Automatic Discrimination between Inherited and Borrowed Latin Words in Romance Languages

Alina Maria Cristea, Liviu P. Dinu, Simona Georgescu,
Mihnea-Lucian Mihai, Ana Sabina Uban
University of Bucharest
alina.cristea@fmi.unibuc.ro, ldinu@fmi.unibuc.ro,
simona.georgescu@lls.unibuc.ro,
mihnea.mihai@gmx.com, ana.uban+acad@gmail.com

Abstract

In this paper, we address the problem of automatically discriminating between inherited and borrowed Latin words. We introduce a new dataset and investigate the case of Romance languages (Romanian, Italian, French, Spanish, Portuguese and Catalan), where words directly inherited from Latin coexist with words borrowed from Latin, and explore whether automatic discrimination between them is possible. Having entered the language at a later stage, borrowed words are no longer subject to historical sound shift rules, hence they are presumably less eroded, which is why we expect them to have a different intrinsic structure distinguishable by computational means. We employ several machine learning models to automatically discriminate between inherited and borrowed words and compare their performance with various feature sets. We analyze the models' predictive power on two versions of the datasets, orthographic and phonetic. We also investigate whether prior knowledge of the etymon provides better results, employing n-gram character features extracted from the word-etymon pairs and from their alignment.

1 Introduction and Related Work

> *"When a foreign word falls by accident into the fountain of a language, it will get driven around in there until it takes on that language's colour."*
>
> — Jakob Grimm; cited by Campbell (1998)

All the world's languages are subjected to contact-induced linguistic change (Chamoreau and Léglise, 2012; Grant, 2020). A base assumption of historical linguistics (HL) is that the sound changes throughout a language's evolution were systemic in nature and produced relatively predictable results. For a long time, this hypothesis has been mainly investigated with comparative linguistics methods (Meillet, 1925; Campbell, 1998), which required a lot of manual work and extensive knowledge, and enabled significant advances in many languages.

The last decades have brought a series of computational approaches to many topics of HL, such as the problem of automatically identifying cognate pairs (Kondrak, 2001; Mulloni and Pekar, 2006; Ciobanu and Dinu, 2014; List et al., 2017; List, 2019; Heggarty, 2021), reconstructing protowords (Oakes, 2000; Bouchard-Côté et al., 2009; Ciobanu and Dinu, 2018; Meloni et al., 2019), predicting etymology (Wu and Yarowsky, 2020), discriminating between cognates and borrowings (Ciobanu and Dinu, 2015; Tsvetkov et al., 2015) or identifying lexical borrowings in a language (Miller et al., 2020; Koo, 2015).

Identifying lexical borrowings is considered one of the most difficult and important problems in HL (Carling et al., 2019; Jäger, 2019), for which "the computerised approach" is regarded as the appropriate solution even by classical linguists (Heggarty, 2012). Besides the classical distinction between borrowed words and cognates, another important problem in HL is discriminating between inherited and borrowed words (Campbell, 1998).

We shall approach the distinction between inherited and borrowed Latin words in the Romance languages (Romanian, Italian, French, Catalan, Spanish and Portuguese), with the aim of investigating whether we can automatically discriminate between the two categories, defined as follows:

– Inherited words: lexemes that have been preserved from the mother tongue in the vernacular languages by uninterrupted oral usage, taking thus part in the process of language formation; in the case of the Romance languages, we can only speak of inherited words when referring to Latin lexemes that have been part of their vocabulary ever since their "birth" (an outcome of the diversification of

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2845–2855

	Catalan	French	Italian	Spanish	Portuguese	Romanian
Inherited	dret	droit	dritto	derecho	direito	drept
Borrowed	direct	direct	diretto	directo	direto	direct

Table 1: A Latin word – *directus* (meaning *right/direct*) – both inherited and borrowed in all six Romance languages.

Latin, that had already started in the Roman period – i.e. before the 5th century AD (cf. Lausberg (1969); Adams (2007));

– Borrowed words (also known as 'loanwords'): lexical items that have been adopted in language A from language B after the language A had completed its formation period (cf. Reinheimer Ripeanu (2001)); we shall thus speak of Latin borrowings when referring to those words that, still being of Latin origin, have penetrated the Romance languages in a later period, most of them not before the 12th century AD.

There is a considerable number of cases where the same Latin word has been both inherited and borrowed. For instance, Ro. *drept* (meaning *right*), It. *dritto*, Fr. *droit*, Ca. *dret*, Es. *derecho*, Pt. *direito* are all inherited from Lat. *directus*. On the other hand, Ro. *direct* (meaing *direct*), It. *diretto*, Fr. *direct*, Ca. *direct*, Es. *directo*, Pt. *directo* have been borrowed from the same etymon, Lat. *directus*, in a period that varies from the 13th century for French, to the 19th century for Romanian (see Table 1). Most of the Latin borrowings are the effect of the so-called "relatinization" of Romance languages (starting as early as the 13th century in Western Europe): in this case, the relation between the Romance languages and Latin – as a non-contemporary source of lexical enrichment – does not count as genetic, but artificial, resulting in learned words (cf. (Reinheimer Ripeanu, 2004)).

Given the twofold relationship between Latin and the Romance languages, it is not always easy to distinguish between an inherited and a borrowed word by using only the classical methods, and the disputes between linguists increase proportionally with the uncertain cases. The importance of this subject is manyfold, having implications in important HL research problems such as protolanguage reconstruction, word dating (Campbell, 1998, pp. 299, 315, 328), or socio-cultural reconstruction (Epps, 2014).

Firstly, while we try to reconstruct a protolanguage – in this case, Protoromance –, it is essential to compare only the inherited words that form an etymological series (knows as 'real cognates'), putting aside all the borrowings (or 'virtual cognates') that may interfere and thus lead to a false protoword reconstruction.

Secondly, the distinction between inherited and borrowed words can facilitate the process of word dating, by automatically placing a lexeme among the ones that were part of a certain Romance language lexicon from the very beginning or among the lexical items that penetrated in a later period.

From a socio-cultural point of view, a word's status can shed light on the speakers' conceptual universe, by allowing us to reconstruct their everyday talk: a word is inherited only if the concept it verbalizes is needed, and it is not borrowed unless at some point the concept becomes necessary. Thus, by carefully separating between these categories, we find evidence of what topics concerned people at certain points in time.

Although linguists have successfully applied the comparative method to build classifications – to distinguish between "internal and external change" (Pat-El, 2013) – there is a fine line between the two categories and we consider that a computational method could aid in better predicting the expected classification of a term given its intrinsic structure. Since the unique application of the traditional methods has still left many uncertainties concerning the status of Romance words (easily noticeable in the Romance dictionaries), we investigate whether by applying machine learning algorithms to this problem the distinction between the two categories becomes more easily detectable.

The research shows that there is an inherent distinction between inherited and borrowed Latin words. Further introspection of the models reveals relevant features which provide useful information to linguists. The tools could be used for parallel investigations in different linguistic families, by automatically showing which category a given word "fits" better.

2 Methodology

Borrowings from Latin are supposedly easily recognized because they are presented in forms close to the Latin form in all Romance languages. For example, Lat. *attestationem* (meaning *testimony/certification*) became *attestazione* (It.), *attestation* (Fr.), *atestació* (Ca.), *atestación* (Es.), *atestação* (Pt.); Lat. *auctor* (meaning *author*) became *autore* (It.), *auteur* (Fr.), *autor* (Es., Pt.); Lat. *cultura* (meaning *culture*) became *cultura* (It., Es., Pt.), *culture* (Fr.).

There are formal differences between inherited and borrowed words from Latin: the inherited lexemes have undergone an evolutionary process that has changed their phonetic appearance (e.g. Lat. *noctem* (meaning *night*) > Ro. *noapte*, Fr. *nuit*, Sp. *noche*), while the borrowed words are generally only adapted to the Romance languages' system (cf. Reinheimer Ripeanu (2001)). Having entered the language at a later stage, borrowings are presumably less eroded and thus consistently exhibit different phonetic features, which is why we expect them to have a different intrinsic structure distinguishable by computational means. As previously discussed, in Table 1 we give an example of a Latin word both inherited and later borrowed in all Romance languages, where it can be noticed that the borrowed terms more closely resemble the original.

From a morphological point of view, one cannot reveal systematic distinct features that characterize the inherited words versus the borrowed ones. The Romance nouns' form – be they inherited or borrowed – is, in the great majority of cases, based on the accusative-ablative structure of the Latin word: for instance, both Es. *razón* (inherited, meaning *reason*) and Es. *ración* (borrowed, meaning *portion*) are originated in the Latin accusative-ablative *ratione(m)* (meaning *calculation/proportion*). It is true, though, that in a few cases of borrowing the adoption of the nominative form results in the presence of word endings that are not attested in inherited words: e.g. -o in French (*écho*, *lumbago*), -i in Italian for feminine singular nouns (e.g. *aferesi*, *crisi*), or -u in Spanish (e.g. *espíritu*, *ímpetu*).

Given that the phonetic form is the interface that we shall mainly consider in our attempt to automatically distinguish between inherited and borrowed words, we need to make a preliminary statement concerning the relation between orthography and pronunciation in the studied languages. Among the Romance idioms, only French has a deep orthography and the most conservative spelling system, while the others use a phonemic orthography. Although all the Romance languages have preserved certain orthographic traits that, far from reflecting the current pronunciation, encode historical features, French is the only language where this characteristic is general and defining. Consequently, for an accurate result we must compare both the actual phonetic transcription and the approximation of phonetic structure by orthography.

To approach our research question, we apply various machine learning models in two scenarios: in the first one we are looking only at the surface forms of inherited and borrowed words, without any other helpful supplementary information, and in the second one we have access to the etymon of the modern Romance words as well.

2.1 Algorithms

We experiment with several machine learning algorithms for the binary classification task of discriminating between inherited and borrowed Latin words: Random Forests (RF), Gradient Boosting (GB), Multi-layered Perceptron (MLP), XGBoost, Recursive Neural Networks (RNN) and Support Vector Machines (SVM). For SVM we used the radial basis function kernel (RBF), which maps samples non-linearly into a higher dimensional space, being thus able to handle the case when the relation between class labels and attributes is non-linear. Given two instances x_i and x_j, where $x_{i,j} \in \mathbb{R}^n$, the RBF kernel function for x_i and x_j is defined as follows:

$$K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2), \gamma > 0,$$

where γ is a kernel parameter. The RNN model is a character-level BiLSTM with attention with 32 units, where input characters are encoded with embedding layer of 16 units. We use dropout (with 0.1 probability) for regularization, and a learning rate of 0.005. We put our system together using several machine learning frameworks: Weka (Hall et al., 2009), Scikit-learn (Pedregosa et al., 2011), TensorFlow (Abadi et al., 2015) and Keras (Chollet et al., 2015). We split the data in two stratified subsets, for training and testing, with a 3:1 ratio, and we perform grid search and 3-fold cross validation over the training set in order to optimize hyper-parameters.

2.2 Features

For the first experiment, we use as input only the modern word forms, without knowledge of the Latin etymon. In this case, we use n-gram features (character n-grams with $n \in \{1, 2, 3\}$).[1] We mark the beginning and the end of the words with a special character $. For the second experiment, we include the Latin etymon in the input data, along with the modern word forms in Romance languages. We experiment with n-gram features extracted around mismatches in the aligned pairs (Ciobanu and Dinu, 2019), using the Needleman and Wunsch (1970) alignment algorithm. In case of multiple alignments with equal scores, we choose the first one. We also use the edit distance (Levenshtein, 1965) between the words and their etymons as an additional feature. In Figure 1 we provide a workflow example for obtaining n-gram features for the Spanish word *sacudir* (meaning *to shake off*) inherited from the Latin word *succutere*. With this approach, the system could capture transformations that occur much more often in inherited words than in borrowed words (such as letter *t* from Latin becoming *d* in Spanish) or the reduction of double consonants (such as *cc* from Latin becoming *c* in Spanish).

In addition, we apply a set of diachronic general features that characterize the sound evolution of inherited words in the Romance languages. We focused on the consonant shifts that can be defined as "sound laws" in the transition from Latin to the Romance languages, leaving aside the vowel behavior, which includes a larger number of variables difficult to systematize.

We synthesize the consonant shifts by treating them as part of a general process of lenition ("weakening"), which overarches most of the particular "sound laws" identifiable in the different Romance languages. The opposite process, of fortition, is much less frequent in the Romance languages and cannot be circumscribed to certain phonetic contexts. Intervocalic consonants (or consonant + *R*) are prone to undergo a process of lenition, materialized as a transition from a stronger articulation to a weaker one; their recurrent trajectory can thus be defined as: voiceless occlusive (*p/t/k*) → voiced occlusive (*b/d/g*) → voiceless affricate (*ts/tS*) → voiced affricate (*dz/dZ*) → voiceless fricative (e.g. *f, s*) → voiced fricative (e.g. *v, z*)→ glide (*w, j*) → disappearance; it goes without saying that it is not necessary for a consonant to go through all the stages involved by the process of lenition, easily skipping steps: e.g. *p* → *b* → *v*, cf. Lat. ***ri**p**a*** (meaning *bank*) > Es. ***ri**b**a*** // Fr. ***ri**v**e***; Lat. ***ca**p**ra*** > Es. ***ca**b**ra*** // Fr. ***chè**v**re***. The weakening process can involve the loss of the original place of articulation, leading to palatalization (the change into a palatal sound) and assibilation (the change into a sibilant): e.g. *k* → *tʃ*(→ *ʃ*), cf. Lat. *caput* (meaning *head*) > Fr. ***ch**ef* [ʃef] // *k* → *ts* → *s*, cf. Lat. *caelum* [kelum] (meaning *sky*) > Fr. *ciel* [siel]. The lenition process includes as well the simplification of geminate consonants: e.g. *pp* → *p*, Lat. *cu**pp**a* (meaning *cask*) > Ro. *cu**p**ă*, Es. *co**p**a*. The consonant shifts are represented in Table 2.

Taking this general recurrent trajectory as a starting point, we extract all possible particular sound shifts and encoded them as binary features with 0/1 values denoting their presence or absence in the input words.

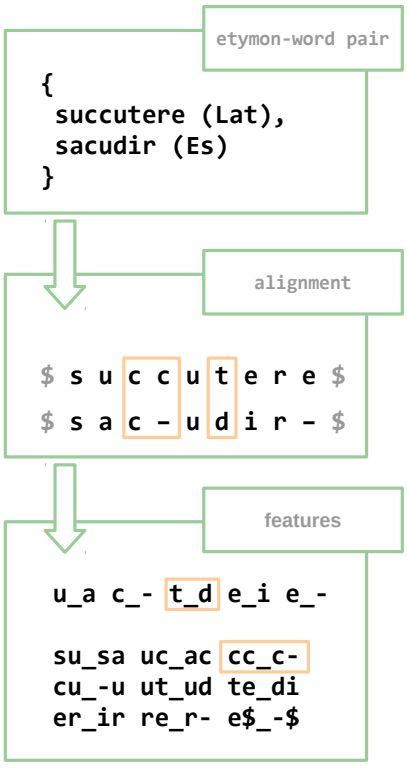

Figure 1: An example for obtaining n-gram features ($n \in \{1, 2, 3\}$) for the Spanish word *sacudir* (meaning *to shake off*) inherited from the Latin word *succutere*. Highlighted in red are transformations that occur much more often in inherited words than in borrowed words and might have high discriminative power.

[1] We ran cross-validations experiments on the training set with different ranges of n-grams with $n \in \{1, 2, 3\}$ and the results are reported for the optimum configuration.

FORTIS				————-»»»»»»»			LENIS	
Voiceless occlusives	**Voiced occlusives**	**Voiceless affricates**	**Voiced affricates**	**Voiceless fricatives**	**Voiced fricatives**	**Nasals**	**Liquids**	**Glides**
p	*b*			*f*	*v* *β*	*m*		*w*
t	*d*	*ts*	*dz*	*s*	*z* *θ* *δ*	*n* *ñ*	*l* *λ* *r*	
k	*g*			*h*	*γ*	*n* velar		*j*
		tʃ	*dʒ*	*ʃ*	*ʒ*			

Table 2: Consonant shifts: "sound laws" in the transition from Latin to the Romance languages.

2.3 Phonetic Transcriptions

We obtain automatic phonetic transcriptions for all datasets using the eSpeak NG[2] library. This tool employs a mainly rule-based approach with lookup enrichment for exceptions and annotations, and has been successfully used in previous historical linguistics applications.

Since we make use of phonetic transcriptions for the Romance languages, we consider that a similar processing of the Latin data would be appropriate. Using the comparative method, linguists were able to very reliably define the phonetic representation of Latin. It is proved that the written variety of Latin, used in the majority of etymological works, sometimes obscures the phonetic form of words – that is, the one that is truly inherited –, as well as the true relation between Romance cognates (real vs. virtual). On the contrary, since the Romance languages come from the spoken (oral) Latin language and not from the classical one as registered by dictionaries, it is preferable for our investigation to take as a starting point the phonetic representation of Latin. This method was also adopted by the Romance etymological dictionary, Dictionnaire Ètymologique Roman.[3]

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

[3]DÉRom, cf. www.atilf.fr/DERom

3 Experiments

In this section we describe and discuss experiments on automatically discriminating between inherited and borrowed Latin words.

Data	Language	Features
Wiki	Catalan	ió ci ó$ $i ll ac ic ia ct di
	French	ou io ti at $i st on ch ic ré
	Italian	zi ne on az io ul $i si cl pl
	Portuguese	çã lh ic ia ul ei ha nh ci ão
	Spanish	ió ci ón n$ ic $i ac ll ct ul
	Romanian	ți en on it $e il ie an $i ân
DEX	Romanian	ân $î on it il en ți în i$ $e

Table 3: Most informative bi-gram features (highest entropy) for each dataset and language.

3.1 Data

We extract datasets of inherited and borrowed words from Wiktionary,[4] which provides Wikitext templates that systematically specify etymological information, taking into account the original inherited word forms as well (for example, accusative Latin structures instead of the dictionary nominative forms). We capture etymons using regular expressions and, scraping the latest database version, we obtain datasets for all six Romance

[4]https://www.wiktionary.org

Data	Language	Size			Avg word length			Avg word-etymon dist		
		inherited	borrowed	all	inherited	borrowed	all	inherited	borrowed	all
Wiki	Catalan	1,536	889	2,425	5.36	7.42	6.12	0.46	0.28	0.39
	French	2,003	2,367	4,370	5.91	7.87	6.97	0.54	0.31	0.42
	Italian	3,087	1,585	4,672	6.74	8.00	7.17	0.38	0.28	0.34
	Portuguese	1,972	1,672	3,644	5.77	7.48	6.55	0.48	0.31	0.40
	Spanish	2,283	1,795	4,078	6.07	7.71	6.80	0.50	0.29	0.40
	Romanian	2,104	859	2,963	5.60	7.16	6.05	0.56	0.28	0.48
DEX	Romanian	1,397	4,631	6,028	5.39	7.69	7.16	0.48	0.25	0.30

Table 4: Dataset characterization: distribution of inherited and borrowed words, average word length and average normalized edit distance between modern words and their etymons. Values are computed only on the training subset and are reported in the following format: inherited | borrowed | all (total).

languages investigated. Additionally, for Romanian we also prepare a comprehensive and accurate dataset extracted from the digitalized version of the language's most reputed dictionary DEX,[5] taking advantage of structural regularities used in the etymology section. We use two versions of the dataset – one raw and another one with several linguistically-motivated edits.[6] In Table 3 we report the most informative bigram features for each language (obtained based on entropy) and in Table 4 we provide a characterization of the extracted datasets, which we make available publicly.[7]

The datasets do not include information about the time of borrowing. The borrowed words are not likely to show very different characteristics even if borrowed at different times, but this also depends on the language: a word that was borrowed in French in the 13th century has undergone more sound shifts than another word borrowed in the 20th century. On the other hand, a Latin word borrowed in Spanish in the 13th century did not experience severe phonetic changes, because most of the "sound laws" specific to Spanish had already ended their active period by that time. The significant phonetic changes in the the Romance languages had mostly taken place before Latin borrowings started entering their lexicons, the "relatinization" process coinciding with the official attempts to normalize the vernacular languages. Once standardized, the Romance languages slowed down the process of change, thus preventing the newly borrowed Latin words to undergo a noteworthy formal evolution. For Spanish, to give an example, one cannot identify any systematic formal features that would allow us to distinguish between earlier and later borrowings from Latin.

3.2 Baselines

We compare our results with two baselines: a majority class baseline that always predicts the most frequent label (accounting, thus, for the class imbalance) and a more informed baseline – a decision tree classifier with only one node that uses the edit distance between the modern word and its etymon

Data	Language	B1	B2
Wiki	Italian (ort)	66.0	65.3
	Italian (phon)	66.0	65.3
	Portuguese (ort)	54.1	69.0
	Portuguese (phon)	54.1	62.6
	Catalan (ort)	63.3	69.0
	Catalan (phon)	63.3	62.6
	Spanish (ort)	55.9	73.4
	Spanish (phon)	55.9	57.2
	French (ort)	54.1	80.3
	French (phon)	54.1	70.2
	Romanian (ort)	70.9	81.2
	Romanian (phon)	70.9	70.8
DEX	Ro (raw, ort)	76.7	84.1
	Ro (raw, phon)	76.7	76.7
	Ro (edit, ort)	76.7	79.7
	Ro (edit, phon)	76.7	76.7

Table 5: Baselines accuracy for discriminating between inherited and borrowed words.

[5] https://dexonline.ro/sursa/dex09

[6] We apply linguistic normalization techniques such as deleting the final *-s/-m* from Latin roots, appending a historically accurate *-u* to Romanian nouns – e.g., *foc* (meaning *fire*) > *focu*, cf. Es *fuego* < Lat *focus* – and the *-re* long infinitive for Romanian verbs which mirrors the Latin etymon.

[7] https://nlp.unibuc.ro/projects/cotohili.html

Data	Language	RF		GB		SVM		RNN		SVM (+ etymons)	
		F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
Wiki	Italian (ort)	81.3	82.2	81.9	82.6	84.5	84.9	83.8	84.2	**86.0**	**86.0**
	Italian (phon)	81.7	82.5	81.8	82.2	81.3	82.1	80.0	80.3	**85.9**	**86.0***
	Portuguese (ort)	82.2	82.3	84.3	84.4	84.6	84.7	81.7	81.7	**85.7**	**85.6**
	Portuguese (phon)	84.9	85.0	86.0	86.0	83.1	83.2	82.0	82.0	**86.2**	**86.2***
	Catalan (ort)	84.4	85.1	84.2	84.8	86.1	86.4	83.4	83.3	**91.7**	**91.7***
	Catalan (phon)	84.2	84.8	85.2	85.6	86.0	86.3	86.1	86.3	**89.4**	**89.5***
	Spanish (ort)	83.9	84.1	83.6	83.7	86.2	86.2	80.9	80.9	**88.5**	**88.4***
	Spanish (phon)	82.8	82.9	82.8	82.8	86.1	86.1	79.0	79.0	**87.9**	**87.9***
	French (ort)	87.9	87.0	87.6	87.6	88.3	87.6	86.4	86.5	**91.0**	**90.9***
	French (phon)	83.7	83.7	85.9	85.9	86.8	86.7	84.0	84.0	**90.8**	**90.8***
	Romanian (ort)	87.3	88.1	87.8	88.2	89.3	89.6	83.0	84.4	**90.5**	**90.6**
	Romanian (phon)	89.0	89.6	86.9	87.5	90.2	90.4	86.6	87.0	**90.8**	**90.9**
DEX	Romanian (raw, ort)	90.7	91.1	91.0	91.3	91.6	91.6	90.9	90.1	**93.6**	**93.6***
	Romanian (raw, phon)	90.4	90.7	91.3	91.5	92.1	92.1	92.5	92.6	**94.0**	**94.0***
	Romanian (edit, ort)	90.3	90.7	91.0	91.2	92.1	92.0	92.2	92.3	**92.9**	**92.9**
	Romanian (edit, phon)	90.5	90.8	91.6	91.8	92.2	92.2	**95.2**	**95.2**	93.5	93.5

Table 6: Results for automatic discrimination between inherited and borrowed Latin words (orthographic -ort, and phonetic -phon). The last column represents SVM results using features extracted from the word-etymon pairs. We marked with * accuracy results for which the difference to SVM without etymons is statistically significant (99% confidence level, performed on 10,000 iterations of bootstrap resampling (Koehn, 2004)).

as single feature. The latter baseline is motivated by the observation (reported in Table 4 on the training subset) that borrowings are generally closer to the form of their etymon than inherited words.

3.3 Results

In Table 5 we report the results of the two baselines. The more informed baseline (B2) outperforms the majority class baseline (B1) in most cases. In Table 6 columns "RF", "GB", "RNN", "SVM" we report the results of our systems in the first scenario – using only the surface forms of the modern Romance words as input. We report results on Wiktionary datasets for six Romance languages and on the additional DEX dataset for Romanian for two versions of each dataset – orthographic and phonetic. We measure the performance of the models with the accuracy and weighted average F1 values (that is, the average is weighted by the number of true instances for each class, taking thus the class imbalance into account).[8]

Comparing results from Table 5 and Table 6, we observe that the proposed systems outperform both baselines significantly, obtaining an increase of up to $\sim$ 36 percentage points over the first one, and up to $\sim$ 20 percentage points over the second one.

The best results are obtained by SVM in most cases. The high performance (F1 between 84.5 and 92.1 at orthographic level and between 81.3 and 92.2 at phonetic level) shows that there are discriminating features that can be learned automatically. We attribute the lower results of the RNN compared to some of the other models on the Wiktionary data to the insufficient data size compared to the model's complexity. A similar RNN architecture was previously used by Miller et al. (2020) for identifying lexical borrowings in monolingual wordlists. In their setup, RNN was reported to perform best, while in our setting RNN was, in most cases, outperformed by the SVM system using features extracted from the etymons.

For the most part, the results at phonetic and orthographic level are comparable. The best results (in F1 terms) on Wiktionary data are obtained for Catalan, followed by French, Romanian, Spanish, Portuguese, and Italian. As a general observation, the inherited words are classified better than the

[8] Since the MLP and XGBoost models did not outperform the best classifiers, we omit them from the table due to lack of space.

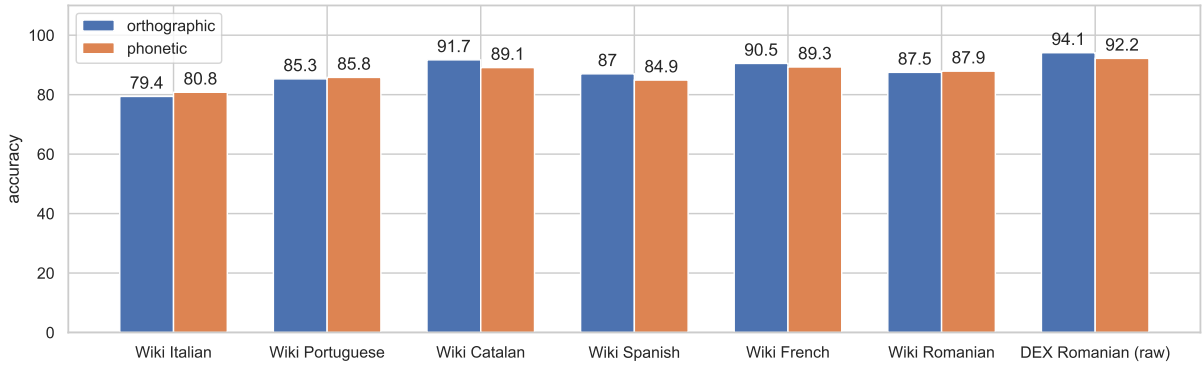

Figure 2: Accuracy for discriminating between inherited and borrowed words on balanced subsets of equal size for all languages, for the best performing system, SVM (+ etymons).

borrowed words.[9] We do not consider this to be caused by the unbalanced datasets (more training material available for inherited words), because an additional experiment with equal training/test data sizes across all languages exhibited the same behavior. Moreover, this is not the case for French, where we have more borrowings than inherited words in the dataset, but the accuracy is still better for the inherited words. The accuracy obtained for equal subsets (850 borrowings and 850 inherited words, split for training/test with a 3:1 ratio) is reported in Figure 2. We observe that, for some languages, the results with orthographic forms are better than with phonetic forms. The orthography tends to be conservative, which allows an easier confrontation between the Romance lexemes and their etymons, hence a better automatic interpretation of the sound evolution. At the same time, the orthographic form facilitates the direct observation of the degree of proximity between the etymon and its Romance descendants, thus allowing its inclusion in the right category. The phonetic form can sometimes distort its automatic interpretation, as the pronunciation is always ahead of the orthography, and can, not infrequently, coincide with the result of the sound laws that intervened in the evolution of the inherited form.

In Table 6 column "SVM (+ etymons)" we report the results of our best-performing system in the second scenario – using the {word, etymon} pairs as input (F1 between 85.7 and 93.6 at orthographic level and between 85.9 and 93.5 at phonetic level). We have experimented with different combinations of features (described in Section 2.2)

[9]Due to lack of space, we report here only the average F1 score; the confusion matrix shows that more borrowed words were incorrectly classified as inherited words than vice-versa.

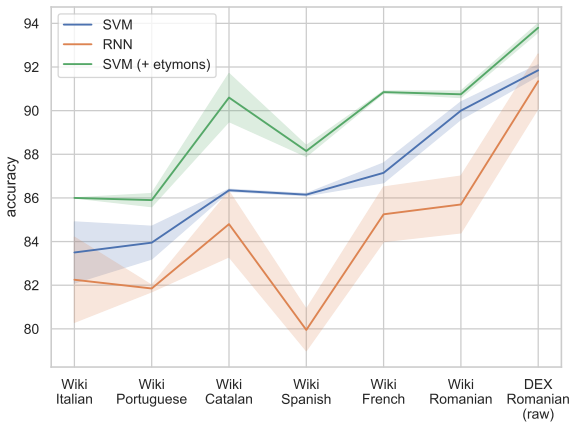

Figure 3: Accuracy for discriminating between inherited and borrowed words: SVM, RNN, SVM (+ etymons).

and we report here the best performing combination, which includes all features: n-grams extracted from the modern words, n-grams extracted from the alignment of the words and their etymons, the edit distance and the linguistic features regarding consonant shifts. We report results on DEX for Romanian with and without linguistically-motivated edits, and both versions in orthographic and phonetic form. The top 3 performing systems are also represented in Figure 3, for a better visualisation. This setup outperforms our previous results for all datasets except for DEX phonetic form, with linguistic edits. Introducing the etymons in the input data lead to a performance increase of $\sim$ 2 percentage points, which was further slightly improved by the linguistic edits. Taking a closer look at the misclassified instances (see Figure 4) we observe that, overall, the edit distance between the misclassified borrowed words and their etymons does not differ significantly from the edit distance be-

tween the misclassified inherited words and their etymons.The difference in edit distance between correctly classified inherited and borrowed words is not significant either.

Figure 4: Normalized edit distance between words and their etymons for misclassified instances.

3.4 Error Analysis

Upon examining the predictions from the held-out test sets, we are able to identify three underlying error sources.

Lack of distinguishing information: Some words are simply not characterised by distinguishing features. While specific discriminative features exist for both classes, it is not guaranteed that each and every test sample will exhibit any such feature. For example, the Romanian term *suc* (meaning *juice*; misclassified by our model as inherited) is actually borrowed from Latin *succus* via the French *suc*. However, had *succus* descended directly from Latin, the result according to the comparative method would still have been *suc*, which shows the limitations of a purely phonetic-based model.

Influence of existing words on borrowings (so-called semi-learned borrowings): A significant number of Latinate borrowings in Romance languages have been artificially influenced by already existing terms inherited from the same Latin root. As such, they phonetically resemble an inherited word despite not being actually inherited. Our model misclassified French *discourir* (meaning *to discourse/talk*) as inherited (although it is borrowed from the Latin *discurrere*) because it was heavily adapted according to the inherited *courir* $<$ *currere*. Another example of phonetic assimilation is the Romanian word *demn* (meaning *dignified*; misclassified as inherited although it is borrowed from the Latin *dignus*), because its phonetic form was heavily altered under the pressure of other inherited roots such as *semn* $<$ *signum* (meaning *sign*) or *lemn* $<$ *lignum* (meaning *wood*). This influence simulates the term having suffered the same diachronic sound shifts although it was not present in the language at that time.

Disputed etymologies: Our model classified the Spanish term *clavo* (meaning *nail*) as borrowed, although it is directly inherited from the Latin *clavus*. This mistake is actually not a fully detrimental trait of the model, because it proves the model learned expected phonetic behaviour, as linguists themselves struggled to explain why the initial consonant cluster *cl-* failed to shift into *ll-* as is usually the case with inherited Spanish words.

4 Conclusions

In this paper we have analyzed the automatic discrimination between inherited and borrowed Latin words in Romance languages, both in orthographic and phonetic form. We have obtained an average F1 over all languages $\sim 90\%$ at orthographic level. We have built a dataset of inherited and borrowed Latin words from two sources (Wiktionary and DEX) in multiple Romance language (Catalan, French, Italian, Portuguese, Romanian, Spanish). We have augmented the data with features provided by linguists in order to increases the system's performance, based on the idea that the optimal approach to computational historical linguistics is to combine the experience and intuitions of linguists with the intelligent processing and automation capabilities of computational tools.

Ethics Statement

All our data are extracted from publicly available sources. There are no ethical issues in our work.

Acknowledgments

We would like to thank the reviewers for their helpful comments. All authors contributed equally to this work. This research is supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI UEFISCDI, project number 108, COTOHILI, within PNCDI III.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

James Noel Adams. 2007. *The regional diversification of Latin 200 BC – AD 600*. Cambridge University Press.

Alexandre Bouchard-Côté, Thomas L. Griffiths, and Dan Klein. 2009. Improved Reconstruction of Protolanguage Word Forms. In *Proceedings of NAACL 2007*, volume 7, pages 65–73.

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

Gerd Carling, Sandra Cronhamn, Robert Farren, Elnur Aliyev, and Johan Frid. 2019. The causality of borrowing: Lexical loans in eurasian languages. *PLOS ONE*, 14(10):1–33.

Claudine Chamoreau and Isabelle Léglise. 2012. *A multi-model approach to contact-induced language change*, pages 1–16. De Gruyter Mouton.

François Chollet et al. 2015. Keras.

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. 2018. Ab initio: Automatic latin proto-word reconstruction. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1604–1614. Association for Computational 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.

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

Anthony P. Grant. 2020. Contact-Induced Linguistic Change: An Introduction. In *The Oxford Handbook of Language Contact*. Oxford University Press.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

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. 2021. Cognacy databases and phylogenetic research on indo-european. *Annual Review of Linguistics*, 7(1):371–394.

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

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 388–395.

Grzegorz Kondrak. 2001. Identifying Cognates by Phonetic and Semantic Similarity. In *NAACL*.

Hahn Koo. 2015. An unsupervised method for identifying loanwords in korean. *Lang. Resour. Evaluation*, 49(2):355–373.

Heinrich Lausberg. 1969. *Romanische Sprachwissenschaft, 3-vol*. Berlin, De Gruyter.

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

Johann-Mattis List. 2019. Automatic Inference of Sound Correspondence Patterns across Multiple Languages. *Computational Linguistics*, 45(1):137–161.

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.

A. Meillet. 1925. *La Méthode Comparative en Linguistique Historique*. H. Aschehoug & Co. Oslo.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2019. Ab Antiquo: Proto-language Reconstruction with RNNs. *CoRR*, abs/1908.02477.

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.

Andrea Mulloni and Viktor Pekar. 2006. Automatic detection of orthographic cues for cognate recognition. In *In Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, pages 2387–2390.

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.

Michael P. Oakes. 2000. Computer Estimation of Vocabulary in a Protolanguage from Word Lists in Four Daughter Languages. *Journal of Quantitative Linguistics*, 7:233–243.

Na'ama Pat-El. 2013. Contact or Inheritance? Criteria for distinguishing internal and external change in genetically related languages. *Journal of Language Contact*, 6:313–328.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sanda Reinheimer Ripeanu. 2001. *Lingvistica Romanica: Lexic, Morfologie, Fonetica*. Ed. All. Bucuresti.

Sanda Reinheimer Ripeanu. 2004. *Les emprunts latins dans les langues romanes.* Editura Universităţii din Bucureşti.

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

Winston Wu and David Yarowsky. 2020. Computational Etymology and Word Emergence. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3252–3259.