Using lexical language models to detect borrowings in monolingual wordlists

John E. Miller[1*], Tiago Tresoldi[2+], Roberto Zariquiey[3], César A. Beltrán Castañón[1], Natalia Morozova[2], Johann-Mattis List[2]

1 Artificial Intelligence/Engineering, Pontificia Universidad Católica del Perú, San Miguel, Lima, Peru

2 Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History, Jena, Germany

3 Linguistics/Humanities, Pontificia Universidad Católica del Perú, San Miguel, Lima, Peru

* jemiller@pucp.edu.pe + tresoldi@shh.mpg.de

Abstract

Native speakers are often assumed to be efficient in identifying whether a word in their language has been borrowed, even when they do not have direct knowledge of the donor language from which it was taken. To detect borrowings, speakers make use of various strategies, often in combination, relying on clues such as semantics of the words in question, phonology and phonotactics. Computationally, phonology and phonotactics can be modeled with support of Markov n-gram models or – as a more recent technique – recurrent neural network models. Based on a substantially revised dataset in which lexical borrowings have been thoroughly annotated for 41 different languages of a large typological diversity, we use these models to conduct a series of experiments to investigate their performance in borrowing detection using only information from monolingual wordlists. Their performance is in many cases unsatisfying, but becomes more promising for strata where there is a significant ratio of borrowings and when most borrowings originate from a dominant donor language. The recurrent neural network

performs marginally better overall in both realistic studies and artificial experiments, and holds out the most promise for continued improvement and innovation in lexical borrowing detection. Phonology and phonotactics, as operationalized in our lexical language models, are only a part of the multiple clues speakers use to detect borrowings. While improving our current methods will result in better borrowing detection, what is needed are more integrated approaches that also take into account multilingual and cross-linguistic information for a proper automated borrowing detection.

Introduction

Problem and Motivation

Lexical borrowing, the direct transfer of words from one language to another, is one of the most frequent processes of language evolution [1]. We can easily observe the process in real time, especially regarding vocabulary from religion or technology, since words are often transferred along with other cultural practices or innovations. While it took scientists a long time to find out that languages constantly change [2], it was already clear in ancient times that languages acquire lexical material from their neighbors [3], as evidenced in Plato's *Kratylos* dialog (409d-10a) [4] where Socrates discusses the problem that lexical borrowings impose on studies in word etymology. Nonetheless, the process is still regarded as one of the outstanding problems in historical linguistics, as it needs to "infer or determine shared traits among two or more languages, and then determine conflicts in these traits, taking geographical closeness and borrowability into account" [5].

Discrimination between inherited and borrowed words (also called "loanwords") is crucial for the successful application of both the *comparative method* in historical linguistics [2], which seeks to identify genetically related languages and reconstruct their ancestral stages which are not recorded in written sources, and in *phylogenetic reconstruction*, which seeks to identify the most plausible phylogenies (often represented by a family tree) by which languages in a given language family evolved into their current shape [6]. Native speakers are often assumed to be remarkably efficient in such discrimination task [7, 8].

Similar to linguists [9], laypeople use an arsenal of different methods to detect borrowed words. When multilingual speakers observe that words denoting similar concepts sound alike in otherwise different languages, they may conclude that the similarity is due to borrowing. Even when the donor language of a word is not known, speakers may detect recent borrowings in their native language due to specific phonological or phonotactic characteristics. In many Hmong-Mien languages, for example, some Chinese words are borrowed with a very specific tone that only occurs in Chinese words [10]. Similarly, it is easy for German speakers to identify *job* as a loan from English, since only in borrowed words the grapheme *j* is pronounced as [dʒ] in German. Apart from specific sounds and tones, evidence for borrowings may include peculiar constructions, specific phonotactic elements (such as certain consonant clusters or vowel combinations), unusual stress patterns [11, 12], or even specific semantics. However, already upon entering the language, speakers adapt borrowed words to the phonological conditions of the recipient language, and the more time has passed since a word was first borrowed, the harder it is to detect it from its external characteristics alone [13]. This process, called nativization, "provides a direct window for studying how acoustic cues are categorized in terms of the distinctive features" relevant to phonology and phonotactics of the native speaker [14, p. 1].

Despite the obvious limitations of speakers' intuition about inherited and borrowed material in their native languages, it seems worthwhile to test to what degree automated borrowing detection in linguistics could be based on monolingual data alone. Assuming that the major source of native speakers' intuition regarding their native languages' lexicons lies in phonology and phonotactics, we can use computational approaches to model phonology and phonotactics derived from annotated wordlists of a given language and then calculate to which degree a word resembles a typically inherited or a typically borrowed word.

To model phonology and phonotactics of a language, we make use of different *lexical language models*. Assuming that a *language model* refers to "any system trained only on the task of string prediction, whether it operates over characters, words or sentences, and sequentially or not" [15], our *lexical language models* are specific cases of language models derived from lexical data typically provided in the form of a wordlist, with words being represented by phonetic transcriptions. Having trained lexical language

models for inherited and borrowed words with the help of a given annotated wordlist representing a given language variety in a supervised learning setting, we can then try to measure to which degree words that were not used to train a given model can be classified as either being inherited or borrowed.

In this study, we test how well three different lexical language models – one non-sequential model based on a support vector machine, and two sequential models, one based on Markov chains and one based on recurrent neural networks – perform in detecting borrowed words. We apply our models to the World Loanword Database [16], a large, cross-linguistic sample of wordlists in which borrowed words are annotated, which we considerably improved by adding harmonized phonetic transcriptions instead of the original orthographic representations of word forms.

Results in many cases are unsatisfying for borrowings attested in wordlists from the World Loanword Database, but become more promising when there is a significant ratio of borrowings, and even more so when borrowings come predominantly from a single donor language. The recurrent neural network performs marginally better than the Markov chain method in the case of borrowing from wordlists (where the support vector machine method fares poorly), and marginally better than the support vector machine in the extreme case of simulated significant borrowing from a single donor (where the Markov chain method fares less ably). A review of the distributions of differences between inherited and borrowed word entropies, the basis for Markov chain and recurrent neural network methods, indicates further opportunities for improvement and innovation.

State of the art

Although the detection of borrowed words is one of the major tasks in historical language comparison, the classical, non-computational techniques which linguists use to identify borrowings have never been properly formalized or explicitly described [9]. Similar to native speakers, who employ specific kinds of evidence (phonological, phonotactic, or semantic), classical linguists extensively use *proxies* to assess whether or not a given word has been borrowed. Apart from direct evidence, derived from the documentation of the same language at different times, these proxies include

(a) conflicts with genealogical explanations (e.g., similar words between otherwise unrelated languages), (b) conflicts within the borrowed traits (irregular sound correspondence patterns in seemingly cognate words in related languages), and (c) distributional properties of shared traits (specific semantics of a group of words in a given language) [9]. While most of the evidence linguists employ to detect borrowed words is based on the comparison of *several* languages, conflicts in phonology and phonotactics are also routinely used for borrowing detection, specifically when dealing with recent borrowing events.

Similar to the prevalence of multilingual approaches to borrowing detection in classical historical linguistics, most recent attempts to detect borrowings automatically have also been based on comparative rather than monolingual evidence. Various authors have tried to detect borrowings by searching for phylogenetic conflicts [17–23]. Other approaches identify similar words in unrelated languages [24–26]. Occasionally, authors have tried to detect borrowings by relying on the idea that some words can be more easily borrowed, because of the meanings they express [27]. While the detection of words borrowed between unrelated languages seems to work relatively well [26], all other approaches that have been proposed in the past, have never been rigorously tested.

In contrast to multilingual approaches to borrowing detection, monolingual approaches in which borrowings are identified by relying on the (annotated) data of one language alone, have been rarely applied so far, and the rare exceptions we know of, where scholars have tried to model native speakers' borrowing detection competence computationally, involve very particular settings for individual languages, as opposed to generic approaches that could be generally applied [28, 29].

Although – to our knowledge – language models have not yet been used to identify borrowings in exclusively monolingual wordlists, the idea to use lexical language models for specific tasks in comparative linguistics is not new. Language identification, for example, which seeks to identify the natural language in which a given document is written [30], shows certain similarities with the task of monolingual borrowing detection. Distinguishing foreign words within a paragraph or sentence is similar to the problem of detecting recently borrowed words in a wordlist.

Materials and methods

Materials

We use the multilingual wordlist collection provided by the World Loanword Database (WOLD) [16], which we modified by adding harmonized phonetic transcriptions. Each of the 41 wordlists in this collection provides translation equivalents for 1,460 distinct concepts (see the Concepticon resource for details on this concept list [31]). Since translations may lack or one concept may have been represented by more than one word form, the resulting wordlists comprise between 956 and 2,558 word forms. While word forms were provided in orthographic form or phonological transcriptions in the original data, we added phonetic transcriptions which follow the unified *Broad IPA* transcription system proposed by the Cross-Linguistic Transcription Systems reference catalog [32, 33] with the help of *orthography profiles* [34]. Orthography profiles can be best thought of as a specific look-up table, which allows to convert transcriptions from one orthography into another one (compare the presentation in Wu et al. [35] for details). Each word form is given a so-called *borrowed score*, indicating the rating of a linguistic expert that the item was borrowed on a five-point scale. To make sure that we only consider clear-cut borrowings in our tests, we treated as borrowed only the words which were labeled as *clearly borrowed*.

The derived database with phonetic transcriptions for all 41 wordlists was curated with the help of the CLDFBench toolkit [36], which allows for a convenient, test-based data curation workflow in which the resulting dataset is offered in the formats recommended by the Cross-Linguistic Data Formats initiative (CLDF, `https://cldf.clld.org` [37]). These format specifications have proven very useful in the past, as they allow not only for a quick aggregation of data from different sources [38], but also for their convenient integration in computational workflows [35].

For testing purposes, we created an additional German wordlist, taken from an etymological dictionary of German [39], with phonetic transcriptions added with modifications from the CELEX database [40]. While the enhanced WOLD database has been curated on GitHub (`https://github.com/lexibank/wold`) and archived with Zenodo [41], the German wordlist is available as part of the software package we wrote for monolingual borrowing detection, curated on GitHub.

Lexical language models

For the purpose of testing how well borrowed words in a wordlist can be detected through language-internal information alone, we employ three different lexical models which reflect unique characteristics of cues that native speakers could take into account when identifying borrowings in their native tongue. The Bag of Sounds method represents words internally as a set of the sounds of which they consist, the Markov Model represents words by their sound n-grams, and the Neural Network represents words in the form of sequences of learned vector representations of sounds.

We perform borrowing detection on each wordlist individually, modeling word expectedness with Bag of Sounds, Markov Model [42], and Neural Network [43] methods. The Bag of Sounds is a baseline method, which uses a support vector machine to directly detect borrowings based only on the set of sounds. The Markov Model and Neural Network produce sequential sound segment probability estimates, which we transform into word entropies and use to predict borrowed words. The Markov Model serves as the standard approach and the Neural Network as an improved alternative to borrowing detection with entropy methods. The Markov Model and Neural Network methods focus on phonotactics, while the Bag of Sounds method focuses on phonology.

Bag of Sounds

Since the word forms in our data are available as harmonized phonetic transcriptions, it is straightforward to represent each word form in a given language as a vector indicating the presence and absence of distinct sound segments. Since the order of these sound segments is not important, and neither is their frequency considered, this vector can be thought of as a simple bag of sounds, in which the sounds making up a given word form are represented as a set. The task of distinguishing borrowed from inherited words can then be pursued with the help of a support vector machine with a linear kernel [44, 45]. The support vector machine identifies the plane which optimally separates native from borrowed words based on the set of sound segments. The Bag of Sounds method does not consider the order or the frequency of elements in a given sound sequence, and we did not expect it to perform extraordinarily well in all languages in our sample. The advantage of the model is that it is simple and fast in application. It also provides a

baseline for those cases where peculiar sounds provide enough information to identify a given borrowed word.

Markov Model

An $n-1$ order Markov Model, emits a sound segment with probability dependent on the $n-1$ previous sound segments (an n-gram model). It transforms the product of sound segment probabilities estimated by the Markov Model method into word entropies which are then used in borrowing detection.

We use a second order Markov Model (emission probability dependent on the previous 2 segments – a tri-gram model) from the Natural Language Toolkit (NLTK) [46]. The second order Markov Model is local with longer range effects resulting from the second order probabilistic process.

We can approximate the probability of a sequence of sound segments by the product of the second order conditional probabilities:

$$P(c_1^n) \approx \prod_{k=1}^{n} P(c_k | c_{k-2}^{k-1}).$$

We transform word probabilities to a per sound segment word entropy,

$$H(w) = -(1/N) \log P(c_1^n),$$

which typically exhibits a smooth distribution with moderate right skew for wordlists. The second order model with a sound segment vocabulary size V requires V^3 probability parameters for sound segment emission probabilities conditioned on the previous two sound segments.

With wordlists of just 1,000 to 2,500 word forms and a typical sound segment vocabulary size of $V \approx 50$, estimating $50^3 = 125,000$ parameters by maximum likelihood would cause sparse parameter estimation with problems of both undefined conditional probabilities and overfitting. We use interpolated Kneser-Ney smoothing to accommodate unseen tri-grams, reduce overfitting, and reduce the number of estimated parameters to less than the V^3 required under maximum-likelihood.

Recurrent Neural Network

Recurrent Neural Networks provide word length order conditioning via the recurrent layer with memory. Word probabilities are expected to be better estimated, i.e., better approximating human performance, than for the Markov Model.

Conditional character probabilities are estimated based on all earlier sound segments of the current word:

$$P(c_n|c_1^{n-1}) = f(c_{n-1}, ..., c_1).$$

We can approximate the probability of a sequence of segments as the product of the segment probabilities:

$$P(c_1^n) \approx \prod_{k=1}^{n} P(c_k|c_1^{k-1}).$$

Word probabilities are again transformed to a per sound segment word entropy.

$$H(w) = -(1/n) \log P(c_1^n).$$

The challenge and advantage of the recurrent Neural Network method is in the estimation of the conditional sound segment probabilities, with the function $f(c_{n-1}, ..., c_1)$, using a more complex architecture but with fewer parameters (figure 1b) than the second order Markov model. Sparse indicator vectors, c_k, representing sound segments are transformed into dense real input vectors, x_k. In the recurrent layer, input vectors, x_k, and prior hidden state vectors, h_{k-1}, are linearly transformed and passed through a $tanh$ activation function to produce current hidden state, h_k, and output, o_k, vectors. Resulting output vectors are linearly transformed in a dense output layer of logits, y, representing possible output segments. The $softmax$ activation function transforms logit values y_k into sound segment probability estimates,

$$\widehat{P}(c_n|c_{n-1}, ...c_1) = e^{y_{c_n}} / \sum_k e^{y_k}.$$

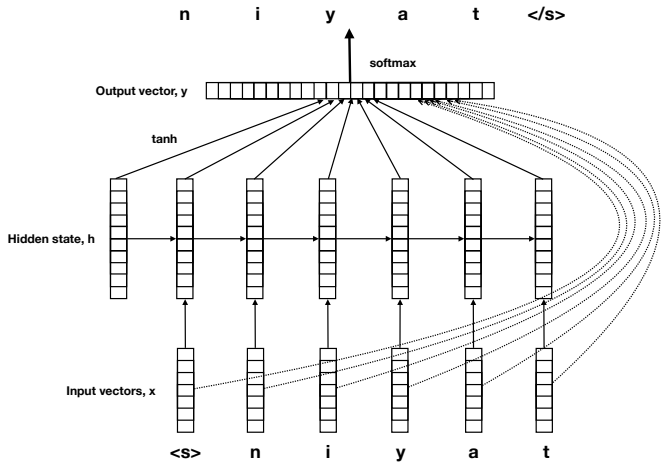
While the recurrent Neural Network model requires a high baseline number of parameters given its embedding length and recurrent layer length, the growth in number of parameters is just linear with the vocabulary size. As a result, the number of

```
# Architecture parameters
embedding_len = 32
rnn_output_len = 32
rnn_cell_type = GRU
rnn_levels = 1

# Regulation parameters
embedding_dropout = 0.0
recurrent_l2 = 0.001
rnn_activity_l2 = 0.0
recurrent_dropout = 0.0
rnn_output_dropout = 0.2
merge_embedding_dropout = 0.2

# Model fitting parameters
epochs = 45
learning_rate = 0.01
learning_rate_decay_factor = 0.95
val_split = 0.0
```

(a) Recurrent neural network configuration



(b) Recurrent neural network - lexical model architecture

Fig 1. Recurrent neural network - lexical model

parameters in the Neural Network is on the order of 10,000, and this does not change much with the vocabulary size. Furthermore, the number of parameters does not increase with word length in sound segments even though the conditioning is on all previous sound segments.

We implement our recurrent Neural Network in Tensor-Flow 2.2 [47] and parameterize the model to permit ready changes in architecture, regulation, and fitting parameters during experimentation. The configuration used in this study is shown in figure 1a. Neural network models, even with just thousands of parameters, may suffer from substantial variance between training and test due to overfitting, especially when the amount of training data is comparatively small as in this case. We apply methods of dropout and l2 regulation to reduce overfitting.

Decision procedures

Models are trained on labeled data and then used to predict whether unlabeled test words are inherited or borrowed. For the Bag of Sounds method, we train a model to distinguish borrowed from inherited words directly from sound segments. For the Markov Model and Neural Network methods, we fit models based on inherited and borrowed words separately, estimate word entropies on test data using both models, and designate the word as inherited or borrowed depending on which model has the lesser entropy.

We assume that for a model trained on inherited words, the entropy estimates for unobserved inherited words will be less than for borrowed words. Similarly, for a model trained on borrowed words, entropy estimates for unobserved borrowed words will be less than for inherited words. The choice of the model with the lesser entropy can be expressed as the difference of entropies compared to a critical value, in this case zero:

$$borrowed = (entropy(w)_{native} - entropy(w)_{borrowed}) > 0.$$

Assessing detection performance

We assess detection performance using *precision*, *recall*, and *harmonic mean* (*F1 score*), as well as *accuracy* measures based on frequency counts of borrowing detection by true borrowing status as defined in table 1. Following [48], *precision* is the proportion of true positive borrowings out of all detected positives,

$$precision = tp/(tp + fp),$$

recall is the proportion of true positive borrowings out of all borrowings,

$$recall = tp/(tp + fn),$$

F1 score is the harmonic mean of precision and recall, and

$$F1 = (2 * precision * recall)/(precision + recall),$$

accuracy is the proportion of all detections that are correct,

$$accuracy = (tp + tn)/(tp + fp + fn + tn).$$

We consider F1, since it combines both precision and recall, as the primary measure. Accuracy does not specifically focus on borrowing detection and is of secondary importance.

Borrowing Detection	True borrowing status	
	Borrowed	**Inherited**
Positive	tp=true positive	fp=false positive
Negative	fn=false negative	tn=true negative

Table 1. Frequency counts of borrowing detection by true borrowing status.

Experiments and studies

We run several experiments and studies as follows. First, we simulate detection of recent borrowings by artificially seeding wordlists with various proportions of words from a foreign language and then apply borrowing detection methods to test detection performance. Second, we test borrowed word detection more realistically by using wordlists without alteration and performing a 10-fold cross validation of borrowed word detection. Third, we perform correlation analysis to diagnose real world performance as a function of phonological variables of the wordlists. Fourth, we stratify language wordlists by number of borrowed words and presence of a dominant donor language and analyze the 10-fold cross validation of borrowed word detection by strata. Last, we examine entropy distributions for a few exemplary wordlists, and see how the entropy method works.

Implementation

Methods for borrowing detection and evaluation have been implemented in the form of a Python package and is available along with supplemental information accompanying this study at `https://osf.io/69ak5/`. The Python package contains the code, access to data, and examples that replicate all studies here presented and illustrate how to perform new analyses.

Results

Detection of artificially seeded borrowings

To simulate a situation in which foreign words have recently entered a language without being modified by loanword nativization processes, we designed an experiment in which the wordlists in our base datasets were artificially mixed with words from another wordlist which was not part of the original WOLD collection. The idea to use

"artificially seeded" borrowings instead of borrowings attested in actual language was originally proposed for evaluating methods for lateral gene transfer detection in biology [49], and later tested on linguistic data in order to assess the power of phylogenetic methods for borrowing detection across multiple languages [22]. The advantage of this procedure is that it creates simulated data without requiring the efforts of detailed simulation experiments.

Artificial borrowings were seeded into a wordlist in three steps. We first removed all borrowed words from the wordlist to guarantee that no recent borrowings from other languages could influence the results. We then added inherited words from the additional German list, which we created for testing purposes. Here, we tested three different proportions of borrowed words, 5%, 10%, and 20%, in order to allow to compare different degrees of contact. In a final step, we then split the resulting wordlist into a training and a test set (reserving 80% of the data for training and 20% for testing) and ran the three methods for monolingual borrowing detection, Bag of Sounds, Markov Model, and Neural Network.

The results of this experiment are given in Table 2, where the borrowing detection results are provided in form of *precision*, *recall*, and *F1 scores* for the three different borrowing rates. Fig 2 presents a plot for 5% and 10% borrowing rates. Accuracy results, not shown, were all above 0.95 and varied little over methods and rates. Individual results indicating the scores achieved by method and borrowing rate for each language are provided as supporting information in S1 Seeded borrowings.

Method	**Rate%**	**Prec.**	**Recall**	**F1**
Bag of Sounds	5	0.80	1.00	0.88
Markov Model	5	0.96	0.67	0.76
Neural Network	5	0.97	0.84	0.90
Bag of Sounds	10	0.87	0.99	0.92
Markov Model	10	0.96	0.87	0.91
Neural Network	10	0.97	0.93	0.95
Bag of Sounds	20	0.91	0.99	0.94
Markov Model	20	0.97	0.94	0.95
Neural Network	20	0.99	0.97	0.98

Table 2. Borrowing detection results for artificially seeded borrowings, averaged from all datasets for all three methods and three different borrowing rates.

As can be seen from the results, all methods perform well when artificially seeded borrowings amount to 20%. With a borrowing rate of 10%, all methods still achieve F1

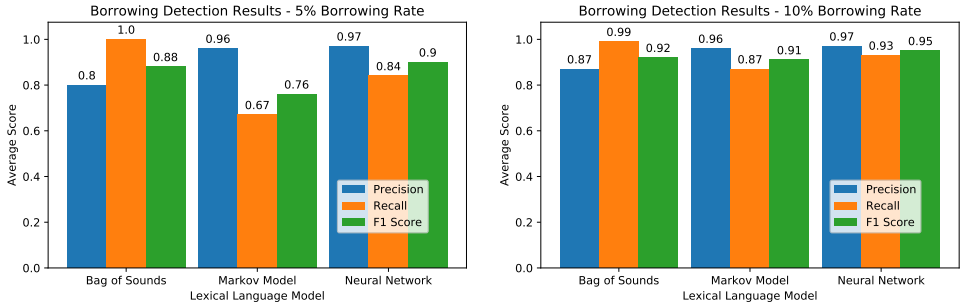


Fig 2. Borrowing detection results for borrowing rates of 5% (left) and 10% (right) in the experiment on artificially seeded borrowings.

scores of more than 0.90, with the Bag of Sounds showing the lowest precision and the Markov Model showing the lowest recall. When borrowings only amount to 5%, we can observe the same trend of low precision for the Bag of Sounds and low recall for the Markov Model. However, while the Bag of Sounds still comes close to the performance of the Neural Network with respect to the F1 score (0.88 vs. 0.90), the Markov Model shows a drastic drop here, resulting from the dramatic loss in recall (0.67).

Cross validation of borrowing detection on real language data

Our experiment on artificially seeded borrowings was simulating an ideal situation of language contact in which new words were recently introduced into a given language without being adjusted to the recipient language's target phonology. While this experiment provided high scores in our evaluation experiment, the experiment does not allow us to estimate how well the three borrowing detection methods will perform when being exposed to "real" data. For this reason, we designed a second experiment on the WOLD data in their original form. Given that the wordlists are quite small, while specifically Markov Model and Neural Network language models tend to require larger amounts of data, we used cross validation techniques, in which the data are repeatedly partitioned into training and test data and evaluation results are measured for each trial and later summarized. We employed *ten-fold cross validation* for this experiment, where each word list was partitioned into 10 parts, and over 10 successive trials, one part was successively designated the test set while the remaining nine parts were designated the training set. This resulted in 10 separate estimates of borrowing detection performance, with each word appearing once in test sets and nine times in training sets.

Table 3 shows the averages and standard deviations of results (*precision*, *recall*, *F1 score*, *accuracy*) of this experiment for each of our three methods. Fig 3 summarizes the averaged results. Individual results indicating the scores achieved by method for each language are provided as supporting information in S2 Cross validation.

Method	Statistic	Prec.	Recall	F1	Acc.
Bag of Sounds	Mean	0.286	0.578	0.349	0.843
	Language SD	0.250	0.287	0.268	0.081
	Pooled SD	0.078	0.226	0.088	0.030
Markov Model	Mean	0.678	0.521	0.578	0.828
	Language SD	0.136	0.181	0.170	0.060
	Pooled SD	0.114	0.088	0.082	0.034
Neural Network	Mean	0.697	0.546	0.603	0.844
	Language SD	0.164	0.191	0.181	0.062
	Pooled SD	0.100	0.082	0.072	0.030

Table 3. Results of the cross validation experiment, for each method over all languages.

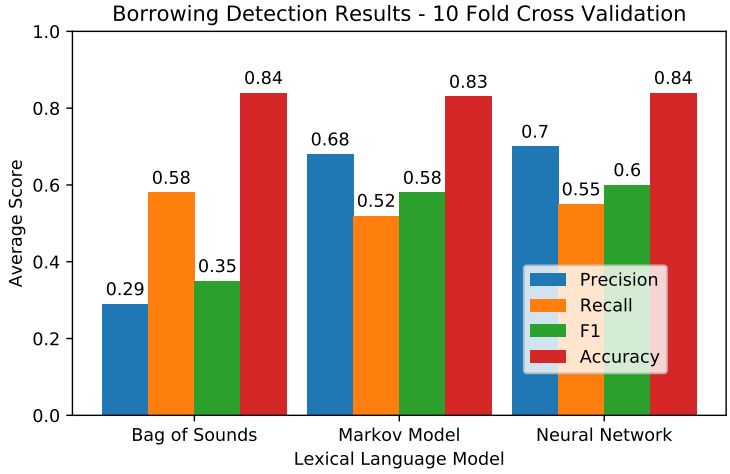


Fig 3. Results of the cross validation experiment, averaged for each model over all languages in our sample.

As can be seen from the table and the figure, the Neural Network marginally outperforms the Markov Model, while both the Neural Network and the Markov Model clearly outperform the Bag of Sounds. The strength of the entropy-based methods lies in their high precision, while the Bag of Sounds shows the highest recall, but an extremely low precision.

When examining the individual results achieved by each method for each individual language in our sample, one can find a rather huge variation in the results, ranging from results which one may consider as satisfying (such as the performance of the Neural

Network on Zinacantán Tzotzil with an F1 score of 0.81) up to extremely bad results (such as the performance of all methods on Mandarin Chinese, with F1 scores below 0.02). The reasons for the underwhelming results on Mandarin Chinese are twofold. On the one hand, the language barely borrows words directly, but rather resorts to *loan translation*, by which new concepts are rendered with the help of the lexical material in the target language. As a result, Mandarin has the lowest amount of direct borrowings in our sample. On the other hand, Mandarin Chinese (as well as all Chinese dialects and many languages from Southeast Asia) has an extremely restricted syllable structure that makes it impossible to render most foreign words truthfully [50]. As a result, words are usually directly adjusted to Chinese phonotactics when being borrowed and also written with existing Chinese characters, which again further masks their foreign origin [51]. However, this very specific situation also makes it also difficult if not impossible for most Mandarin Chinese speakers to identify borrowings when considering phonotactic criteria alone.

Factors that influence borrowing detection performance

Given that the performance of our supervised borrowing detection methods varied substantially, ranging from poor performance with F1 scores below 0.5, average performance with F1 scores between 0.5 and 0.8, and acceptable performance with F1 scores above 0.8, we ran two tests to assess to which degree certain factors might influence the borrowing detection methods.

In concrete, we computed specific characteristics of each language variety in our sample and then checked to which degree these characteristics correlated with the test performance. As characteristics, we chose the proportion of borrowed words in a given language wordlist, the proportion of unique sounds inside borrowed words, and the proportion of unique sounds in inherited words. Statistical analysis, correlational study, matrix plots, and regression, were performed with Minitab® Statistical Software [52]. The correlation results, based on all wordlists in our sample taken from the WOLD database, are reported in Table 4, and accompanied by detailed plots in Figs 4, 5, and 6.

As can be seen from the correlations and the plots, there is a positive correlation between the proportion of borrowed words and the evaluation scores for all tests. The

	Bag of Sounds			Markov Model			Neural Network		
Proportion of	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**	**Prec.**	**Recall**	**F1**
Borrowed words	0.584	0.337	0.539	0.387	0.736	0.654	0.399	0.690	0.600
Borrowed sounds	0.185	0.345	0.199	0.345	0.274	0.297	0.377	0.268	0.301
Inherited sounds	-0.006	-0.010	-0.004	0.035	-0.330	-0.263	-0.075	-0.178	-0.148

Table 4. Correlations between phonological factors and performance of borrowing detection methods.

effect of proportion of borrowed words appears non-linear for the entropy methods, where less than 5% borrowings has much worse borrowing detection than expected in the linear correlation plot from Figs 5, and 6. For the other factors, the proportion of sounds occurring exclusively in borrowed words, and the proportion of sounds occurring exclusively in inherited words, the results are less clear. While we observe a moderate correlation between the proportion of exclusively borrowed sounds with the recall for the Bag of Sounds, there is a higher correlation with the precision for the other two methods.

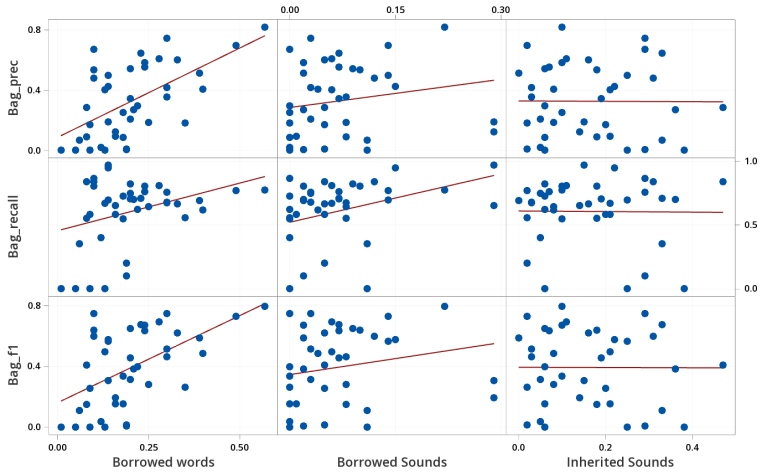


Fig 4. Determining factors that influence the performance of the Bag of Sounds.

In order to further investigate the influence of the three factors on the borrowing detection performance, we further analyzed them by fitting a multiple regression model to them. Our major goal was to check whether exclusively borrowed and exclusively inherited sound proportions can help us explain the methods' performance beyond the overall proportion of borrowed words in each wordlist. By fitting a full second order regression model to predict F1 scores from our three factors, using Minitab's forward information criteria for model selection, we found that all three phonological variables contribute to explain the F1 scores for the borrowing detection performance for the

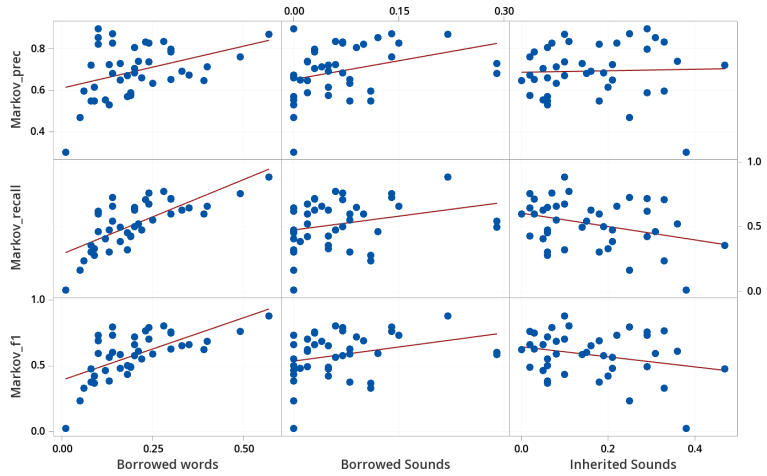


Fig 5. Determining factors that influence the performance of the Markov Model.

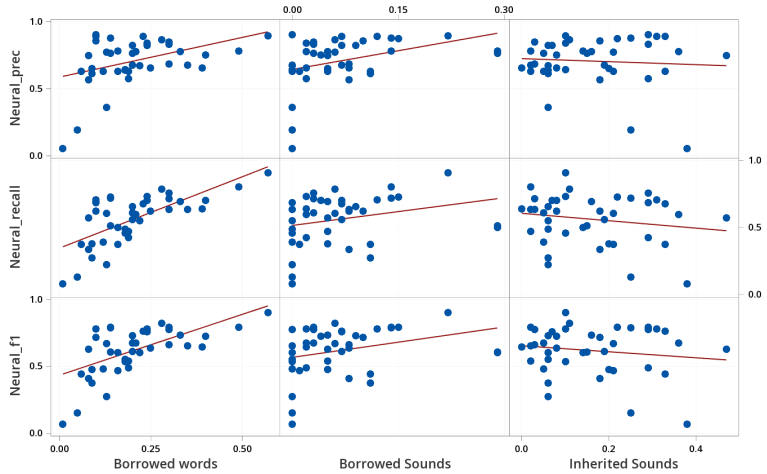


Fig 6. Determining factors that influence the performance of the Neural Network.

Markov Model and the Neural Network, while only the proportion of borrowed words seems to be the dominant factor for the Bag of Sounds.

Method	Regression model	R^2
Bag of Sounds	$F1 = -0.040 + 1.53bw + 0.76ns$	29.9%
Markov Model	$F1 = 0.141 + 2.66bw + 2.05bs - 3.38bw^2 - 5.05bs^2$	48.8%
Neural Network	$F1 = 0.032 + 3.12bw + 2.43bs + 0.43ns - 3.93bw^2 - 6.35bs^2$	49.9%

Table 5. Regression analysis of factors that influence borrowing detection performance as reflected in F1 scores.

Borrowings from a single donor in intensive contact situations

Testing our lexical language models on the WOLD data in their entirety could be considered as unfair to the methods, given that we know well that monolingual evidence

for borrowing in phonotactics may get lost easily and that the WOLD database was never restricted to recent borrowings alone. Another problem of the data is that the distinction between inherited words on the one hand and borrowings on the other hand is as well a simplifying assumption, since we know that in intensive contact situations borrowings come from a specific donor language. As a result, it seems to be justified to test the three methods for monolingual borrowing detection with the help of more specific experiments in which the task consists in the detection of borrowings when there is a single or dominant language donor, i.e., intensive contact, versus the case when no language donor dominates.

To test whether our methods show an improved performance when there is a dominant language donor as opposed to detecting borrowed words *per se*, we first created two subsets of the WOLD database, one containing languages with 300 and more borrowed words (17 language varieties), and one containing languages with 100 and more borrowed words (37 language varieties). We then searched for "dominant donor languages" in all wordlists in each sample, with dominant donor languages being defined as those donor languages (as identified in the WOLD database) that would account for two-thirds of all borrowings identified for a given language variety. For our sample of language varieties with 300 and more borrowings, this yielded a partition of the data into 8 language varieties for which a dominant donor could be identified and 9 for which none could be found. For the sample of language varieties with 100 and more borrowings, the partition yielded 20 language varieties with a dominant donor and 17 without. We were able to apply results of the 10-fold cross validation study for these two subsets of the data, which we had previously applied to all language varieties in the WOLD database. In order to test whether the observed differences between dominant donor and no dominant donor categories were significantly different, we also performed randomization resampling tests of 5,000 iterations each, using Student's independent t statistic with unequal variances as our test statistic. We report p-values from the empirical distribution of t statistics calculated under the hypothesis of no difference due to dominant donor, i.e., dominant and no dominant categories are exchangeable.

As can be seen from the results in Table 6, the performance of all borrowing detection methods improves when the vast majority of the borrowings come from a single donor language. The performance also improves, as we saw previously, with more

borrowed words. While performing worse than the other two methods, the Bag of Sounds method shows a strong increase in performance, which is mostly owed to a strong increase in precision, when most borrowings come from a single donor language.

Borrowed	Method	Donor	Precision	p<	Recall	p<	F1	p<
≧ 300	Bag of Sounds	Dominant (8)	0.536	.0300	0.739	.0200	0.588	.0400
		No dominant (9)	0.308		0.672		0.390	
	Markov Model	Dominant	0.785	.0030	0.722	.0020	0.749	.0030
		No dominant	0.672		0.585		0.622	
	Neural Network	Dominant	0.810	.0002	0.722	.0070	0.760	.0030
		No dominant	0.690		0.606		0.642	
≧ 100	Bag of Sounds	Dominant (20)	0.418	.0030	0.737	.0020	0.490	.0010
		No dominant (17)	0.192		0.498		0.252	
	Markov Model	Dominant	0.762	.0002	0.600	.0300	0.661	.0060
		No dominant	0.639		0.505		0.558	
	Neural Network	Dominant	0.787	.0002	0.619	.0200	0.685	.0060
		No dominant	0.655		0.523		0.567	

Table 6. 10-fold cross validation - dominant versus no dominant donor.

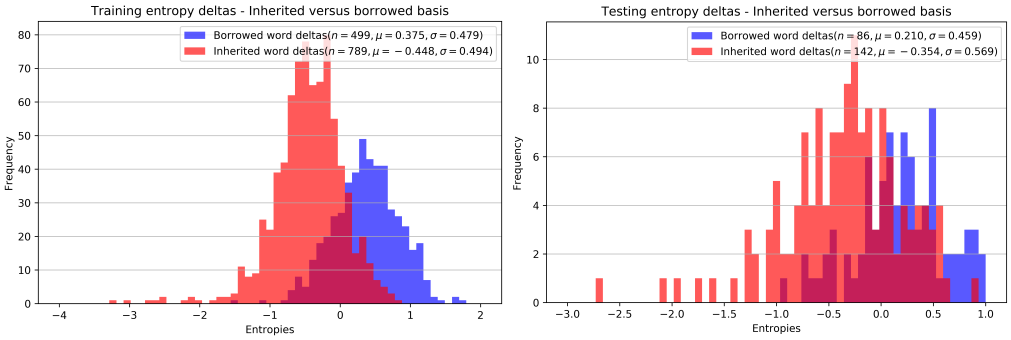
Comparing entropy distributions to investigate the performance of the Markov Model and Neural Network methods

The Markov Model and the Neural Network methods estimate word entropy on a per sound basis given the inherited or borrowed words on which they are trained. Models trained on inherited words should estimate lower entropies for inherited words, and models trained on borrowed words should estimate lower entropies for borrowed words. However, since words are borrowed over time and potentially also from various donor languages, using a single language model for borrowed words is not always optimal.

Our decision procedure for the Markov Model and the Neural Network methods requires the comparison of competing entropies for a given word, the entropy of the lexical language model derived from inherited words and the entropy of the lexical language model derived from borrowed words. If the difference between the entropies is greater than zero, we designate the word as borrowed, and if it is smaller than or equal to zero, we designate the word as inherited.

In order to investigate the *discriminative force* of this procedure, it is useful to compare entropy difference distributions of inherited and borrowed words for a given language variety. The distributions for training and test data from the English wordlist in the WOLD database are shown in Fig 7. While there is a certain overlap between entropy difference distributions for inherited and borrowed words, the problem of

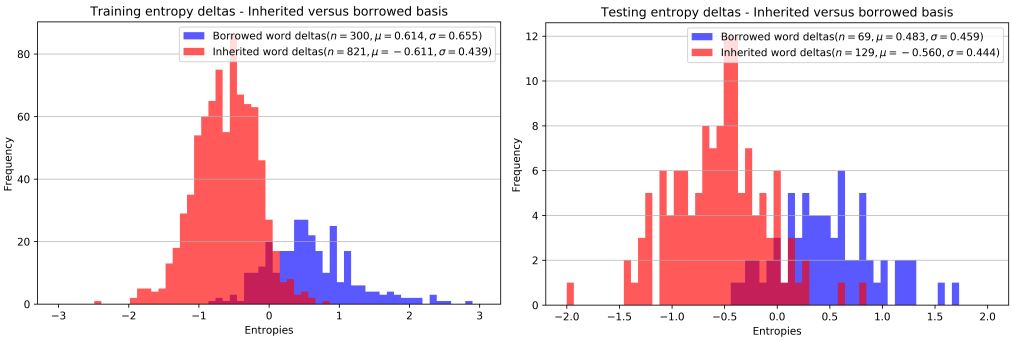
discriminating between them based on entropy differences seems tractable, and we can assume that improvements in entropy estimation would have an immediate benefit on prediction.



(a) Training entropy deltas **(b)** Testing entropy deltas

Fig 7. Distribution of entropy differences – training (85%) and testing (15%) data for English – Neural Network method.

Since both the Markov Model and the Neural Network performed considerably well on Imbabura Quechua, a Quechua language spoken in parts of Ecuador, Columbia, and Northern Peru, with an F1 score above 0.8, it is not surprising that we find a good separation between the entropy difference distributions for inherited and borrowed words, as shown in Fig 8.

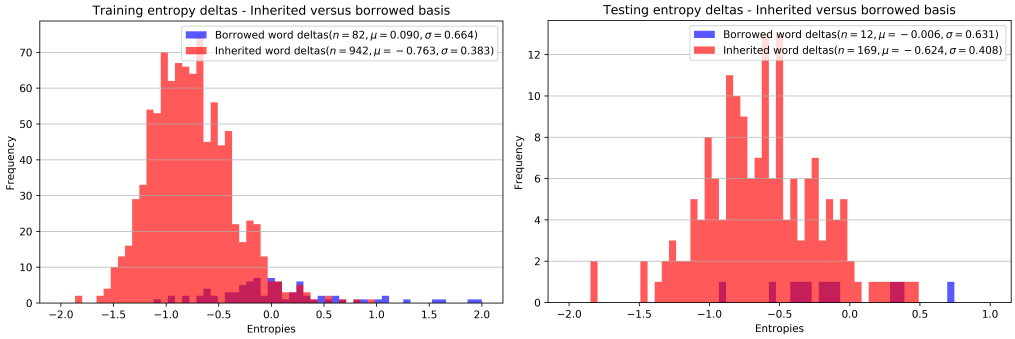


(a) Training entropy deltas **(b)** Testing entropy deltas

Fig 8. Distribution of entropy differences – training (85%) and testing (15%) data for Imbabura Quechua – Neural Network method.

Neither method performed very well on Oroqen, a Northern Tungusic language spoken in the Mongolian region of the People's Republic of China, with F1 scores below 0.36. Consequently, as can be seen in Fig 9 the entropy difference distributions for inherited and borrowed words are not well separated.

This strong relationship between the distribution of entropy differences and



(a) Training entropy deltas **(b)** Testing entropy deltas

Fig 9. Distribution of entropy differences – training (85%) and testing (15%) data for Oroqen – Neural Network method.

borrowing detection, indicates a tactic for improving monolingual lexical borrowing detection – increase the separation of difference distributions for inherited versus borrowed words. An examination of our sample cases reveals: 1. English and Imbabura Quechua, even though there were substantial borrowings, have reduced separation between inherited and borrowed word difference distributions for testing, resulting in reduced discriminative power, and 2. Oroqen, with few borrowings, has almost no separation between inherited and borrowed word distributions for testing, resulting in little discriminative power. Identification of problems permits trying to solve them, such as through improved controls for training of Neural Networks, and by obtaining more borrowings, real or simulated, for training.

Discussion

Artificially seeded borrowings

In our experiment on artificially seeded borrowings, we used a very straightforward approach to simulate data that would reflect a situation of very close and intensive language contact during which a larger amount of words are being transferred without being further altered in their phonology or phonotactics. While all methods performed well when the proportion of artificially borrowed words was high, they developed specific problems when the proportion of borrowings was decreased.

While the Bag of Sounds outperformed the other two methods regarding recall, the Markov Model and the Neural Network outperformed the Bag of Sounds method in

precision. Since the core strategy of the Bag of Sounds lexical language model is to identify borrowed words by their specific sounds, while the order of sounds itself is ignored, it is not surprising that the method performs better in identifying artificially seeded borrowings, i.e., better recall, since the direct transfer of words from one wordlist to another wordlist, as it was done in our experiment, will always introduce a larger number of sounds which were not present in the recipient wordlist prior to the transfer.

Cross validation of borrowing detection methods

In our 10-fold cross validation experiment, which was carried out on the full wordlist data as provided by the WOLD database, we tried to check to what degree the methods would be able to detect borrowings in a more realistic setting.

Here, the Neural Network performed marginally better than the Markov Model. A major factor favoring the Neural Network seems to be that it includes conditional dependencies from all previous sound segments, without having to explicitly estimate numerous extra parameters for this dependency.

Both the Markov Model and the Neural Network methods performed much better than Bag of Sounds. Similar to the previous experiment, the Bag of Sounds method showed a high recall, but suffered from a low precision as well. So while the Bag of Sounds *suspects* considerably many words of being borrowings, it does not necessarily always pick the right ones and shows a rather high rate of false positives (as can be seen from the low rates of precision). In contrast, the Markov Model and the Neural Network methods show a lower recall, but also a much higher precision. They are therefore much more *conservative* than the Bag of Sounds method. When the overall proportion of borrowed words in wordlists is small, all models perform poorly. This is not necessarily surprising, since low borrowing proportions make it difficult to learn the phonotactics or phonology of borrowed words (if these can be identified after all), and it is also not clear to which degree native speakers would be able to identify borrowed words in the respective languages.

Factors determining borrowing detection performance

Given the disappointing results of our cross validation study, we tried to determine the major factors that might influence the performance of the monolingual borrowing detection methods. Besides selecting the proportion of borrowings as one potentially important factor, we also chose the number of sounds uniquely attested in borrowed words and the number of sounds uniquely attested in inherited words as potential factors. Our assumption was that the latter two factors should have some effect on the performance of the Bag of Sounds, given that this method explicitly deals with sounds, while ignoring all phonotactic aspects. While the effect of the proportion of borrowed words was remarkable, showing a strong linear increase in performance for all methods when the proportion of borrowed words was 5% and more, the impact of the proportions of sounds occurring exclusively in borrowed words and sounds occurring exclusively in inherited words was much lower than we would have expected, especially for the Bag of Sounds method. However, what we may have overestimated was that – even if a given language has many sounds occurring exclusively in borrowed words – this does not mean that these sounds need to occur in each and every borrowed word. Thus, while the presence of specific sounds may be a powerful indicator of a borrowing or an inherited word, this evidence may be too sparse in comparison with the full lexicon of a given language.

Detecting borrowings from a single donor language

Since we create lexical language models for borrowed and inherited words, it is straightforward to question why our basic approach would treat all borrowed words as if they represented a single donor language. While it may hold for specific contact situations that a given language is heavily influence by one single, dominant donor language, it is also possible that borrowings form distinct layers in the lexicon of a given language, reflecting borrowings from different donor languages and different times. If the majority of the borrowings attested in a given language stem from a single donor, however, we would assume that our lexical language model approaches to monolingual borrowing detection would perform better, since the donor language which we access through the recipient language would provide a much more coherent picture than would

a mix of words from different donor languages.

We therefore systematically tested whether the performance of our methods would increase for those wordlists in our sample for which a dominant donor language could be identified. Our assumption, that the methods should show an increased performance for languages with a dominant donor language were largely confirmed, as reflected in substantially increased F1 scores of ≈ 0.75 for the Markov Model and the Neural Network methods in cases of high contact with more than 300 borrowings. While we still consider the overall performance of the monolingual borrowing detection disappointing, this experiment reflects the importance of having a consistent sample of the donor language when dealing with monolingual borrowing detection.

Comparing entropy distributions

Our final evaluation was intended to demonstrate how the Markov Model and Neural Network methods discriminate between inherited and borrowed words. We showed how plots of the distribution of entropy differences between competing inherited and borrowed word models served to explain borrowing detection results. When comparing the distributions of entropy differences, we found that in those cases where the proportion of borrowings was small, the discriminative force of the word entropy differences seemed to drop drastically for testing. Even when borrowings for training seemed adequate we saw a reduction in discriminative force for testing due to reduced separation of inherited and borrowed word entropy difference distributions. This provided additional evidence that monolingual borrowing detection heavily depends on the presence of a large enough proportion of borrowed words, and also that modest improvements might be possible with improved training controls.

Conclusion

We presented three supervised methods for the detection of borrowings in monolingual wordlists. These methods are based on lexical language models which are intended to model specific aspects of phonology and phonotactics in the lexicon of spoken languages. Assuming that phonological and phonotactic properties of words in the lexicon of a spoken language can provide enough clues to identify borrowings by language-internal

comparison of words alone, we designed workflows in which the lexical language models could be trained with monolingual wordlists in which borrowings are annotated and then used to detect borrowings when being confronted with so far unobserved words.

While tests on artificially seeded borrowings showed promising results, tests on real wordlists taken from the WOLD database revealed a rather disappointing performance for all three methods. Consecutive attempts to identify the potential reasons for this mediocre performance revealed two main factors that considerably influence how well the methods performed, namely (1) the amount of borrowings in a given language variety, (2) the uniformity of the borrowings in a given language variety (as reflected in the presence of a dominant donor language). While first factor reflects the importance of having enough training data when working in supervised learning frameworks, the second factor reflects the very specific linguistic conditions of monolingual borrowing detection. Assuming that speakers who can identify borrowings in their native language make use of primarily phonological and phonotactic clues, it seems that the salient factor lies not only in the properties of the inherited words, but also in the specific properties of the borrowed words, which can be much better identified when they come from a uniform sample.

While our results do not recommend any of the three methods represented here as a replacement for previously proposed methods for borrowing detection, we believe that the methods we created offer a valuable and promising base for further exploration, and we are even convinced that they may be useful in some current applications. The recurrent Neural Network method offers more promise, both for its marginally better detection performance than the Markov Model, and for its opportunities for improvement via better control in training or leading edge algorithms. Given that we know that our methods rely heavily on a sufficiently large sample of training data, our methods may be useful for those studies in which borrowed words or sentences need to be identified in large amounts of data, preferably in situations where borrowings are considerably young. Here, especially, larger linguistic corpora could be analyzed and tagged for inherited and borrowed words. Our methods might also be attractive for scholars working on code switching, where multilingual language users switch between different varieties based on sociolinguistic contexts.

Additionally, we think that – given that by now no single method for borrowing

detection has been proposed that exhibits satisfactory performance – our methods add to the growing pool of automated approaches to borrowing detection which could ideally be later combined into an integrated workflow in which evidence from multiple sources can be combined to form a unified picture of language contact.

Last not least, we also emphasize that it is very well possible to further improve our methods: 1. Our comparison of distributions of entropy differences suggests improved control of Neural Network training is possible. 2. Improved detection for dominant donors suggests that using multiple donor models instead of just one borrowing model might offer better detection results. While improving our current methods will result in better borrowing detection, there is much more to this problem than individual monolingual wordlists. Minimally what is needed are more integrated approaches that also take into account multilingual and cross-linguistic information for a proper automated borrowing detection. We hope that the software library which implements all three approaches and which we supplement with this study will make it easy for us and our colleagues to build and improve upon, and use to further explore borrowed word detection.

Acknowledgments

The author, JEM, has received funding and encouragement from the Graduate School of the Pontificia Universidad Católica del Perú (PUCP) thorugh the Huiracocha scholarship program. The authors, TT, JML, have received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. ERC Grant #715618, "Computer-Assisted Language Comparison"). We thank Mei-Shin Wu for work on the White Hmong and Mandarin profiles used to convert WOLD word forms to IPA sound segments.

Author Contributions

- Conceptualization: JEM TT CB RZ JML.
- Data curation: TT NM JML.
- Formal analysis: JEM TT JML.
- Funding acquisition: CB JML.
- Investigation: JEM TT JML.
- Methodology: JEM TT JML.
- Project administration: CB RZ JML.
- Software: JEM TT JML.
- Validation: JEM TT JML.
- Visualization: JEM.
- Writing – original draft: JEM TT JML.
- Writing – review & editing: JEM TT CB RZ JML.

References

1. Grant AP. Lexical Borrowing. In: Taylor JR, editor. The Oxford Handbook of the Word. Oxford: Oxford University Press; 2014.

2. Campbell L. Historical Linguistics: An Introduction. Edinburgh University Press; 2013.

3. Geisler H, List JM. Do languages grow on trees? The tree metaphor in the history of linguistics. In: Fangerau H, Geisler H, Halling T, Martin W, editors. Classification and evolution in biology, linguistics and the history of science. Concepts – methods – visualization. Stuttgart: Franz Steiner Verlag; 2013. p. 111–124.

4. Plato. Plato in Twelve Volumes. vol. 12. Cambridge, MA/ London: Harvard University Press/William Heinemann Ltd.; 1921.

5. List JM. Automatic detection of borrowing (Open problems in computational diversity linguistics 2); 2019. Web blog at: http://phylonetworks.blogspot.com/2019/03/automatic-detection-of-borrowing-open.html.

6. Gray RD, Greenhill SJ, Atkinson QD, et al. Phylogenetic models of language change: three new questions. Cultural Evolution: Society, Technology, Language, and Religion. 2013;.

7. Heien LG. Loanword recognition in Russian. Russian Language Journal. 1984;38(131):51–61.

8. Cho I. Recognition of English loanwords by learners of Korean. The Korean Language in America. 2001;6:69–74.

9. List JM. Automated methods for the investigation of language contact situations, with a focus on lexical borrowing. Language and Linguistics Compass. 2019;13(e12355):1–16.

10. Chen Q. Miooyao yuwen [Miao and Yao language]. Beijing: Central Institute of Minorities; 2012. Available from: `https://en.wiktionary.org/wiki/Appendix:Hmong-Mien_comparative_vocabulary_list`.

11. Maddieson I. Borrowed sounds. In: Ferguson CA, Fishman JA, editors. The Fergusonian Impact: In Honor of Charles A. Ferguson on the Occasion of His 65th Birthday. v. 1. Berlin: Mouton de Gruyter; 1986.

12. Grossman E, Eisen E, Nikolaev D, Moran S. SegBo: A Database of Borrowed Sounds in the World's Languages. In: Proceedings of The 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association; 2020. p. 5316–5322. Available from: https://www.aclweb.org/anthology/2020.lrec-1.654.

13. Kiparsky P. New perspectives in historical linguistics. In: Bowern C, Evans B, editors. The Routledge Handbook of Historical Linguistics. Routledge; 2014. p. 64–102.

14. Calabrese A, Wetzels WL. Loan phonology: Issues and controversies. In: Calabrese A, Wetzels WL, editors. Loan Phonology. vol. 307 of Current Issues in Linguistic Theory. John Benjamins; 2009. p. 1–10.

15. Bender E, Koller A. Climbing towards NLU: On meaning, form, and understanding in the age of data. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsberg: Association for Computational Linguistics; 2020. p. 5185–5198.

16. Haspelmath M, Tadmor U, editors. World Loanword Database (WOLD). Leipzig: Max Planck Institute for Evolutionary Anthropology; 2009.

17. Minett JW, Wang WSY. On detecting borrowing. Diachronica. 2003;20(2):289–330.

18. Nakhleh L, Ringe D, Warnow T. Perfect Phylogenetic Networks: A new methodology for reconstructing the evolutionary history of natural languages. Language. 2005;81(2):382–420.

19. Nelson-Sathi S, List JM, Geisler H, Fangerau H, Gray RD, Martin W, et al. Networks uncover hidden lexical borrowing in Indo-European language evolution. Proceedings of the Royal Society of London B: Biological Sciences. 2011;278(1713):1794–1803. doi:https://doi.org/10.1098/rspb.2010.1917.

20. List JM, Nelson-Sathi S, Geisler H, Martin W. Networks of lexical borrowing and lateral gene transfer in language and genome evolution. Bioessays. 2014;36(2):141–150.

21. List JM, Nelson-Sathi S, Martin W, Geisler H. Using phylogenetic networks to model Chinese dialect history. Language Dynamics and Change. 2014;4(2):222–252.

22. List JM. Network Perspectives on Chinese Dialect History: Chances and Challenges. Bulletin of Chinese Linguistics. 2015;8:27–47.

23. Willems M, Lord E, Laforest L, Labelle G, Lapointe FJ, Sciullo AMD, et al. Using hybridization networks to retrace the evolution of Indo-European languages. BMC Evolutionary Biology. 2016;16(1):180.

24. van der Ark R, Mennecier P, Nerbonne J, Manni F. Preliminary identification of language groups and loan words in Central Asia. In: Proceedings of the RANLP Workshop on Acquisition and Management of Multilingual Lexicons; 2007. p. 13–20.

25. Mennecier P, Nerbonne J, Heyer E, Manni F. A Central Asian language survey. Language Dynamics and Change. 2016;6(1):57–98. doi:10.1163/22105832-00601015.

26. Zhang L, Manni F, Fabri R, Nerbonne J. Detecting loan words computationally; 2019. Draft, submitted to the Contact Language Libraries series.

27. McMahon A, Heggarty P, McMahon R, Slaska N. Swadesh sublists and the benefits of borrowing: An Andean case study. Transactions of the Philological Society. 2005;103:147–170.

28. Mi C, Yang Y, Zhou X, Wang L, Li X, Jiang T. Recurrent Neural Network Based Loanwords Identification in Uyghur. In: Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers. Seoul, South Korea; 2016. p. 209–217.

29. Mi C, Yang Y, Wang L, Zhou X, Jiang T. Toward Better Loanword Identification in Uyghur Using Cross-lingual Word Embeddings. In: Proceedings of CoLing 2018. Seoul, South Korea; 2018. p. 3027–3037.

30. Jauhiainen T, Lui M, Zampieri M, Baldwin T, Lindén K. Automatic Language Identification in Texts: A Survey. J Artif Int Res. 2019;65(1):675–682.

31. List JM, Rzymski C, Greenhill S, Schweikhard N, Pianykh K, Tjuka A, et al. Concepticon. A resource for the linking of concept lists (Version 2.3.0). Jena: Max Planck Institute for the Science of Human History; 2020. Available from: https://concepticon.clld.org/.

32. Anderson C, Tresoldi T, Chacon TC, Fehn AM, Walworth M, Forkel R, et al. A Cross-Linguistic Database of Phonetic Transcription Systems. Yearbook of the Poznań Linguistic Meeting. 2018;4(1):21–53. doi:https://doi.org/10.2478/yplm-2018-0002.

33. List JM, Anderson C, Tresoldi T, Rzymski C, Greenhill S, Forkel R. Cross-Linguistic Transcription Systems. Version 1.3.0. Jena: Max Planck Institute for the Science of Human History; 2019.

34. Moran S, Cysouw M. The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles. Berlin: Language Science Press; 2018. Available from: http://langsci-press.org/catalog/book/176.

35. Wu MS, Schweikhard NE, Bodt TA, Hill NW, List JM. Computer-Assisted Language Comparison. State of the Art. Journal of Open Humanities Data. 2020;6(2):1–14.

36. Forkel R, List JM. Proceedings of the Twelfth International Conference on Language Resources and Evaluation. In: Proceedings of the Twelfth International Conference on Language Resources and Evaluation. Luxembourg: European Language Resources Association (ELRA); 2020-05-11/2020-05-16). p. 6997-7004. Available from: http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.864.pdf.

37. Forkel R, List JM, Greenhill SJ, Rzymski C, Bank S, Cysouw M, et al. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. Scientific Data. 2018;5(180205):1–10.

38. Rzymski C, Tresoldi T, Greenhill S, Wu MS, Schweikhard NE, Koptjevskaja-Tamm M, et al. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross- linguistic polysemies. Scientific Data. 2020;7(13):1–12.

39. Kluge F, editor. Etymologisches Wörterbuch der deutschen Sprache. 24th ed. Berlin: de Gruyter; 2002.

40. Baayen RH, Piepenbrock R, Gulikers L, editors. The CELEX Lexical Database. Philadelphia: Linguistic Data Consortium; 1995.

41. Tresoldi T, Forkel R, Morozova N. CLDF dataset derived from Haspelmath and Tadmor's "World Loanword Database" from 2009. Geneva: Zenodo; 2019. Available from: `https://doi.org/10.5281/zenodo.3537579`.

42. Jurafsky D, Martin JH. Speech and Language Processing (2nd Edition). Upper Saddle River, NJ, USA: Prentice-Hall, Inc.; 2009.

43. Bengio Y, Ducharme R, Vincent P, Janvin C. A Neural Probabilistic Language Model. J Mach Learn Res. 2003;3:1137–1155.

44. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.; 2001.

45. Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. 1st ed. Cambridge University Press; 2000.

46. Steven Bird EK, Loper E. Natural Language Processing with Python. O'Reilly Media; 2019.

47. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al.. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems; 2015.

48. Manning CD, Schütze H. Foundations of Statistical Natural Langage Processing. Cambridge, MA, USA: MIT Press; 2001.

49. Dessimoz C, Margadant D, Gonnet GH. DLIGHT – Lateral gene transfer detection using pairwise evolutionary distances in a statistical framework. In: Vingron M, Won L, editors. Research in Computational Molecular Biology. Berlin and Heidelberg: Springer; 2008. p. 315–330.

50. Norman J. Chinese. Cambridge: Cambridge University Press; 1988.

51. Sun C. Chinese: A linguistic introduction. Cambridge: Cambridge University Press; 2006.

52. Minitab L. Minitab® Statistical Software, version 19.; 2020. Available from minitab.com.

Supporting information

S1 Seeded borrowings Detection results by language for seeded borrowings

	Recurrent neural net				Markov model				Bag of sounds			
Language	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**
Archi	0.89	0.53	0.67	0.96	1.00	0.39	0.56	0.95	0.71	1.00	0.83	0.98
Bezhta	1.00	0.80	0.89	0.99	1.00	0.41	0.58	0.93	0.75	1.00	0.86	0.99
Ceq Wong	1.00	1.00	1.00	1.00	1.00	0.23	0.38	0.76	0.83	1.00	0.91	0.99
Dutch	0.71	0.71	0.71	0.97	0.94	0.60	0.73	0.96	0.67	1.00	0.80	0.98
English	0.86	0.75	0.80	0.98	0.89	0.24	0.37	0.86	0.80	1.00	0.89	0.99
Gawwada	0.93	0.82	0.87	0.98	1.00	0.54	0.70	0.95	0.73	1.00	0.84	0.99
Gurindji	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.93	1.00	0.96	0.99
Hausa	1.00	0.85	0.92	0.99	1.00	0.83	0.91	0.99	0.71	1.00	0.83	0.98
Hawaiian	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Hup	0.88	1.00	0.93	1.00	1.00	0.31	0.48	0.95	0.86	1.00	0.92	1.00
Imbabura Quechua	1.00	1.00	1.00	1.00	0.89	0.89	0.89	0.99	0.89	1.00	0.94	0.99
Indonesian	1.00	0.95	0.97	1.00	0.92	0.92	0.92	0.99	0.92	1.00	0.96	1.00
Iraqw	1.00	0.83	0.91	0.99	0.93	0.81	0.87	0.98	0.80	1.00	0.89	0.99
Japanese	1.00	1.00	1.00	1.00	1.00	0.81	0.89	0.99	1.00	1.00	1.00	1.00
Kali'na	1.00	1.00	1.00	1.00	0.94	1.00	0.97	1.00	0.93	1.00	0.96	1.00
Kanuri	1.00	0.93	0.96	1.00	0.88	0.70	0.78	0.99	0.76	1.00	0.87	0.99
Ket	1.00	0.73	0.85	0.98	0.93	0.57	0.70	0.95	0.79	1.00	0.88	0.99
Kildin Saami	1.00	0.81	0.90	0.99	1.00	0.45	0.62	0.96	0.86	1.00	0.92	0.99
Lower Sorbian	1.00	0.87	0.93	0.99	0.92	0.57	0.71	0.97	0.71	1.00	0.83	0.98
Malagasy	1.00	1.00	1.00	1.00	1.00	0.88	0.93	0.99	1.00	1.00	1.00	1.00
Manange	1.00	0.79	0.88	0.99	0.88	0.41	0.56	0.90	0.82	1.00	0.90	0.99
Mandarin Chinese	1.00	0.94	0.97	1.00	1.00	0.91	0.95	1.00	0.74	1.00	0.85	0.99
Mapudungun	1.00	1.00	1.00	1.00	1.00	0.38	0.55	0.96	0.91	1.00	0.95	1.00
Old High German	0.82	0.60	0.69	0.97	1.00	0.54	0.70	0.96	0.81	1.00	0.90	0.99
Oroqen	1.00	0.47	0.64	0.96	1.00	0.48	0.65	0.94	0.41	1.00	0.58	0.96
Otomi	1.00	0.90	0.95	1.00	0.97	0.88	0.92	0.99	0.78	1.00	0.88	0.99
Q'eqchi'	0.94	0.70	0.80	0.98	0.94	0.62	0.74	0.97	0.71	1.00	0.83	0.99
Romanian	0.86	0.63	0.73	0.97	0.86	0.60	0.71	0.97	0.58	1.00	0.73	0.97
Sakha	1.00	0.83	0.91	0.98	1.00	0.68	0.81	0.98	0.71	1.00	0.83	0.98
Saramaccan	1.00	0.67	0.80	0.98	1.00	0.81	0.90	0.98	0.55	1.00	0.71	0.97
Selice Romani	1.00	1.00	1.00	1.00	0.89	0.33	0.48	0.89	0.71	1.00	0.83	0.99
Seychelles Creole	1.00	0.88	0.93	0.99	0.90	0.76	0.83	0.98	0.78	1.00	0.88	0.99
Swahili	1.00	0.93	0.97	1.00	1.00	0.90	0.95	0.99	0.91	1.00	0.95	1.00
Takia	1.00	0.80	0.89	0.99	0.82	0.47	0.60	0.94	0.91	1.00	0.95	1.00
Tarifiyt Berber	1.00	0.64	0.78	0.97	1.00	0.39	0.56	0.94	0.56	1.00	0.71	0.98
Thai	1.00	0.84	0.91	0.99	0.86	0.86	0.86	0.99	0.76	1.00	0.87	0.99
Vietnamese	1.00	0.63	0.77	0.97	1.00	0.50	0.67	0.97	0.77	1.00	0.87	0.99
White Hmong	1.00	0.85	0.92	0.99	1.00	1.00	1.00	1.00	0.95	1.00	0.97	1.00
Wichí	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.82	1.00	0.90	0.99
Yaqui	1.00	1.00	1.00	1.00	1.00	0.95	0.97	1.00	0.93	1.00	0.96	1.00
Zinacantán Tzotzil	1.00	0.92	0.96	1.00	0.91	0.77	0.83	0.98	0.91	1.00	0.95	1.00
Mean over languages	0.97	0.84	0.90	0.99	0.96	0.67	0.76	0.96	0.80	1.00	0.88	0.99

Table 7. Fake words - 5% borrowing - metrics by language.

	Recurrent neural net				Markov model				Bag of sounds			
Language	**Prec**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**
Archi	0.95	0.95	0.95	0.99	0.95	0.58	0.72	0.94	0.86	0.95	0.90	0.98
Bezhta	1.00	0.85	0.92	0.98	0.95	0.76	0.84	0.97	0.87	1.00	0.93	0.99
Ceq Wong	1.00	0.90	0.95	0.99	0.92	0.57	0.71	0.93	0.78	1.00	0.88	0.99
Dutch	0.82	0.78	0.79	0.94	0.86	0.71	0.77	0.95	0.72	0.95	0.82	0.97
English	1.00	0.83	0.91	0.98	0.88	0.70	0.78	0.94	0.75	1.00	0.86	0.99
Gawwada	0.96	0.92	0.94	0.99	0.96	0.74	0.83	0.96	0.86	1.00	0.92	0.99
Gurindji	1.00	0.96	0.98	0.99	1.00	1.00	1.00	1.00	0.89	1.00	0.94	0.99
Hausa	1.00	1.00	1.00	1.00	1.00	0.91	0.95	0.99	0.87	1.00	0.93	0.99
Hawaiian	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	1.00	0.98	1.00
Hup	1.00	0.92	0.96	0.99	0.96	0.85	0.90	0.98	0.84	1.00	0.91	0.99
Imbabura Quechua	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	1.00	0.98	1.00
Indonesian	1.00	1.00	1.00	1.00	0.95	0.95	0.95	0.99	0.88	1.00	0.94	0.99
Iraqw	1.00	0.96	0.98	1.00	0.96	0.79	0.87	0.97	0.91	1.00	0.95	0.99
Japanese	1.00	1.00	1.00	1.00	1.00	0.94	0.97	0.99	1.00	1.00	1.00	1.00
Kali'na	1.00	1.00	1.00	1.00	0.93	0.97	0.95	0.99	0.93	1.00	0.97	0.99
Kanuri	0.97	1.00	0.98	1.00	0.97	0.97	0.97	0.99	0.86	0.96	0.91	0.98
Ket	0.89	0.86	0.87	0.97	1.00	0.81	0.90	0.98	0.75	0.96	0.84	0.96
Kildin Saami	0.97	0.91	0.94	0.98	0.97	0.78	0.86	0.96	0.71	0.95	0.82	0.97
Lower Sorbian	0.93	0.96	0.95	0.99	1.00	0.97	0.99	1.00	0.88	1.00	0.94	0.99
Malagasy	1.00	1.00	1.00	1.00	1.00	0.94	0.97	0.99	0.85	1.00	0.92	0.98
Manange	0.96	1.00	0.98	1.00	1.00	0.82	0.90	0.98	0.96	1.00	0.98	1.00
Mandarin Chinese	0.98	0.98	0.98	1.00	1.00	0.98	0.99	1.00	0.89	1.00	0.94	0.99
Mapudungun	1.00	0.90	0.95	0.99	1.00	0.96	0.98	1.00	0.95	1.00	0.98	1.00
Old High German	0.85	0.71	0.77	0.96	0.81	0.84	0.82	0.97	0.86	0.95	0.90	0.98
Oroqen	0.89	0.86	0.87	0.97	1.00	0.70	0.82	0.96	0.55	1.00	0.71	0.96
Otomi	1.00	0.98	0.99	1.00	0.98	0.98	0.98	1.00	0.94	0.98	0.96	0.99
Q'eqchi'	0.96	0.90	0.92	0.98	0.97	0.94	0.95	0.99	0.95	1.00	0.97	0.99
Romanian	0.96	0.83	0.89	0.98	0.97	0.85	0.91	0.97	0.86	1.00	0.93	0.99
Sakha	0.90	0.96	0.93	0.98	0.92	0.83	0.87	0.97	0.80	1.00	0.89	0.99
Saramaccan	1.00	0.87	0.93	0.98	1.00	0.94	0.97	0.99	0.93	1.00	0.97	0.99
Selice Romani	0.94	0.88	0.91	0.98	0.84	0.94	0.89	0.98	0.80	1.00	0.89	0.98
Seychelles Creole	0.95	0.91	0.93	0.99	0.95	0.84	0.89	0.98	0.95	1.00	0.97	1.00
Swahili	0.97	1.00	0.99	1.00	1.00	0.97	0.98	1.00	0.92	0.96	0.94	0.99
Takia	1.00	0.80	0.89	0.98	0.93	0.93	0.93	0.98	0.95	1.00	0.97	1.00
Tarifiyt Berber	0.94	0.94	0.94	0.99	0.95	0.74	0.83	0.96	0.85	1.00	0.92	0.98
Thai	0.95	0.82	0.88	0.97	0.97	0.85	0.91	0.98	0.67	0.96	0.79	0.97
Vietnamese	0.97	1.00	0.99	1.00	1.00	0.97	0.98	1.00	0.90	1.00	0.95	0.99
White Hmong	1.00	1.00	1.00	1.00	1.00	0.96	0.98	1.00	0.92	1.00	0.96	0.99
Wichí	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Yaqui	0.95	1.00	0.98	1.00	1.00	1.00	1.00	1.00	0.91	1.00	0.96	0.99
Zinacantán Tzotzil	1.00	0.97	0.98	1.00	0.90	0.78	0.84	0.97	0.80	1.00	0.89	0.98
Mean over languages	0.97	0.93	0.95	0.99	0.96	0.87	0.91	0.98	0.87	0.99	0.92	0.99

Table 8. Fake words - 10% borrowing - metrics by language.

	Recurrent neural net				Markov model				Bag of sounds			
Language	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**
Archi	0.98	0.98	0.98	0.99	0.96	0.94	0.95	0.98	0.88	0.93	0.91	0.96
Bezhta	0.98	0.97	0.97	0.99	0.96	0.96	0.96	0.98	0.96	0.96	0.96	0.98
Ceq Wong	0.97	0.85	0.91	0.96	0.94	0.91	0.92	0.97	0.89	1.00	0.94	0.98
Dutch	0.88	0.77	0.82	0.92	0.83	0.78	0.81	0.91	0.78	1.00	0.88	0.96
English	0.93	0.89	0.91	0.97	1.00	0.80	0.89	0.96	0.83	1.00	0.91	0.97
Gawwada	0.97	1.00	0.98	0.99	0.95	0.93	0.94	0.98	0.93	1.00	0.96	0.98
Gurindji	1.00	0.98	0.99	1.00	1.00	0.97	0.99	1.00	0.92	1.00	0.96	0.98
Hausa	0.98	1.00	0.99	1.00	0.98	0.98	0.98	0.99	0.93	0.99	0.96	0.98
Hawaiian	1.00	1.00	1.00	1.00	0.98	1.00	0.99	1.00	0.99	1.00	0.99	1.00
Hup	1.00	1.00	1.00	1.00	0.98	0.98	0.98	0.99	0.82	1.00	0.90	0.97
Imbabura Quechua	1.00	1.00	1.00	1.00	0.97	1.00	0.99	1.00	0.93	1.00	0.96	0.99
Indonesian	1.00	1.00	1.00	1.00	0.97	0.97	0.97	0.99	0.97	1.00	0.99	0.99
Iraqw	0.96	0.96	0.96	0.99	0.96	0.96	0.96	0.99	0.88	1.00	0.94	0.98
Japanese	1.00	1.00	1.00	1.00	0.99	0.97	0.98	0.99	0.94	1.00	0.97	0.99
Kali'na	1.00	1.00	1.00	1.00	0.98	1.00	0.99	1.00	0.96	1.00	0.98	0.99
Kanuri	0.98	0.98	0.98	0.99	0.98	1.00	0.99	1.00	0.90	1.00	0.95	0.98
Ket	1.00	1.00	1.00	1.00	0.95	0.90	0.92	0.97	0.93	0.98	0.95	0.98
Kildin Saami	1.00	0.96	0.98	0.99	0.91	0.84	0.87	0.95	0.92	0.98	0.95	0.98
Lower Sorbian	1.00	0.99	0.99	1.00	0.98	0.96	0.97	0.99	0.92	1.00	0.96	0.98
Malagasy	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.97	1.00	0.98	0.99
Manange	1.00	0.98	0.99	1.00	1.00	0.89	0.94	0.98	0.92	1.00	0.96	0.98
Mandarin Chinese	1.00	0.97	0.98	0.99	1.00	0.98	0.99	1.00	0.93	1.00	0.96	0.98
Mapudungun	1.00	0.98	0.99	1.00	1.00	1.00	1.00	1.00	0.96	0.98	0.97	0.99
Old High German	0.96	0.82	0.89	0.96	0.91	0.81	0.86	0.95	0.89	0.94	0.92	0.97
Oroqen	1.00	0.98	0.99	1.00	0.86	0.85	0.85	0.94	0.67	0.92	0.78	0.93
Otomi	0.99	0.97	0.98	0.99	1.00	0.99	1.00	1.00	0.91	0.98	0.95	0.98
Q'eqchi'	1.00	0.97	0.98	0.99	0.99	0.95	0.97	0.99	0.90	1.00	0.95	0.98
Romanian	0.94	0.93	0.93	0.97	0.90	0.90	0.90	0.96	0.85	0.92	0.88	0.96
Sakha	0.98	0.97	0.98	0.99	0.96	0.94	0.95	0.98	0.81	1.00	0.89	0.96
Saramaccan	1.00	0.94	0.97	0.99	1.00	0.96	0.98	0.99	0.90	0.93	0.92	0.98
Selice Romani	1.00	0.95	0.97	0.99	1.00	0.87	0.93	0.97	0.91	1.00	0.95	0.98
Seychelles Creole	0.99	0.99	0.99	1.00	0.96	0.93	0.95	0.98	0.89	1.00	0.94	0.98
Swahili	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00	0.89	0.98	0.94	0.98
Takia	0.96	0.96	0.96	0.98	1.00	0.89	0.94	0.98	0.91	1.00	0.95	0.98
Tarifiyt Berber	1.00	0.98	0.99	1.00	0.91	0.89	0.90	0.96	0.94	0.92	0.93	0.98
Thai	0.98	0.94	0.96	0.98	0.96	0.88	0.92	0.97	0.88	1.00	0.94	0.98
Vietnamese	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	0.96	1.00	0.98	0.99
White Hmong	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Wichí	1.00	0.96	0.98	0.99	1.00	0.98	0.99	1.00	1.00	1.00	1.00	1.00
Yaqui	0.99	1.00	0.99	1.00	0.98	0.98	0.98	0.99	0.94	1.00	0.97	0.99
Zinacantán Tzotzil	1.00	0.98	0.99	1.00	0.98	1.00	0.99	1.00	0.91	1.00	0.96	0.98
Mean over languages	0.99	0.97	0.98	0.99	0.97	0.94	0.95	0.98	0.91	0.99	0.94	0.98

Table 9. Fake words - 20% borrowing - metrics by language.

S2 Cross validation Detection results by language for 10-fold cross validation on WOLD wordlists

	Recurrent neural net				Markov model				Bag of Sounds				
Language	**Prec.**	**Recall**	**F1**	**Acc**	**Prec.**	**Recall**	**F1**	**Acc**	**Prec.**	**Recall**	**F1**	**Acc**	**Native**
Archi	0.747	0.603	0.664	0.840	0.740	0.530	0.615	0.803	0.249	0.687	0.359	0.814	0.786
Bezhta	0.799	0.759	0.778	0.863	0.770	0.704	0.735	0.834	0.689	0.757	0.720	0.840	0.701
Ceq Wong	0.770	0.710	0.736	0.818	0.690	0.619	0.650	0.753	0.581	0.649	0.610	0.754	0.667
Dutch	0.625	0.492	0.545	0.807	0.557	0.432	0.484	0.778	0.010	0.200	0.019	0.815	0.814
English	0.651	0.624	0.637	0.716	0.630	0.601	0.613	0.696	0.510	0.685	0.584	0.720	0.614
Gawwada	0.661	0.367	0.457	0.864	0.590	0.357	0.438	0.862	0.179	0.603	0.254	0.916	0.909
Gurindji	0.428	0.232	0.291	0.753	0.525	0.301	0.376	0.796	0.000	0.000	0.000	0.875	0.875
Hausa	0.654	0.527	0.579	0.813	0.702	0.520	0.595	0.814	0.288	0.664	0.395	0.830	0.802
Hawaiian	0.783	0.430	0.554	0.797	0.727	0.475	0.572	0.826	0.022	0.400	0.041	0.843	0.839
Hup	0.869	0.629	0.727	0.936	0.807	0.527	0.625	0.903	0.464	0.908	0.593	0.939	0.899
Imbabura Quechua	0.861	0.780	0.816	0.892	0.859	0.791	0.821	0.897	0.615	0.762	0.676	0.839	0.720
Indonesian	0.689	0.619	0.651	0.778	0.657	0.608	0.630	0.768	0.265	0.636	0.371	0.732	0.698
Iraqw	0.787	0.560	0.647	0.894	0.690	0.459	0.545	0.855	0.377	0.760	0.493	0.901	0.870
Japanese	0.822	0.723	0.767	0.853	0.767	0.703	0.733	0.835	0.413	0.680	0.513	0.768	0.704
Kali'na	0.714	0.506	0.590	0.857	0.660	0.537	0.589	0.868	0.208	0.967	0.335	0.885	0.855
Kanuri	0.635	0.444	0.518	0.794	0.598	0.429	0.493	0.788	0.055	0.683	0.101	0.830	0.823
Ket	0.779	0.470	0.582	0.906	0.683	0.402	0.486	0.885	0.207	0.750	0.315	0.930	0.916
Kildin Saami	0.560	0.454	0.497	0.790	0.562	0.403	0.469	0.758	0.003	0.050	0.006	0.805	0.810
Lower Sorbian	0.744	0.605	0.664	0.856	0.713	0.602	0.649	0.849	0.215	0.752	0.328	0.831	0.803
Malagasy	0.602	0.373	0.452	0.821	0.559	0.365	0.437	0.821	0.000	0.000	0.000	0.875	0.875
Manange	0.593	0.293	0.380	0.881	0.638	0.274	0.360	0.859	0.045	0.300	0.079	0.937	0.935
Mandarin Chinese	0.050	0.006	0.011	0.955	0.190	0.008	0.016	0.811	0.000	0.000	0.000	0.993	0.993
Mapudungun	0.816	0.664	0.727	0.879	0.801	0.636	0.707	0.868	0.534	0.832	0.645	0.885	0.800
Old High German	0.351	0.189	0.241	0.887	0.450	0.176	0.246	0.860	0.000	0.000	0.000	0.947	0.947
Oroqen	0.530	0.271	0.354	0.857	0.497	0.267	0.342	0.856	0.054	0.350	0.091	0.925	0.922
Otomi	0.908	0.709	0.793	0.954	0.929	0.629	0.749	0.939	0.673	0.854	0.751	0.957	0.902
Q'eqchi'	0.852	0.651	0.733	0.935	0.820	0.597	0.689	0.923	0.540	0.816	0.647	0.939	0.895
Romanian	0.724	0.698	0.710	0.764	0.716	0.668	0.690	0.743	0.412	0.623	0.493	0.663	0.600
Sakha	0.632	0.599	0.610	0.800	0.620	0.543	0.577	0.782	0.196	0.595	0.290	0.766	0.751
Saramaccan	0.622	0.589	0.603	0.714	0.632	0.596	0.611	0.718	0.089	0.669	0.149	0.652	0.645
Selice Romani	0.872	0.905	0.888	0.874	0.875	0.878	0.876	0.859	0.829	0.746	0.784	0.740	0.427
Seychelles Creole	0.568	0.272	0.364	0.828	0.606	0.323	0.420	0.854	0.000	0.000	0.000	0.911	0.911
Swahili	0.783	0.680	0.723	0.857	0.700	0.623	0.658	0.825	0.536	0.786	0.635	0.851	0.758
Takia	0.808	0.618	0.697	0.839	0.762	0.617	0.680	0.834	0.047	0.700	0.087	0.780	0.768
Tarifiyt Berber	0.764	0.795	0.778	0.788	0.765	0.772	0.768	0.774	0.695	0.773	0.731	0.750	0.511
Thai	0.655	0.526	0.581	0.799	0.622	0.450	0.521	0.754	0.104	0.630	0.175	0.794	0.785
Vietnamese	0.668	0.463	0.544	0.795	0.579	0.411	0.477	0.770	0.101	0.601	0.166	0.821	0.817
White Hmong	0.597	0.373	0.457	0.785	0.607	0.354	0.443	0.767	0.025	0.400	0.046	0.846	0.845
Wichí	0.873	0.705	0.773	0.931	0.848	0.729	0.781	0.935	0.518	0.698	0.583	0.900	0.857
Yaqui	0.819	0.736	0.773	0.885	0.839	0.764	0.798	0.897	0.567	0.794	0.658	0.861	0.760
Zinacantán Tzotzil	0.906	0.751	0.815	0.940	0.836	0.675	0.744	0.919	0.430	0.935	0.584	0.913	0.857
Mean over languages	0.697	0.546	0.603	0.844	0.678	0.521	0.578	0.828	0.286	0.578	0.349	0.843	0.797

Table 10. 10-fold cross-validation by language - Means.

	Recurrent neural net				Markov model				Bag of Sounds			
Language	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**	**Prec.**	**Recall**	**F1**	**Acc.**
Archi	0.073	0.071	0.052	0.029	0.107	0.093	0.087	0.045	0.078	0.099	0.084	0.031
Bezhta	0.052	0.035	0.035	0.030	0.037	0.054	0.040	0.029	0.045	0.070	0.044	0.033
Ceq Wong	0.044	0.104	0.072	0.047	0.115	0.089	0.089	0.066	0.083	0.085	0.072	0.053
Dutch	0.149	0.105	0.103	0.051	0.076	0.069	0.061	0.042	0.016	0.350	0.031	0.030
English	0.046	0.051	0.042	0.021	0.062	0.073	0.055	0.043	0.029	0.047	0.024	0.018
Gawwada	0.177	0.093	0.080	0.042	0.101	0.107	0.100	0.029	0.158	0.379	0.201	0.021
Gurindji	0.170	0.094	0.099	0.046	0.234	0.126	0.151	0.031	0.000	0.000	0.000	0.036
Hausa	0.081	0.066	0.048	0.028	0.079	0.095	0.086	0.033	0.098	0.138	0.109	0.036
Hawaiian	0.064	0.054	0.057	0.032	0.086	0.077	0.078	0.037	0.031	0.516	0.057	0.030
Hup	0.122	0.086	0.089	0.019	0.138	0.158	0.120	0.035	0.170	0.102	0.151	0.024
Imbabura Quechua	0.065	0.046	0.035	0.022	0.053	0.062	0.037	0.018	0.107	0.092	0.084	0.037
Indonesian	0.035	0.061	0.045	0.029	0.058	0.059	0.047	0.034	0.067	0.085	0.078	0.041
Iraqw	0.155	0.146	0.134	0.031	0.134	0.091	0.095	0.025	0.093	0.126	0.080	0.022
Japanese	0.041	0.060	0.034	0.022	0.045	0.017	0.021	0.022	0.068	0.092	0.076	0.047
Kali'na	0.095	0.087	0.084	0.029	0.087	0.122	0.104	0.035	0.084	0.105	0.118	0.027
Kanuri	0.089	0.071	0.060	0.041	0.126	0.077	0.085	0.038	0.039	0.328	0.068	0.024
Ket	0.107	0.099	0.100	0.023	0.150	0.178	0.148	0.033	0.132	0.362	0.180	0.023
Kildin Saami	0.093	0.099	0.087	0.024	0.042	0.038	0.036	0.043	0.010	0.158	0.019	0.028
Lower Sorbian	0.089	0.084	0.072	0.031	0.103	0.059	0.062	0.035	0.084	0.110	0.106	0.041
Malagasy	0.137	0.095	0.083	0.035	0.087	0.097	0.089	0.035	0.000	0.000	0.000	0.020
Manange	0.162	0.126	0.137	0.026	0.176	0.140	0.143	0.038	0.073	0.483	0.127	0.025
Mandarin Chinese	0.158	0.019	0.033	0.019	0.341	0.014	0.026	0.041	0.000	0.000	0.000	0.007
Mapudungun	0.105	0.105	0.087	0.040	0.092	0.103	0.089	0.037	0.092	0.124	0.085	0.027
Old High German	0.156	0.060	0.078	0.023	0.147	0.066	0.081	0.024	0.000	0.000	0.000	0.025
Oroqen	0.137	0.091	0.101	0.035	0.173	0.103	0.117	0.031	0.078	0.474	0.130	0.021
Otomi	0.074	0.112	0.091	0.019	0.057	0.059	0.048	0.016	0.090	0.054	0.069	0.011
Q'eqchi'	0.076	0.101	0.075	0.023	0.073	0.094	0.084	0.020	0.106	0.114	0.104	0.018
Romanian	0.028	0.037	0.022	0.019	0.037	0.045	0.031	0.030	0.056	0.060	0.038	0.030
Sakha	0.073	0.074	0.047	0.029	0.112	0.116	0.110	0.035	0.069	0.154	0.088	0.027
Saramaccan	0.074	0.093	0.080	0.037	0.069	0.058	0.053	0.044	0.048	0.224	0.068	0.044
Selice Romani	0.022	0.029	0.017	0.019	0.021	0.031	0.018	0.019	0.042	0.033	0.026	0.025
Seychelles Creole	0.089	0.048	0.046	0.025	0.114	0.048	0.064	0.016	0.000	0.000	0.000	0.031
Swahili	0.072	0.089	0.057	0.020	0.074	0.054	0.057	0.027	0.061	0.070	0.051	0.028
Takia	0.059	0.070	0.044	0.019	0.081	0.090	0.084	0.056	0.046	0.483	0.082	0.033
Tarifiyt Berber	0.054	0.042	0.033	0.029	0.039	0.045	0.031	0.033	0.053	0.062	0.045	0.042
Thai	0.055	0.052	0.038	0.018	0.069	0.045	0.045	0.030	0.054	0.091	0.077	0.019
Vietnamese	0.073	0.064	0.053	0.030	0.095	0.119	0.110	0.035	0.061	0.262	0.095	0.035
White Hmong	0.139	0.073	0.091	0.031	0.150	0.101	0.110	0.038	0.043	0.516	0.076	0.038
Wichí	0.103	0.123	0.090	0.024	0.048	0.089	0.060	0.016	0.150	0.129	0.129	0.024
Yaqui	0.067	0.069	0.054	0.031	0.046	0.066	0.035	0.024	0.072	0.069	0.053	0.018
Zinacantán Tzotzil	0.060	0.097	0.041	0.021	0.060	0.098	0.077	0.025	0.102	0.093	0.102	0.028
Pooled std. dev.	0.100	0.082	0.072	0.030	0.114	0.088	0.082	0.034	0.078	0.226	0.088	0.030

Table 11. 10-fold cross-validation by language - Standard Deviations.