in different scenarios indicate inherent difficult cases.

Most of the asymmetries and weaker results appear for the language pairs where the imbalance is severe. Aside from borrowings from Romanian which we exclude, there are a few other cases with few examples per class, which are reflected in the models' performance: specifically, words borrowed from Portuguese to Italian, French, and Spanish are more difficult to correctly classify by the models, while borrowings in the opposite directions (where data is more abundant) are detected with high F1-scores. In cases with very low number of examples, such as for borrowings from Portuguese to the other languages, there are models which classify none of the examples correctly, obtaining null F1-scores, shown as a missing bar in the chart, as is the case with all three ensemble models for the Portuguese to French borrowings (only 81 pairs), and with the two transformer models for the Portuguese to Spanish borrowings (only 104 examples), and with the three phonemic models for the Portuguese to Italian borrowings (only 62 pairs). In addition, there is the case of borrowings from Spanish to

	Es				Pt				Fr			
	M	F1	Pr	Rec	M	F1	Pr	Rec	M	F1	Pr	Rec
It	TG	1	1	1	TG	.908	.908	.908	TG	.976	.974	.979
Es					EG	.772	.974	.700	EGP	.774	.942	.719
Pt									TP	.694	.823	.641

Table 3: Performance metrics for the best models in terms of macro-averaged F1-scores, precision and recall, for each language pair, and the models and settings for which they were obtained (TG = Transformer (Graphic); TP = Transformer (Phonemic); EG = Ensemble (Graphic); EGP = Ensemble (Graphic and Phonemic); M=Model; Pr=Precision; Rec =Recall).

Portuguese in which, despite abundant data, two models (Ensemble Graphic and Phonemic, and Ensemble Phonemic) obtain an F1 score of 0. This evidence suggests that phonemic information is confusing for the models and does not discriminate well between Spanish and Portuguese, a hypothesis which is also backed up by the results of the other borrowing direction, where phonemic models scored an F1 close to 0 (0.15).

In terms of features, the graphic features seem to generally outperform phonemic features, especially in the case of language pairs involving Italian, where the graphic-based transformer outperforms the other models drastically and consistently (see figure 2), but also in the case of Spanish Portuguese pair. On the other hand, the phonemic features outperform the graphical features in both model architectures for language pairs involving French: Portuguese-French, Spanish-French as well as French-Spanish and French-Portuguese, with the exception of French-Italian, for which the graphical features override the phonemic ones. This contrast between French and the other Romance languages might be explained by the fact that French has a poor correlation between the orthography and phonology of the language, so phonemic features add extra information and help in discriminating the direction of borrowing.

As for the best performing model, the transformer model on graphic word forms seems to be the most robust across language pairs and directions, being the best performing model in 7 out of 12 experiment scenarios.

We also report macro-average results across the two borrowing directions (showed in Figure 2 and Table 3). The best result is reported for Spanish-Italian pair, with an F1 score of 1, obtained with the graphic transformer model. The transformer model with graphic features obtains the best results for 3 language pairs, followed by the phonemic transformer (1 pair), graphic ensemble (1 pair) and ensemble with graphic and phonemic features (1 pair).

We additionally computed the average performance across language pairs for each of the 5 models, in terms of macro-averaged F1-score, in order to facilitate a direct comparison between models and settings. The best average F1-score is obtained for the graphic transformer (0.78), followed by the phonemic transformer (0.65), ensemble with phonemic and graphic features (0.648), graphic ensemble (0.53) and phonemic ensemble (0.52). While the average scores per model don't seem to be spectacular overall, if we consider only the best results for each language pair, the mean of the best F1-scores reaches 0.854. This happens because, for this particular task of historical linguistics, there is no universal model for all pairs of Romance languages, the best model being different for different pairs of languages. This finding contrasts with the results from another important task in historical linguistics, namely automatic cognate identification (Dinu et al., 2023), where it is reported that the best model for all the Romance languages pairs investigated is always the ensemble with graphic and phonemic features.

In Table 5 we show the most informative features for the ensemble models for each language pair and borrowing direction. The placement of the misalignments seems to vary across different language pairs, as we can see important features at the start of the words, the end, and, most frequently, strictly inside of the two words in the pair.

A few examples of predictions (both correct and incorrect classifications) of the best performing models for each language pair are shown in Table 2. One interesting example is that of the Portuguese *caramelo* and Spanish *caramelo*, which is a controversial case in lexicography: all models except for the ensemble based on phonemic features predict Portuguese *caramelo* is borrowed from Spanish *caramelo*, which is in accordance to the etymology

	It				Es				Pt				Fr			
	M	F1	Pr	Rec	M	F1	Pr	Rec	M	F1	Pr	Rec	M	F1	Pr	Rec
It					TG	1	1	1	TG	.833	.833	.833	TG	.987	.994	.979
Es	TG	1	1	1					EG	.571	1	.4	EGP	.938	.881	1
Pt	TG	.982	.982	.982	EG	.973	.948	1					TP	.987	.979	.995
Fr	TG	.973	.957	.989	EGP	.608	1	.437	TP	.4	.666	.285				

Table 4: Performance metrics for the best performing models in terms of F1-scores, precision and recall, for each language pair and borrowing direction, and the models and settings for which they were obtained (TG = Transformer (Graphic); TP = Transformer (Phonemic); EG = Ensemble (Graphic); EGP = Ensemble (Graphic and Phonemic); M=Model; Pr=Precision; Rec =Recall. The information in a cell represents the performance and the model for identifying the borrowing direction from the language on the same column to the language on the same line.

found in the RoBoCoP database. Another interesting case is that of Spanish *yuca* > French *yucca*, which is classified correctly only by the ensemble using both graphic and phonemic features.

4.2 Error Analysis

By looking at the misclassified cases, we observed as a possible source for error the use of an older form of words borrowed from one language or another. For instance, languages such as Spanish, Portuguese or Italian have not only borrowed from modern French, but also from old French. As a result, the machine did not recognize certain modern French terms as the origin of Portuguese, Italian or Spanish words (either because they have a different spelling today, or because they have fallen out of use, or because they are in very limited use). Here are some examples: It. *battifredo* < Old Fr. *berfroi* (contemporary Fr. *beffroi*); It. *berlengo* < Old Fr. *berlenc* (out of use); It. *brandistocco* < Fr. *brindestoc* (out of use and absent from TLFi).

It is also possible that some errors are due to homonymy, like in the case of Fr. *palanquin*, with two different sources, one Portuguese, the other Italian, or in the case of Fr. *morfil*, with two different sources, one Spanish, the other autochthonous.

Furthermore, the machine sometimes gives feminine forms as coming from another language, whereas the feminine is formed by suffixation, and the borrowing concerns only the masculine form of the noun (e.g. the pair Italian *buffalo* - French *bufflesse*; correct: *buffle*).

By looking at the misclassifications generated by the various experimental settings (i.e. using only graphic features, only phonemic ones, or a combination of both), we can infer that, in general, the ensemble models trained on both categories of features outperform their equivalents trained on only one category. For most language pairs, we observed that the multi-featured ensembles do not generate new errors on top of the ones displayed by their graphic-only or phonemic-only counterparts.

5 Conclusions

We obtained the first results, to the best of our knowledge, for the task of automatically predicting the direction of borrowing between pairs of words in four Romance languages (Italian, Spanish, Portuguese, and French), based on orthographic and phonemic features. At this stage, we designed the experiments specifically to limit the input to the models to only word forms, in order to understand how well machine learning can work for our purpose in this setup. Our results show that the task can be solved with machine learning with high performance (obtaining perfect predictions on the test set for certain language pairs), and that results can vary depending on the language pair. The average F1 score for the best models for all the 6 pairs of Romance languages analyzed was 85.4%

The challenge of this task formulated in this way is specifically to understand whether the problem can be solved without any additional information. The promising results already obtained show us that we can hope for even better success in identifying the direction of borrowing by including new features, such as chronological parameters (the oldest record of the words in the two languages), cultural ones (in which culture does the concept fit better), or semantic ones (Cristea et al., 2021a; Uban et al., 2021). The semantic criterion we expect to have an important effect on classification performance, given that, in cases of polysemous words, the number of meanings can elucidate the antiquity of a word in a given language: a rich semantic expansion of a lexeme in L_1 versus a poor or absent polysemy in L_2 is more eloquent than simple phonemic features. At a technical level, jointly learning some of the parameters for words in each language across language pairs could improve the models' capacity

	It	Es	Pt	Fr
It		il>ig, t->tt, l->li	az>as, z->ss, ab>av	at>ad, ta>de, -s>es
Es	ill>igl, et->ett, ll->gli		ur>or, zu>so, –>en	in>–, no>–, me>–
Pt	naz>nas, az->ass, nha>n-a	zur>sor, uru>oru, $zu>$so		du>–, re>–, es>–
Fr	ata>ade, ta$>de$, $-s>$es	llo>lle, lla>lle, arr>ar-	dub>—, $du>$–, ubr>—	

Table 5: Top 3 informative graphic alignment bigrams and trigrams according to χ^2 feature selection. N-grams are separated by commas, $>$ marks where the n-gram for the first word in the pair ends and where the n-gram for the second word begins, $-$ marks an insertion/deletion computed by the alignment algorithm. Bigrams are shown above the main diagonal, whereas trigrams are shown below it.

to encode the linguistic structures, and could help improve performance for language pairs with less training data.

Ethics 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 (Dinu et al., 2023). We will make our code public and the data available for research purposes upon request (so as to comply with copyright constraints in the original data sources).

Limitations

A limitation of our models' capacity to accurately predict the borrowing direction stems from the scarcity of data for some of the language pairs. However, this limitation is inherent to the task, rather than to the data source or to our solution: for example, in the case of Romanian, there are very few borrowings into other Romance languages, which makes it theoretically difficult to study borrowings from Romanian with statistical methods. One limitation in terms of the significance of our results does stem from the construction of the RoBoCoP database: it is the case of the borrowing pairs for which RoBoCoP reports as an etymon the old form of the word, which is then used in training the models (as discussed in Section 4).

Acknowledgements

This research is supported by the POCIDIF project in Action 1.2 "Romanian Hub for Artificial Intelligence" MySMIS no. 310483"

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.

Lyle Campbell. 2003. How to Show Languages are Related: Methods for Distant Genetic Relationship. In Brian D. Joseph and Richard W. Janda, editors, *The Handbook of Historical Linguistics*. Blackwell.

Lyle Campbell. 2013. *Historical linguistics*. Edinburgh University Press.

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.

Alina Maria Ciobanu and Liviu P. Dinu. 2015. Automatic discrimination between cognates and borrowings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 431–437. The Association for Computer Linguistics.

Alina Maria Ciobanu and Liviu P. Dinu. 2019. Automatic identification and production of related words for historical linguistics. *Computational Linguistics*, 45(4):667–704.

Alina Maria Cristea, Anca Dinu, Liviu P. Dinu, Simona Georgescu, Ana Sabina Uban, and Laurentiu Zoicas. 2021a. Towards an etymological map of Romanian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 315–323, Held Online. INCOMA Ltd.

Alina Maria Cristea, Liviu P. Dinu, Simona Georgescu, Mihnea-Lucian Mihai, and Ana Sabina Uban. 2021b. Automatic discrimination between inherited and borrowed Latin words in Romance languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2845–2855, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Louis Deroy. 1956. *Les emprunts linguistique*. Belles Lettres, Paris.

Liviu P. Dinu and Alina Maria Ciobanu. 2014. Building a dataset of multilingual cognates for the romanian lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 1038–1043. European Language Resources Association (ELRA).

Liviu P. Dinu, Ana Sabina Uban, Alina Maria Cristea, Anca Dinu, Ioan-Bogdan Iordache, Simona Georgescu, and Laurentiu Zoicas. 2023. Robocop: A comprehensive romance borrowing cognate package and benchmark for multilingual cognate identification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 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. 2024. 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, 20-25 May, 2024, Torino, Italy*, pages 12657–12667. ELRA and ICCL.

Michael Dunn. 2015. Language phylogenies. *The Routledge handbook of historical linguistics*, pages 190–211.

Patience Epps. 2014. Historical linguistics and sociocultural reconstruction. In *The Routledge Handbook of Historical Linguistics*, pages 579–597. London: Routledge.

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.

Martin Haspelmath. 2009. The typological database of the world atlas of language structures. *The Use of Databases in Cross-Linguistic Studies*, 41:283.

Einar Haugen. 1950. The analysis of linguistic borrowing. *Language*, 26:210–231.

Paul Heggarty. 2015. Prehistory through language and archaeology. In *The Routledge Handbook of Historical Linguistics*, pages 598–626. Routledge.

Gerhard Jäger. 2019. Computational Historical Linguistics. *Theoretical Linguistics | Volume 45: Issue 3-4*, 45: Issue 3–4.

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.

Johann-Mattis List. 2019. Open problems in computational historical linguistics. *Invited talk presented at the 24th International Conference of Historical Linguistics, Canberra, Australian National University*, pages 1–26.

James P Mallory. 1992. In search of the indo-europeans/language, archaeology and myth. *Praehistorische Zeitschrift*, 67(1):132–137.

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.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab antiquo: Neural proto-language reconstruction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4460–4473. Association for Computational Linguistics.

John E. Miller and Johann-Mattis List. 2023. Detecting lexical borrowings from dominant languages in multilingual wordlists. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2591–2597. Association for Computational Linguistics.

John E. Miller, Tiago Tresoldi, Roberto Zariquiey, César A. Beltrán Castañón, Natalia Morozova, and Johann-Mattis List. 2020. Using lexical language models to detect borrowings in monolingual wordlists. *PLOS ONE*, 15(12):1–23.

Saul B Needleman and Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Brigitte Pakendorf. 2015. Historical linguistics and molecular anthropology. *The Routledge handbook of historical linguistics*, pages 627–642.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

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.

Uri Tadmor. 2009. Loanwords in the world's languages: Findings and results. In *Loanwords in the world's languages. A comparative handbook*, pages 55–76. De Gruyter Mouton.

Yulia Tsvetkov, Waleed Ammar, and Chris Dyer. 2015. Constraint-Based Models of Lexical Borrowing. In *Proceedings of NAACL-HLT 2015*, pages 598–608.

Ana-Sabina Uban, Alina Maria Ciobanu, and Liviu P Dinu. 2021. Cross-lingual laws of semantic change. *Computational approaches to semantic change*, 6:219.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Winston Wu, Kevin Duh, and David Yarowsky. 2021. Sequence models for computational etymology of borrowings. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4032–4037. Association for Computational Linguistics.

A Appendix

A.1 Classification Experiments. Experimental Settings and Training Details

A.1.1 Ensemble Models

For all of our experiments involving ensemble architectures we needed a way to select the best standalone models out of a collection of classical machine learning algorithms. Our implementation is based on the *scikit-learn* Python library (version 1.2.0) and the employed models, along with the tested hyper-parameters are the following (note that if not specified otherwise, other hyper-parameters have been defaulted to the library's provided values):

- Multinomial Naive Bayes (`MultinomialNB`)
- Linear Support Vector Machine (`LinearSVC`): $C \in \{0.1, 1, 10\}$
- Linear classifiers trained using Stochastic Gradient Descent (`SGDClassifier`): loss $\in$ { hinge, log_loss, modified_huber, squared_hinge, perceptron }; class_weight = balanced; max_iter = $10,000$

All of these models were evaluated using a series of 3-fold cross-validation experiments (using only the training split of the dataset). Also, each model was trained on alignment features computed on the graphic representation, the phonemic representation, or both representations using various values for the length of the feature n-grams (we experimented with values for n between 1 and 3). Thus, for one model category with its set hyper-parameters, we performed 9 versions of the experiment for the various ways of selecting the features.

For each language pair (regardless of the attempted task: binary classification, or any of the two ternary classification tasks) we selected the configurations for the best performing five models and we ensembled them into a `StackingClassifier` and trained it on the whole training set. This is the final model used for predicting the test labels.

Variations of the ensemble were created (as reported in the main paper) were the selected base models were reduced to those trained only on a category of features. We also experimented with oversampling techniques for balancing the dataset, but for most situations they either performed similarly well, or even degraded the ensemble's performance.

Resources:

- CPU: Ryzen 5 3600X, 6 cores, 3.8 GHz
- Memory: 16GB RAM
- Time: $\approx 2h$ (for training, validation, and testing)

A.1.2 Transformer Models

Model Architecture. The character-level Transformer architecture was implemented using the `TransformerEncoder` implementation provided by the *torch* Python library (version 1.13.1) and it has the following structure:

- embedding size: 200
- hidden state size: 200
- number of attention heads: 8
- number of layers: 4
- dropout layer after positional encoding, probability: 0.2
- trainable parameters: $\approx 10^6$

Resources:

- CPU: Ryzen 5 3600X, 6 cores, 3.8 GHz
- Memory: 16GB RAM
- GPU: Nvidia RTX 2060 Super, 1470 MHz, 8GB VRAM
- Time: $\approx 1.5h$ (for training, validation, and testing)

Training Details. To avoid overfitting we compute the model's performance (i.e. the cross entropy loss) after each epoch on a validation subset randomly extracted from the training set, and if we see no improvement after the last epoch we decay the optimizer's learning rate with a γ coefficient. After a number of consecutive epochs without improvement (i.e. maximum "patience") we stop the training. The training parameters are the following:

- number of epochs: 50
- batch size: 64

- loss function: cross entropy loss
- optimizer: Adam
- initial learning rate: 10^{-3}
- γ: 0.6
- patience: 5 epochs