Pater incertus? There is a Solution: Automatic Discrimination between Cognates and Borrowings for Romance Languages

Liviu P. Dinu[♡♠], **Ana Sabina Uban**[♡♠], **Bogdan Iordache**[♡],
Alina Maria Cristea[♡], **Simona Georgescu**[♡♣], **Laurentiu Zoicas**[♡♣]
[♡]Human Languages Technologies Research Center, University of Bucharest
[♠] Faculty of Mathematics and Computer Science, University of Bucharest
[♣]Faculty of Foreign Languages and Literatures, University of Bucharest
{ldinu,auban,alina.cristea}@fmi.unibuc.ro, iordache.bogdan1998@gmail.com,
{simona.georgescu,laurentiu.zoicas}@lls.unibuc.ro

Abstract

Identifying the type of relationship between words (cognates, borrowings, inherited) provides a deeper insight into the history of a language and allows for a better characterization of language relatedness. In this paper, we propose a computational approach for discriminating between cognates and borrowings, one of the most difficult tasks in historical linguistics. We compare the discriminative power of graphic and phonetic features and we analyze the underlying linguistic factors that prove relevant in the classification task. We perform experiments for pairs of languages in the Romance language family (French, Italian, Spanish, Portuguese, and Romanian), based on a comprehensive database of Romance cognates and borrowings. To our knowledge, this is one of the first attempts of this kind and the most comprehensive in terms of covered languages.

Keywords: cognates, borrowings, historical linguistics, romance languages

1. Introduction and Related Work

Ever since the founding of comparative historical linguistics with the discovery of the Indo-European language family at the end of the 18th century, one of the most challenging tasks for linguists has been to distinguish between cognates and borrowings. Cognates are words derived from a common ancestor (e.g. Ro. *tânăr* 'young' – Fr. *tendre* 'tender', both coming from Lat. *tener* 'tender'), while borrowings are words derived from one another (e.g. Ro. *coafor* 'hairdresser' was borrowed from Fr. *coiffeur*). The discrimination between these two etymological categories is essential for various domains of diachronic and synchronic linguistics, and it is not always easy: for example, the above-mentioned French word *tendre* 'tender' was also borrowed in Romanian as *tandru* 'loving, gentle'; we thus have a cognate relation between Fr. *tendre* and Ro. *tânăr*, but a borrowing relation between Fr. *tendre* and Ro. *tandru*.

On the one hand, the correct identification of cognates offers a real reflection of the genetic relations between languages, constituting thus the foundation of linguistic phylogeny (Dunn, 2015; Heggarty et al., 2023) and allowing for a deeper insight into the geographic and chronological aspects that define a linguistic community (Heggarty, 2015). Cognate sets (i.e. strings of words coming from the same ancestor, in different languages, e.g. Ro. *tânăr* 'young' – It. *tenero* 'soft' – Fr. *tendre* 'tender' – Es. *tierno* 'id.', Pt. *tenro* 'delicate') represent, in fact, the basis of the comparative grammar – reconstruction method (Chambon, 2007), providing the necessary information concerning the patterns and extension of linguistic change.

On the other hand, an accurate identification of borrowings furnishes data that can create a faithful mirror of the mutual influences between languages across space and time, while permitting what Epps (2014) called a "socio-cultural reconstruction" (cf. (Mailhammer, 2015). Campbell (1998) suggest that, in principle, borrowings must be identified and excluded before the comparative method can be applied. Following the example cited above, if we placed Ro. *tandru* 'gentle' (borrowed from Fr. *tendre* in the 19th century) at the same level as It. *tenero*, Fr. *tendre*, etc., we would draw fallacious conclusions concerning the phonetic and semantic evolutionary patterns from Latin to Romanian, which would consequently lead to a false interpretation of both the history of Romance languages and of social relations, ethnic contact and cultural connections across the territory of the former Roman Empire. Heggarty (2012) highlights the importance of an accurate distinction between cognates and borrowings, stressing that the possible errors, while distorting the real relation between languages, invalidate the whole scientific process.

From a synchronic point of view, the correct discerning of cognates vs borrowings can lead to a more accurate model of automatic translation (Kondrak et al., 2003). It also removes the errors generated by pairs of false friends (Uban and Dinu, 2020), which reside mostly in a mistaken interpretation of the genetic relation between two similar

LREC-COLING 2024, pages 12657–12667

words. Moreover, the implications of a correct categorization of word pairs can be found in language acquisition and bilingual word recognition (Dijkstra et al., 2012), as well as in the analysis of linguistic corpora (Simard et al., 1992).

Distinguishing between cognates and borrowings is not an easy task, and there are many cases of philological disputes about whether a word belongs to one category or the other. For example, Es. *atacar* 'to attack' is considered by some linguists to be borrowed from It. *attacare* 'id.', along with a certain military technique (DRAE[1], DCECH[2]), while other philologists (Georgescu, 2021) argue that both Italian and Spanish have inherited this verb from a Latin protoform **attaccare* 'to bind, to collide', implying that this word had a parallel semantic evolution with no mutual influence, as both communities interpreted war in the same terms. Just to mention a few more examples, it is still questionable if It. *trovare* 'to find' was borrowed from French *trouver* 'id.' or is its cognate, if Es. *botar* 'to jump' comes from Fr. *bouter* 'to throw' or both words were inherited from a Latin non-attested word, just as we are uncertain if Fr. *charabia* 'gibberish' was borrowed from Es. *algarabía* 'indistinguishable noise coming from people shouting', or they were both borrowed from Classical Arabic *'arabiyyah* 'Arabic language', the hypothesis of them being totally unrelated being also invoked. This former case can be particularly interesting for the history of linguistic contacts as well as of cultural stereotypes. In such cases, where the traditional philological approach cannot reach definitive conclusions, an objective computational approach could bring new useful insights in the lexical - and therefore historical and anthropological - relations (Heggarty, 2012).

Nonetheless, Jäger (2019) considers the handling of language contact (and borrowings, more specifically) an "unsolved problem for computational historical linguistics", as "automatic cognate clustering does not distinguish between genuine cognates (related via unbroken chains of vertical descent) and (descendants of) loanwords". The challenge of automatically identifying cognates and borrowings has become a necessity today, considering the large amount of linguistic data that has not yet been processed from a historical perspective (List et al., 2017). The last decade has brought the first attempts to automatically discriminate between cognates and borrowings for various pairs of languages (Ciobanu and Dinu, 2015; Tsvetkov et al., 2015; Mi et al., 2018) and to automatically analyze loanwords at various lexical levels (Cristea et al., 2021; Nath et al., 2022; Miller and List, 2023). Sims-Williams (2018) discuss the computational approaches to the main historical linguistics problems in a more general frame, highlighting the difficulty of differentiating between cognates and borrowings.

Considering the multiple applications of accurately discriminating between cognates and borrowings, the purpose of this work is twofold. On one hand, when analyzing two languages stemming from the same ancestor, we need to establish whether a certain lexical similarity is the result of divergence starting from the same nucleus, or it reflects a convergence through linguistic contact (Heggarty, 2012). On the other hand, in cases of uncertain origin of languages, we could eventually determine if two languages do have a common ancestor or they have simply influenced each other across time: "the method could be applied to languages not yet known to be related, and used precisely to diagnose or establish whether they were. From the information question 'how closely related?', one had passed to the yes/no question 'related or not?'" (Heggarty, 2012).

Starting with these remarks, our main contributions are:

- We investigate whether cognates can be automatically distinguished from borrowings based on their graphic and phonetic forms. More specifically, our task is as follows: given a pair of words (x, y) in two different Romance languages, we want to determine whether x and y are cognates, borrowings or neither. 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 out of the five languages considered.
- We explore which kinds of features and machine learning models are more useful for accurately distinguishing between cognates and borrowings, also investigating whether graphic or phonetic similarities between words are more significant for identifying borrowings as opposed to cognates (or unrelated words).
- We discern for which of the main Romance languages it is more challenging to distinguish between cognates and borrowings and what this can tell us about language similarity and evolution.

The rest of the paper is organized as follows: in Section 2 we present the database that we use and offer details about the processing steps involved, in Section 3 we introduce our approach for the automatic discrimination between cognates and bor-

[1]Diccionario de la lengua española

[2]Diccionario crítico etimológico castellano e hispánico

rowings and present our experiments, and extensive results and error analyses are presented in Section 4. The last section is dedicated to final remarks.

2. Data

We use a comprehensive database of cognates and borrowings for Romance languages (Dinu et al., 2023) thus our results can be considered a benchmark for cognate-borrowing discrimination for the main Romance languages. The dataset is based on the available machine-readable reference dictionaries[3], which contain etymological information.

For each of the five Romance languages (It, Es, Fr, Pt, Ro), the database contains lists of words, with their etymologies. Starting with these data, we obtained new lists of cognate pairs and borrowing pairs between any two Romance languages of the five, by the following procedure. For any triplet <*u, e, L_1*> in language L_1, if we find a triplet <*v, e, L_2*> in L_2 (having the same etymon *e*), add the triplet <*u, v, e*> to the list of cognate pairs of the language pair (L_1, L_2). Borrowings are defined as word pairs <*u,v*> in a tuple <*u,v,e*> where $e = v$ or $e = u$ (the etymon of one word is the other word). For any language pair (A, B), we have merged together into a unique list of borrowings the words borrowed by language A from language B and the words borrowed by language B from language A.

The database comprises a total of 125,598 words across all languages and 90,853 cognate pairs. Table 3 shows the total number of words per language and the top three source languages for borrowings, for each language. The number of cognate and borrowing pairs identified for any language pair is depicted in Table 4. We have also computed the average Levenshtein distance between pairs of cognates and borrowings, for all pairs of Romance languages - shown in Table 1.

Regarding the accuracy of the extraction and cleaning algorithm for obtaining the cognate and borrowings sets, the following accuracies were computed based on 100 randomly sampled examples for each language and language pair respectively: an average accuracy of 98.6% for extracting etymologies (100% for Spanish, 98% for Romanian, 97% for Portuguese, 100% for Italian, and 98% for French) and an average accuracy of 98.2% for cognates extraction (the accuracies for all pairs of languages are shown in Table 2).

[3]Italian: *Il dizionario della lingua italiana De Mauro*, dizionario.internazionale.it. Spanish: *Diccionario de la lengua española*, lema.rae.es/drae. Portuguese: *Dicionário infopédia da Língua Portuguesa*, www.infopedia.pt/lingua-portuguesa. French: *Trésor de la Langue Française informatisé*, www.cnrtl.fr. Romanian: *Dicționarul Explicativ al Limbii Române*, dexonline.ro.

	Ro	It	Es	Pt	Fr
Ro		3.31	2.81	2.95	2.92
It	1.40		2.73	2.71	3.32
Es	0.60	2.13		2.82	2.84
Pt	0.71	1.56	1.04		2.84
Fr	2.39	3.01	2.13	2.20	

Table 1: Average Levenshtein distance between cognate pairs (above the main diagonal) and borrowing pairs (below the main diagonal) for each Romance languages pair.

	Ro	It	Es	Pt
It	98%			
Es	98%	99%		
Pt	99%	99%	97%	
Fr	98%	98%	98%	98%

Table 2: Estimated accuracy for the cognate extraction method used for building the database based on etymology dictionaries.

3. Automatic Cognate-Borrowing Discrimination Methodology

We tackle the problem of distinguishing between pairs of cognates and borrowings through several experiments. There are in fact two categories of tasks defined for each pair of Romance languages considered: the first one implies a binary classification problem meant to discriminate between pairs of borrowings and cognates, the second one adds a third class consisting of unrelated words (which comes with various degrees of difficulty: it is one thing to distinguish between randomly selected pairs of words, and another to discriminate unrelated words that have very similar forms).

As for methods, we employ a variety of classical machine learning algorithms along with deep learning models based on character-level Transformers. Important information about the relationships between words can be found in both their graphical representation and their phonetic transcription. Because of that we experiment with several combinations of such representations.

3.1. Cognates or Borrowings

Our first proposed task can be defined as follows: given a pair of related words from two languages (about which we know *a priori* that they are ei-

Ro	It	Es	Pt	Fr
45465	24257	16458	28180	19822
Fr:35511 Lat:9312 It:3358	Lat:18437 Fr:1981 Gr:1649	Lat:11936 Fr:1366 Es:712	Lat:17446 Gr:2818 Fr:2369	Lat:12804 En:1086 It:912

Table 3: Number of words in dictionaries for each language (upper row), and most frequent source language (lower row) across words in a dictionary.

	Ro	It	Es	Pt	Fr
Ro		4,999	7,588	5,855	7,360
It	3,139		7,863	12,198	7,105
Es	209	770		9,533	10,220
Pt	103	620	1,201		7,783
Fr	33,311	2,896	1,690	2,450	

Table 4: Number of cognate pairs (above the main diagonal) and borrowing pairs (below the main diagonal) for each Romance languages pair.

ther cognates or borrowings), the model needs to decide on its correct class (cognate or borrowing). For example, in the It-Ro pair (gola, gură) the words should be identified as cognates, while a borrowing should be predicted in the case of the Es-Fr pair (bufanda, bouffante).

Experimental Settings. We perform our experiments independently for each of the ten language pairs. For training and testing our classification methods, we randomly generate $90\% : 10\%$ splits from the cognates and borrowings dataset for a given language pair. The splits are stratified, such that for every language pair the ratio between the number of cognate pairs and borrowing pairs is the same for both sides of the split. All of the experiments rely, in some way, on either the graphic representation of the words, the phonetic one, or both. In order to obtain the phonetic representation, similarly to previous studies (Meloni et al., 2021), we used the open-source eSpeak library[4].

Features. Some of our approaches rely on handcrafted features extracted from the graphic and phonetic 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 phonetic representation, as opposed to the mere graphic one. The process is similar (i.e. we align the graphemes, 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.

Ensemble Model. For this set of experiments, we preprocess the graphic representation of the words, by removing the accents, while the phonetic one is left as it is. Using the feature vectors computed from the alignments, we train various machine learning algorithms: Support Vector Machine, Naive Bayes, and other linear classifiers trained using stochastic gradient descent. We employ, for all of these, the implementation provided by the *scikit-learn* library (Pedregosa et al., 2011). The models are trained using the graphic features, the phonetic features, or both. The performance is assessed through 3-fold cross-validation on the training set of each of the language pairs. We then select the best 5 performing models and build a stacking ensemble classifier. This ensemble is then trained on the whole 90% training split. We also evaluate ensembles trained using only graphic and only phonetic base models, respectively, to assess if any category of features outperforms the other, or if their combination is more favorable.

Transformer Model. Our second approach relies on the Transformer model (Vaswani et al., 2017). We train such models either on the graphic, or on the phonetic form of the words. The "tokenization" is performed by splitting the representations into letters or into graphemes, 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% training selection as for the ensembles. The transformer models are ultimately 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 (approximately 10% of the whole dataset size), in order to stop the training process early and avoid divergence.

Metrics. Our main metric for comparing the performance of the trained models (in all of the experiments from this section and the following one) is the macro-averaged F1 score. We chose it since our datasets are unbalanced, but we also compute the accuracy, precision, and recall scores.

[4]https://github.com/espeak-ng/espeak-ng

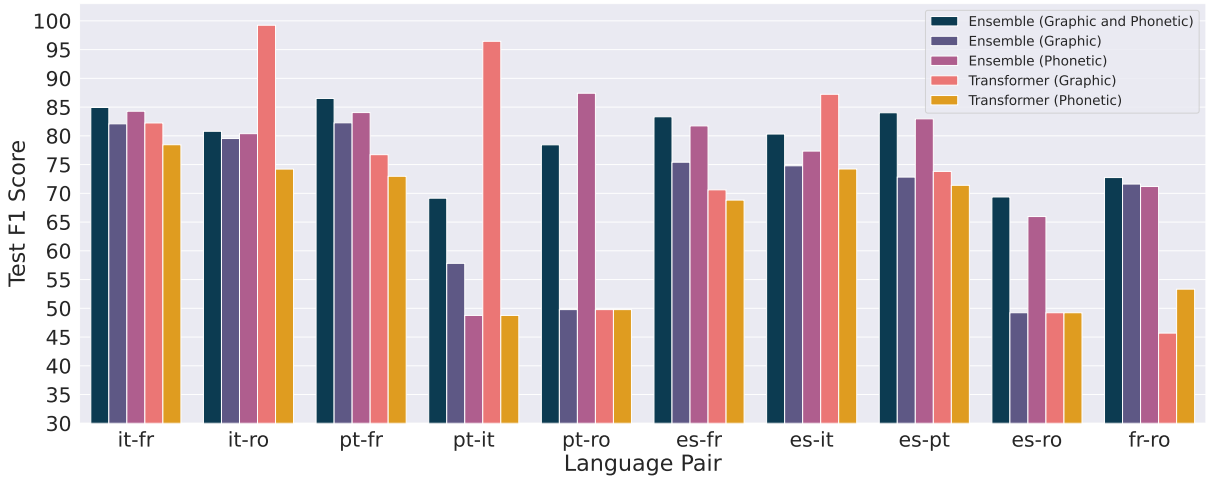

Figure 1: F1 scores for all models and features, measured for the binary classification task.

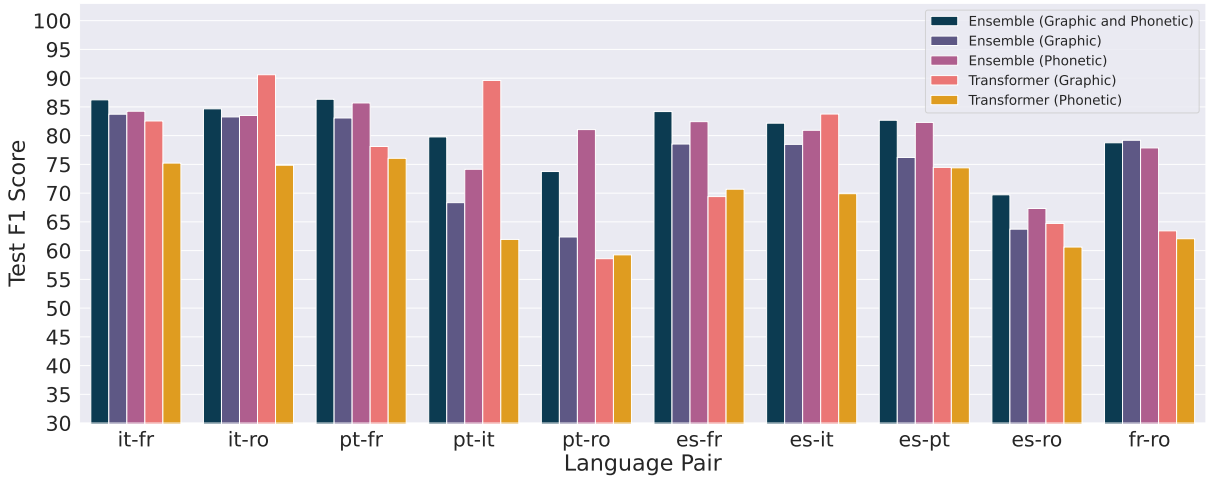

Figure 2: F1 scores for all models and features, measured for the ternary classification task where unrelated examples were selected via Levenshtein distance

3.2. Cognates, Borrowings, or Unrelated

In the first experiment we start from the idea that two words are related (either cognates, or one borrowed from the other), and we have to discriminate between these two categories. However, in a more realistic scenario, we do not know *a priori* if two words are related or not. Some pairs of words from two different languages could have very close forms (e.g. Ro. *molie* 'moth' and It. *moglie* 'wife'), and could be erroneously considered as related words (especially by methods based on computing the distance between the word forms). Thus, we additionally propose an experiment where we attempt to distinguish between cognate pairs, borrowing pairs, and pairs of words which are neither cognates, nor borrowings. In other words, for a pair of words from two different languages, we have to decide if the two words are cognates, borrowings, or unrelated.

In order to design this experiment, we need to complement our datasets with examples of word pairs which are unrelated. It is remarkable that, to our best coverage of the literature, while positive data was generally well documented, negative data lack explanations, with a few exceptions (Ciobanu and Dinu, 2014). The choice of negative examples is essential in informing the interpretation of automatic detection results. For instance, it is easy to decide that two obviously different words in two languages such as Romanian *apă* ('water') and Spanish *cerveza* ('beer') are not cognates, but not so easy for more similar words such as Italian *rumare* ('rumble') and Romanian *rumen* ('ruddy').

Unrelated words selection. To address this issue, we propose two methods of sampling unrelated word pairs. For each 3-way classification experiment, we generate a number of unrelated pairs examples equal to the number of examples in the larger of the two remaining classes: if $|B|$ is the number of borrowings for a given language pair,

and $|C|$ is the number of cognates for the same pair, then the number of unrelated examples generated is equal to $max(|B|, |C|)$. In this way we ensure the three classes are fairly balanced, while still preferring to generate enough examples that the classifiers can better learn from.

Random sampling. In the simpler setting, we generate a pair of unrelated words by randomly extracting word pairs from the two languages and excluding pairs of cognates and borrowings from the selection, if they occur.

Levenshtein-based sampling. We include a second method where we select as unrelated examples word pairs which are graphically similar and do not have common etymology (they are neither cognates, nor borrowings), by conditioning the words in the pair to have a Levenshtein distance (Levenshtein, 1965) smaller than the average Levenshtein distance across cognate sets for that language pair.

For our second set of experiments we use the same features and models as in the binary classification experiments.

4. Results and Error Analysis

In this section we report the results of the experiments carried out, analyse them both from a computational and a linguistic point of view, and extract significant data for the description of Romance languages. A subsection is devoted to error analysis, which is also useful both for future linguistic and computational investigations.

4.1. Results Analysis

In Table 6 we show the results obtained for cognate-borrowing discrimination (two classes) by using ensemble methods and transformers, for each pair of Romance languages considered (a visual representation of these results can be seen in Fig. 2). For ensembles (first row for each language in the table), we show results by using only graphic, phonetic, or mixed features, and for transformers (second row for each language) we show results by using only graphic and phonetic features. The best result (99.2%) were obtained for It-Ro, with graphical transformers, followed by It-Pt (96.4), with the same model architecture. The explanation probably lies in the much clearer phonetic distinctions, in these language pairs, between borrowings and cognates: the latter have undergone a well-defined phonetic evolution that has distanced them from each other, enough that they are not confused with borrowings, but not to such a degree that they cannot be identified as cognates.

Generally, the best results were obtained either via the ensemble model trained on both graphic and phonetic alignment features (outperforming the grahic-only and phonetic-only classical models), or via the character Transformer model trained on the word forms. The average F1 scores for all the pairs of languages, for the best model, for the ensemble models (trained on all, phonetic and graphic features) and for Transformers models (graphic and phonetic) are presented in Table 5. The average score for binary classification was 85.09% for the best model, followed by the ensemble model trained on all features (78.96%). We note that, for the pairs which included French, the best model is based on the ensemble combining the best features (both graphic and phonetic). In the pairs comprising Italian (except for the It-Fr pair), the best results are obtained with graphical transformers, while for (Pt-Ro) the best result is obtained with the phonetic ensemble. The fact that, although we have a majority system (ensemble all), we do not have unanimity for a single system that gets the best results, can only testify that the problem is difficult, and that each language pair comes with its own challenges.

The fact that there are 3 different models that obtain the best results allows us to investigate the specificity of each relationship established at the inter-Romance level. Thus, the etymological spelling in French requires the combination of two systems, since, on the one hand, the spelling does not reflect the pronunciation, on the other hand, the pronunciation sometimes makes the words unrecognisable in relation to their lexical correlations in other Romance languages. The rendering of the pronunciation, i.e. of the phonetic form, is useful in detecting borrowings from French into other languages, since the vast majority of these borrowings conforms to the oral structure of the word, e.g. Fr. *charpente* 'roof truss' pronounced [ʃaʁˈpɑ̃t] is borrowed into Romanian as *şarpantă* [ʃarˈpan.tə] thus reproducing the pronunciation, not the French spelling. On the other hand, the conservative spelling is important in the process of detecting cognates, since in this case the pronunciation in French divergent from any other Romance language in our list would greatly distance the word forms, e.g. Fr. *nuit* 'night' [nɥi] *versus* It. *notte*, Pt. *noite*. Figure 3 shows the 3 most relevant alignment n-grams according to feature selection in the ensemble models. In Table 8 we present the confusion matrix for binary and three classes classification, computed using the best model from each pair of languages.

For three class classification (cognate-borrowing-unrelated) we show the results in Table 7, both for random negative sampling (lower diagonal) and Levenshtein-based negative sampling (upper diagonal). In this ternary system, the results provide interesting insights into machine learning and its ability to respond to this challenge. At first

	french	italian	romanian	spanish	portuguese
french		o$→-$, e$→-$, no→--	--→ne, ˈɔ̃-→ˈun, jˈɔ̃→jˈu	ta→t-, $tʃ→$ʃ, xe→ʒ-	ŋ$→-$, ɹ$→-$, ʊ$→-$
italian	o$→-$, o$→e$, no→n-		o$→-$, jˈo→j-, ne→-e	ˈet→ˈet:, ta→t:a, -$→e$	ɐ$→a$, ˈet→ˈet:, ˈɛt→ˈet:
romanian	-$→e$, -$→a$, nt→--	o$→-$, ne→-e, on→--		ˈað→ˈad, ða→də, ðˈo→dˈo	ɐ$→a$, ˈeɪ-→eˈi, ʁi→re
spanish	a$→e$, je→ge, aj→ag	ll→gl, il→ig, l-→li	ss→s-, i$→-$, a-→as		o$→ʊ$, ʎa→ʎɐ, ʎo→ʎʊ
portuguese	r$→-$, o$→e$, em→e-	ch→cc, h-→ci, t-→tt	mm→m-, av→--, o-→os	ll→lh, la→ha, lo→ho	

Figure 3: Top 3 most relevant alignment n-grams ($n \leq 2$) according to χ^2 feature selection, based on the full training dataset, for the binary classification experiments. Below the diagonal, we show the graphic features, while above we added the phonetic ones.

	Best	EnAll	EnGr	EnPh	TrGr	TrPh
2Cl	85.09	78.92	69.49	76.35	73.06	64.07
3Cl-L	83.24	80.77	75.64	79.9	75.47	68.46
3Cl-R	87.78	84.72	79.68	83.2	81.76	73.62

Table 5: The average scores for the best model, for the ensemble models trained on all features (EnAll), graphic features (EnGr), and phonetic features, respectively, and for the graphic (TrGr) and phonetic (TrPh) Transformers, in all of the three scenarios: binary classification (2Cl), and ternary classification with unrelated pairs selected randomly (3Cl-R) or via Levenshtein distance (3Cl-L).

sight, the results seem weaker: e.g., for the It-Ro pair, where in the binary system the result was 99.2%, an F1 score of 90.5% was reached when we added words chosen with Levenshtein distance and 96.2% when we introduced words chosen randomly. Nevertheless, a closer look at the examples themselves shows that, in fact, the performance in identifying cognate pairs and borrowings was much higher. Comparing the list in the binary system with that of the three class classification, we note that many of the mistakes made by the machine in the first experiment were corrected in the second one. For example, Fr. *crever* 'to burst' – Ro. *crăpa*, both inherited from Lat. *crepare* 'to crack', so cognates, were predicted as borrowings in our binary system experiment, while in the ternary they were correctly identified as cognates. As a general observation, when the three-category classification model was applied, the number of borrowing pairs that were misclassified as cognates in the first experiment dropped by half in non-neighbouring language pairs (Table 8). Furthermore, our intuition on the difficulty of three-way classification was correct, as in all experiments where unrelated examples were sampled randomly, the models performed better than their equivalents using the Levenshtein-based selection.

4.2. Error Analysis

The lowest accuracy rates both for binary and ternary classes were performed for Spanish-Romanian and French-Romanian. The third lowest rate, in the two-class categorization, occurred for the French-Spanish pair, while in the three-class system it was reported for Portuguese-Romanian. We believe that the reasons for this increased difficulties are specific to each language pair's relationship. Thus, in the case of the Spanish-Romanian pair, the confusion between cognates and borrowings is probably determined by the similarity of the phonetic laws that characterise the transfer of words from Latin to Romanian and Spanish, which makes it difficult to recognise borrowed words from Spanish to Romanian. Thus, a word such as Ro. *infante* 'king's son', borrowed from Es. *infante* (in its turn originated in Lat. *infantem* 'child'), would have had the same form, had it been taken directly from Latin. In such cases where the result of a regular sound evolution from the source language to both languages L_1 and L_2 would have been similar, the machine performed a lower rate of discrimination between cognates and borrowings between L_1 and L_2. Consequently, the borrowings from Spanish into Romanian would not have been significantly different from the words taken directly from Latin into Romanian. Moreover, the small number of examples of borrowings from Spanish into Romanian and the absence of borrowings in the opposite direction gave the machine insufficient data to learn any differences. It is noteworthy, however, that the computer identified the phonological features of the words originated in Latin, and thus, in this particular case, placing a loan between cognates would not have altered the interpretation of phonetic features or genetic relationships. In the case of the French-Romanian pair, the lower rate of discrimination is a consequence of the peculiarity of the relationship between the Romanian language and French, more precisely of the category of words with multiple etymologies: the Romanian language is characterised by a large number of borrowings which, for instance, in one region of the country have been taken from French (e.g. by a writer with knowledge of this language, then spread in the current speech) and, in another

		Ro			It			Es			Pt		
		Gr	Ph	All	Gr	Ph	All	Gr	Ph	All	Gr	Ph	All
It	En	79.5	80.3	80.7	-	-	-	-	-	-	-	-	-
	Tr	**99.2**	74.2	-	-	-	-	-	-	-	-	-	-
Es	En	49.2	65.9	**69.3**	74.7	77.3	80.3	-	-	-	-	-	-
	Tr	49.2	49.2	-	**87.2**	74.2	-	-	-	-	-	-	-
Pt	En	49.7	**87.4**	78.4	57.8	48.7	69.1	72.8	82.9	**84.0**	-	-	-
	Tr	49.7	49.7	-	**96.4**	48.7	-	73.8	71.3		-	-	-
Fr	En	71.6	71.1	**72.7**	82.0	84.2	**84.9**	75.4	81.7	**83.3**	82.2	84.0	**86.5**
	Tr	45.6	53.3	-	82.2	78.4	-	70.6	68.8	-	76.7	72.9	-

Table 6: Cognate-borrowing classification F1 (two classes) on the test set using the ensemble model and transformers. For each language pair, the results are displayed on two consecutive rows: the first row show the results using graphical-only (Gr), and phonetic-only (Ph) features, and the best ensemble (En) with combined features (shown on three consecutive columns), while the results using transformers model (only graphical and phonetic) are shown on the second row.

		Ro		It		Es		Pt		Fr	
		En	Tr	En	Tr	En	Tr	En	Tr	En	Tr
Ro	Gr	-	-	83.2	**90.5**	63.7	64.7	62.3	58.5	79.1	63.4
	Ph	-	-	83.5	74.8	67.3	60.6	**81.0**	59.2	77.8	62.0
	All	-	-	84.6	-	**69.7**	-	73.7	-	**78.7**	-
It	Gr	85.7	**96.2**	-	-	78.4	**83.7**	68.3	**89.6**	83.7	82.5
	Ph	85.7	78.8	-	-	80.9	69.9	74.1	61.9	84.2	75.2
	All	86.4	-	-	-	82.1	-	79.7	-	**86.2**	-
Es	Gr	65.0	**79.9**	82.9	**90.5**	-	-	76.2	74.4	78.5	69.3
	Ph	75.9	66.8	82.9	80.6	-	-	82.2	74.4	82.4	70.6
	All	78.2	-	85.2	-	-	-	**82.6**	-	**84.1**	-
Pt	Gr	76.8	62.6	70.2	**92.5**	78.8	78.6	-	-	83.0	78.1
	Ph	81.8	62.5	76.4	64.9	86.9	76.5	-	-	85.6	76.0
	All	**82.5**	-	78.8	-	**88.4**	-	-	-	**86.3**	-
Fr	Gr	**80.7**	72.9	87.3	86.6	82.6	76.4	86.8	81.4	-	-
	Ph	79.2	70.0	87.7	82.2	87.1	75.9	88.4	78.0	-	-
	All	80.6	-	**89.8**	-	**88.0**	-	**89.3**	-	-	-

Table 7: Cognate-borrowing-unrelated classification F1 (three classes) on the test set using the ensemble model and transformers. For each language pair, the results are displayed on two consecutive columns: the first column show the results using graphical-only (Gr), and phonetic-only (Ph) features, and the best ensemble (En) with combined features (shown on three consecutive rows), while the results using transformers model (only graphical and phonetic) are shown on the second column. The results using the Levenshtein-based sampling are shown above the main diagonal, while the results using random sampling are shown below the main diagonal.

		Ro			It			Es			Pt			Fr		
		Co	Bo	Un	Co	Bo	Un	Co	Bo	Un	Co	Bo	Un	Co	Bo	Un
Ro	Co	-	-	-	466	3	-	699	2	-	546	0	-	272	399	-
	Bo	-	-	-	3	336	-	16	6	-	2	3	-	99	3954	-
	Un	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
It	Co	390	0	79	-	-	-	781	5	-	1220	0	-	668	43	-
	Bo	2	326	11	-	-	-	27	50	-	8	54	-	77	214	-
	Un	46	3	420	-	-	-	-	-	-	-	-	-	-	-	-
Es	Co	635	2	64	698	7	81	-	-	-	941	9	-	996	26	-
	Bo	7	4	11	13	47	17	-	-	-	49	71	-	62	107	-
	Un	42	3	656	80	4	702	-	-	-	-	-	-	-	-	-
Pt	Co	507	0	39	1083	2	135	873	15	66	-	-	-	742	37	-
	Bo	1	2	2	4	53	5	36	69	15	-	-	-	60	184	-
	Un	36	0	510	130	4	1086	54	10	890	-	-	-	-	-	-
Fr	Co	238	402	31	636	38	37	955	24	43	705	39	35	-	-	-
	Bo	97	3414	42	65	207	19	57	97	15	59	175	10	-	-	-
	Un	9	31	1893	25	7	679	43	5	974	30	10	739	-	-	-

Table 8: Confusion matrix. For each language pair, the results are displayed for the best system results: upper diagonal are the results for two classes discrimination, lower diagonal are the results for three classes discrimination (cognate-Co, borrowing-Bo, unrelated-Un).

region, from Latin (the cultural model considered by the educated people of the area), and are thus cognate with French words coming from the same Latin word. In such cases, dictionaries mention both origins (e.g. Ro. *tensiune* < Fr. *tension*, Lat. *tensio*; Ro. *persoană* < Lat. *persona*, Fr. *personne*, etc.). This is exactly the category of words where the machine prediction differed from the label in our initial list, but given the situation described above, either of the two options can be considered to be correct (e.g. Fr. *administrateur* - Ro *administrator*, Fr. *adolescent* - Ro. *adolescent*, Fr. *adorable* - Ro. *adorabil*, etc.).

The third pair with poorer results than the rest in the binary classification, French-Spanish, is characterised by a linguistic contact that spans at least three centuries, which means that not a few words borrowed from French into Spanish or vice versa have had sufficient time to adapt to the receiving language, thus becoming phonetically closer to words originated in Latin in both languages.

As for the third pair with a low accuracy rate in the ternary class experiment, Portuguese-Romanian, the explanation resides, contrary to the previously-mentioned situation, in the very reduced contact between these languages (located at the two extremes, on the east-west axis, of the former Roman Empire), which translates into insufficient input for machine learning. In some cases, the choice of the machine helped us to identify and correct errors in the database. For example, in the initial list, Ro. *adjutant* was classified as a borrowing from Es. *ayudante*, but the machine identified them as cognates: the situation is quite complicated, since Es. *ayudante* is a derivative of Es. *ayudar*, which in turn is inherited from Latin; Ro. *adjutant* is borrowed from French, but with a phonetic adaptation in accordance with Latin *adiutare*, which makes it difficult to catalogue the relationship between Es. *ayudante* and Ro. *adjutant*, but in no case are we dealing with a borrowing from one language to the other.

Another example is Es. *borona* - Pt. *broa*, classified in the database as a borrowing situation (due to the explanation provided by the dictionary used in the creation of the database, Infopedia), and categorized by the machine as cognates. Actually, it is impossible that a borrowing from Spanish into Portuguese would have undergone the sound laws specific to a word inherited from Latin into Portuguese: the syncopation of /o/ and the lenition of intervocalic /n/ (as in Lat. *ponere* 'to put' > old Pt. *poer* > Pt. *pôr*). The machine actually had the same interpretation as DELP [5], considering them cognates. A similar situation is encountered in the case of Es. *centollo* / Pt. *santola*, which appear in our initial list as borrowings (due to Infopedia, which considers Pt *santola* as a borrowing from Es. *centollo*, a choice not supported by the phonetic evidence), but which the machine classifies as cognates, again corresponding to the choice made by DELP. It is remarkable, therefore, that the computer contributed to improving the database.

[5]Dicionário etimológico da língua portuguesa

5. Conclusions and Future Work

Taking as a starting point the idea that identifying the type of relationship between words is essential for tracing back the history of languages, we propose a computational approach for discriminating between cognates and borrowings, which is considered to this day one of the most difficult tasks in historical linguistics. For this purpose, we use a comprehensive database of cognates and borrowings for Romance languages, based on the available machine-readable reference dictionaries (Dinu et al., 2023). We employ a variety of classical machine learning algorithms along with deep learning models based on character-level Transformers. Given that important information about the relationships between words can be found in both their graphical representation and their phonetic transcription, we compare the discriminative power of graphic and phonetic features and we analyze the underlying linguistic factors that prove relevant in the classification task. We provide an extensive analysis of results and errors, useful for further linguistic and computational investigations. Considering that, to our knowledge, this is one of the first attempts of this kind and the most comprehensive in terms of covered languages, the presented results can be considered a benchmark for cognate-borrowing discrimination for Romance languages. For further investigation, we aim to provide an extended analysis of machine-selected n-grams, structures that contributed to the selection of the type of relationship between word pairs, and compare them with the information we know from historical Romance linguistics. Also, in order to increase the accuracy in discriminating between the two categories, we intend to go beyond the strictly formal aspect of the word, enriching the input with semantic information, sometimes essential for detecting the genetic relationship between words.

Acknowledgements

Research partially supported by the POCIDIF project in Action 1.2 "Romanian Hub for Artificial Intelligence".

Ethical Considerations and Limitations

There are no ethical issues that could result from the publication of our work. Our experiments com-

ply with all license agreements of the data sources used.

There are a few limitations to our cognates-borrowing discrimination results. First, distinguishing between oral and written Latin can further refine the types of etymological relations between words of Latin origin. Another clear limitation is that the used database only covers the main Romance languages, and does not yet include other Romance varieties nor any other language families.

In terms of cognates-borrowings discrimination experiments, we acknowledge there are different architectures and feature sets to be used for this task which could improve results in the case of deep learning models, and we invite other researchers to propose new methods and test them. An explainability analysis of the deep models could also be interesting to understand to what extent they are capable of identifying "alignment" patterns based only on word forms. A classifier trained on all language pairs together could also reveal interesting commonalities across language pairs, as well as potentially obtain better results due to this.

6. Bibliographical References

Lyle Campbell. 1998. *Historical Linguistics. An Introduction*. MIT Press.

Jean-Pierre Chambon. 2007. Remarques sur la grammaire comparée-reconstruction en linguistique romane (situation, perspectives). *Mémoires de la Société de linguistique de Paris*, 15:57–72.

Alina Maria Ciobanu and Liviu P. Dinu. 2014. Automatic Detection of Cognates Using Orthographic Alignment. In *Proceedings of ACL 2014, Volume 2: Short Papers*, pages 99–105.

Alina Maria Ciobanu and Liviu P. Dinu. 2015. Automatic Discrimination between Cognates and Borrowings. In *Proceedings of ACL 2015, Volume 2: Short Papers*, pages 431–437.

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, Liviu P. Dinu, Simona Georgescu, Mihnea-Lucian Mihai, and Ana Sabina Uban. 2021. 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.

Ton Dijkstra, Franc Grootjen, and Job Schepens. 2012. Distributions of Cognates in Europe as Based on Levenshtein Distance. *Bilingualism: Language and Cognition*, 15:157–166.

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.

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

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

Simona Georgescu. 2021. *La regularidad en el cambio semántico. Las onomatopeyas en cuanto centros de expansión en las lenguas románicas*. ELiPhi Éditions de linguistique et de philologie, Strasbourg.

Paul Heggarty. 2012. Beyond Lexicostatistics: How to Get More out of "Word List" Comparisons. In *Quantitative Approaches to Linguistic Diversity: Commemorating the Centenary of the Birth of Morris Swadesh*, pages 113–137. Benjamins.

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

Paul Heggarty, Cormac Anderson, Matthew Scarborough, Benedict King, Remco Bouckaert, Lechosław Jocz, Martin Joachim Kümmel, Thomas Jügel, Britta Irslinger, Roland Pooth, et al. 2023. Language trees with sampled ancestors support a hybrid model for the origin of indo-european languages. *Science*, 381(6656):eabg0818.

Gerhard Jäger. 2019. Computational historical linguistics. *Theoretical Linguistics*, 45(3-4):151–182.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates can improve statistical translation models. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT-NAACL 2003, Edmonton, Canada,*

May 27 - June 1, 2003. The Association for Computational Linguistics.

Vladimir I. Levenshtein. 1965. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady*, 10:707–710.

Johann-Mattis List, Simon J. Greenhill, and Russell D. Gray. 2017. The Potential of Automatic Word Comparison for Historical Linguistics. *PLOS ONE*, 12(1):1–18.

Robert Mailhammer. 2015. Etymology. In *The Routledge handbook of historical linguistics*, pages 441–459. Routledge.

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.

Chenggang Mi, Yating Yang, Lei Wang, Xi Zhou, and Tonghai Jiang. 2018. Toward better loanword identification in uyghur using cross-lingual word embeddings. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3027–3037. 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*, pages 2599–2605, Dubrovnik, Croatia. Association for Computational Linguistics.

Abhijnan Nath, Sina Mahdipour Saravani, Ibrahim Khebour, Sheikh Mannan, Zihui Li, and Nikhil Krishnaswamy. 2022. A generalized method for automated multilingual loanword detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4996–5013.

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.

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. *Journal of machine learning research*, 12(Oct):2825–2830.

Michel Simard, George F. Foster, and Pierre Isabelle. 1992. Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation*.

Patrick Sims-Williams. 2018. Mechanising historical phonology. *Transactions of the Philological Society*, 116(3):555–573.

Yulia Tsvetkov, Waleed Ammar, and Chris Dyer. 2015. Constraint-based models of lexical borrowing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 598–608.

Ana Sabina Uban and Liviu P. Dinu. 2020. Automatically building a multilingual lexicon of false Friends with no supervision. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3001–3007, Marseille, France. European Language Resources Association.

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.