# Detecting loan words computationally

Liqin Zhang, Franz Manni, Ray Fabri and John Nerbonne[i]

**Abstract**. A loanword is borrowed from one language and adopted into another; e.g. the English words *toboggan*, *skunk*, and *hickory* were all originally Algonquian. Among languages not (closely) related, loan words are recognizable because they are both semantically and phonetically more similar to each other than one would expect by coincidence. We suggest that quantitative measures might be profitably deployed in the study of contact, which Salikoko Mufwene has furthered so impressively. We apply techniques for measuring pronunciation similarity, two edit-distance measures used in dialectology and a sound-class based method. A novel issue in loan-word detection is the circumstance that loan words are usually modified to fit the phonology of the borrowing language, meaning that sensitivity in measuring pronunciation similarity may be deprecated.

**Keywords.** Loan words, automatic detection, edit distance, sound class alignment, language contact

> Neither a borrower nor a lender be;
> For loan oft loses both itself and friend,
> – Hamlet, Act 1, Scene 3

## Introduction

Loan words can provide evidence of social, cultural, commercial or other contact, e.g. when we note that Germanic languages owe their words for paved roads to Latin *via strata* 'way paved' (< Eng. *street*, Germ. *Straße*, Dutch *straat*), suggesting that not just the word, but also the infrastructural innovation was copied from the Romans. Those trying to reconstruct earlier, undocumented stages of languages must at times wish that languages had heeded Polonius's advice to Hamlet, and avoided borrowing altogether, since loan words confuse the historical record, normally suggesting closer phonetic similarity than is actually warranted. This paper wishes to contribute to the automatic detection of loan words in non-related languages and thereby to the study of language contact which Mufwene has so greatly advanced.

The basic idea is simple: if two words from unrelated languages mean roughly the same thing and are similar in pronunciation, then the chances are that one has been borrowed from the other's language. It would be too great a coincidence for the similarity to arise by chance. This very simple characterization reveals, too, that our detection will be symmetric. Words that are semantically and phonetically similar may be classified as involving a borrowing, but we will not attempt to say which language borrowed from the other – or indeed, whether the two languages borrowed from a third.

The restriction to focusing on loan word detection in non-related languages is important because, while loans from related languages may be semantically and phonetically similar, non-loans may also be semantically and phonetically similar due to their historical relatedness. Words are certainly borrowed from related as well as from unrelated languages: the English word *skirt* is a borrowing from the related Old Norse *skyrta* 'shirt', arising from the same older common Germanic word that survived in English *shirt*. Note that *skirt* is phonetically more similar to *skyrta* than *shirt* is, but that *shirt* is more similar in meaning to the hypothetical source. The sorts of procedures we examine below would be unlikely to distinguish these two words since both are phonetically similar to the putative Old Norse source, as well as being semantically related. The procedures would not distinguish the two English words well, even though only one of them is a genuine borrowing. In focusing on borrowings from unrelated languages, we avoid this problem. By working on the simpler problem we hope to make progress more likely. Naturally, an approach that detects borrowings without such a restriction will be superior in the long term.

We will operationalize this idea by examining the results of field work carried out in Central Asia (Mennecier et al. 2016). The informants were speakers of either Turkic languages or Indo-Iranian languages, two unrelated language families. The informants were asked to verbalize concepts found in the 200-word Swadesh list (Swadesh 1952), and their answers were recorded acoustically and transcribed in IPA. Reacting to the same concept will be interpreted here as indicating semantic similarity, and we will examine several ways of estimating phonetic similarity based on phonetic transcriptions, focusing on finding the best algorithm for detecting the pronunciation similarity in loan words.

## Mufwene's perspective

In a series of publications Salikoko Mufwene has urged that we view linguistic evolution on analogy to biological evolution (Mufwene 2001, 2005, 2008). This is a view dating back at least to Darwin and Schleicher (Mufwene 2005), but Mufwene more particularly encourages the view that languages are analogous to species consisting of many, often very different individual organisms, just as languages may be viewed as collections of many, very different idiolects, and he emphasizes the essential role of the environment in understanding how these populations of idiolects develop. This view is sympathetic to that of linguistic variationism, emphasizing the individual variation normally ignored when one compares only the enormous abstractions, the languages. Mufwene (2005) suggests, e.g., that the process of acquisition should not be thought of as a direct transmission from a parent-caretaker to a child, but rather as an active construction of an idiolect by the child, based on the many varieties it is exposed to, emphasizing "the piecemeal way in which speakers develop competence" (p.37). The emphasis is on the enormous variation the learner is confronted with, which leads him or her to select linguistic elements, sometimes in parallel, especially in large, heterogeneous societies.

In emphasizing Mufwene's evolutionary perspective we have elsewhere encountered the objection that modern linguists see little or no evidence of primitive languages from which more sophisticated ones have evolved. The objection is correct: evolutionary progress in this sense is not postulated nor is it necessary to an understanding of how Mufwene sees languages changing.[ii] Instead the emphasis is on how language changes in response to its ecology, and in particular, to the other languages in use nearby, where loanwords are strikingly interesting. Clearly languages adopt elements from other languages and thereby "adapt" – socially and culturally – to their environments.

Mufwene (2001) applies ideas on population genetics to the analysis of colonial varieties, noting that variation in languages, just as in gene pools in biological populations, is likely to be reduced when a relatively small sample of speakers emigrates to a colony. Succeeding generations in the isolated colony normally have to select their variants from the reduced pool, maintaining (the tendency toward) the initial selections. The so-called FOUNDER EFFECT in population genetics (Jobling, Hurles & Tyler-Smith 2004)[iii] thus has an analogue in linguistics: early immigrants have an inordinate influence on the populations they engender. Shackleton (2010) traces New England speech to East Anglia, and Virginian speech to the southwest of Great Britain. The largest overall similarities were found between these two pairs of areas.

This view thus seems congenial with respect to acquisition and the study of colonial varieties, but it is in the study of pidgin and creole languages where it has been found most convincing. Since pidgin and creole languages arise in a multilingual ecology, it is only natural for their speakers to adopt elements from the different languages they hear. The better established pidgin and creole varieties therefore contain elements selected from their original multilingual environment and transmitted to later generations of language learners.

Mufwene's perspective resonates with the research line from which the present paper arises for several reasons, so that we think it is more than just another instance of the "contagion of ideas" (Sperber 1997). Most importantly, we detect resonance because Mufwene's theoretical perspective is best served by a quantitative methodology of the sort we employ here. Given that languages are extremely variable, then comparisons intended to establish genealogical or areal relations (those arising from contact) must be based on large, representative samples analyzed statistically. Finding a feature *f* in a given variety (a sound, a word, an inflectional affix, or a grammatical construction) that is also found in another, potentially influential variety is always a striking observation. But given that languages consist of dozens of sounds, tens of thousands of words, (often) dozens of morphological elements and hundreds of grammatical constructions, it is incumbent on those wishing to demonstrate a genuine relationship that elements have indeed been transmitted from earlier varieties or from others in areal contact. This can be done if large amounts of data are analyzed, preferably from large numbers of varieties. This is exactly our tack in this paper.

Less importantly, Mufwene's perspective shares a good deal with the variationist program our own work has proceeded from. As noted above, his work assumes that there is a great deal of linguistic variation, not only among the dialects or varieties of a language, but also within those. This has consistently been our experience (Nerbonne 2009). Finally, like Mufwene, we have been active in promoting the collaboration between population genetics and (variationist) linguistics (Manni, Heeringa & Nerbonne 2006; Manni 2017), e.g., in showing that the number of loan words is proportional to social contact in neighboring populations (Mennecier et al. 2016).

## Previous work

Greg Kondrak has worked regularly on the task of "cognate identification", not only in machine translation (Kondrak, Marcu & Knight 2003) but also in historical linguistics (Kondrak & Sherif 2006). In Kondrak's work on machine translation, "cognates" include what we call loanwords, so there is a fairly direct connection. He has compared both linguistically inspired methods, such as the ones we focus on here, but also sophisticated machine learning methods, such as pair Hidden Markov Models and dynamic Bayesian Networks, with the latter tending to be more accurate, reducing error by approximately 10% (Kondrak & Sherif 2006, Wieling, et al. 2007). Our work differs in focusing on what linguists regard as loanwords, rather than "cognates", but it is clear that the problem of detecting loanwords is quite similar to that of detecting cognates in historical linguistics and to detecting "cognates" in the broader sense of machine translation.

We are also aware of work done in linguistic phylogenetics (Delz, 2013) and ancestral state reconstruction (Köllner & Dellert, 2016), which undertake extensive historical reconstruction in order to classify words as borrowings (or non-borrowings). These approaches have the advantage of attempting to detect loans on the basis of language history and may check the plausibility of a native (non-loan) source of a word, but they also require that the language histories be reconstructed. In this paper we attempt to avoid that step by checking directly for unexpected semantic and phonetic similarity in the synchronic data. Our approach yields less information than the others, but, being less ambitious, may also be more feasible. There are undoubtedly studies in which identifying loan words is itself interesting, even without an account of the entire history of the languages involved.

Finally, Johannes Dellert has applied causal inference techniques to the problem of cognate detection (Dellert 2017, 2018), combining information about pronunciation and semantic similarity with (induced) models of language relatedness. In contrast to the methods in this paper, Dellert's techniques do not require that data be hand-annotated (in order to set a threshold). We leave it to future work to examine his ideas more thoroughly.

## Data

Mennecier et al. (2016) conducted a survey to explore the language variety of a Central Asian region and then utilized the data to measure the relatedness of languages and to attempt to detect loanwords. The data, which is documented and publicly available for the study of loanwords detection in this experiment,[iv] was collected from 23 sites in three Central Asian countries, namely Uzbekistan, Kyrgyzstan, and Tajikistan. The sites were chosen for their "complex human and linguistic geography". There were 88 informants from the three countries. For reasons having to do with genetics, males over 40 years old were preferred. Linguistic and genetic sampling proceeded in parallel in order to examine the linguistic and genetic histories of the peoples, and in particular to see whether genetic commonalities were paralleled by similarity in culture (language). This aspect of the work will be reported on separately. The native languages of the informants were Kazakh, Kyrgyz, Karakalpak, Uzbek, Tajik, and Yaghnobi, all languages from two language families, Turkic and Indo-Iranian (see Table 1). The informants also understood Russian well since they all went to school during the times of the Union of Soviet Socialist Republics (USSR).

Table 1 The languages examined in this study.

| Turkic | Indo-Iranian |
|---|---|
| Kazakh | Tajik |
| Karakalpak | Yagnobi |
| Kyrgyz | |
| Uzbekh | |

A 200-word extended Swadesh list was first reduced to 178 words, eliminating words that were polysemous or too difficult to understand in the interview context. The reduced list was then presented in Russian to the informants, who were asked to translate the words in the list orally into their native languages. It is clear that presenting the words in Russian introduces a bias in facilitating responses involving Russian loan words. The fieldworker attempted to probe in such cases, but one must be aware of the potential for bias. Each word in the Swadesh list represents a concept. The pronunciations were digitally recorded and catalogued in phonetic transcriptions. In total, each informant was asked to produce 178 pronunciations, resulting in more than 15,000 recordings. There are therefore approximately 88 phonetic transcriptions for each Russian word representing a concept, even if some could not be used, for example, when an informant did not pronounce a word clearly enough for transcription. Philipe Mennecier transcribed all of the data.

An expert classification of the words into cognate classes is available in the dataset as well, which will make our evaluation straightforward (see below). Within a concept, each pronunciation is marked with a code, so that the pronunciations with identical codes are designated cognates. Hence, a word from a Turkic language (or an Indo-Iranian one) bearing the same cognate designation as another word in the Indo-Iranian family (or, respectively, the Turkic family) means that one of the words is a loanword. Notably, it is common that pronunciations of a concept in one language family are assigned to different cognate classes because there are multiple languages in a language family, and because informants may know multiple ways to translate a Russian word representing a concept. Besides, each phonetic transcript is coded according to its original language or language family. The pronunciations in the dataset originated from Turkic, Iranian, Arabic, and Russian.

## Example data

To illustrate the procedure more concretely, we provide a sample of the data in Table 2. Our procedure will compare all the pairs in T×I, where T is the set of 49 Turkic pronunciations and I the set of 39 Indo-Iranian ones. Both in the case of ‘one’ (first column) and in the case of ‘three’ (rightmost column), the last few pronunciations, involving [iː] as ‘one’ and [traj] etc. as ‘three’ suggest that a novel lexical item has entered the Indo-Iranian varieties. But note that in neither of these cases do we find closer similarity to the Turkic realizations.

Table 2 A sample of the data used in comparing pronunciations in order to detect borrowings. The diagonal slash is introduced to separate pronunciations from different sampling sites. The double slash in the first row of the Turkic pronunciations of ‘one’ indicates missing data.

| | ‘one’ | ‘three’ |
|---|---|---|
| Turkic | bɪr / brɪw / bɪrɪw / / brɪw / brɪw / bɪr / bər / bər / bər / bər / bir / bir / bir / bir̯ / bir̯ / bir / bir / bir / bir / bɪr / bər / bɪʃ / biʃ / bir / bər / bər / bər / bər / bər / bər / bər / bər / bər̯ / bər / bər / bɪr / ber / bər / bɪr / bər / bər / bər / bɪr / bɪr / bɪr / bɪr / bɪr / bər / | ʉʃʲ / ʉʃʲ / ʉʃʲjʊː / ʉʃ / ʉʃ / ʃuː / ʉʃ / yʃ / yʧ / øʃ / ʉʃ / yʧ / yʧ / yʧ / yʧ / yʧ / yʧ / yʧ / yʧ / yʧ / ʉʧ / yʧ / yʧ / yʧ / yʧ / ʉʃ / ʉʃ / ʉʃ / yʧ / ʉʃ / yʧ / ʉʃ / ʉʃ / ʉʃ / ʉʃ / ʉʃ / yʧ / ʉʧ / ʉʃ / uʧ / yʧ / yʧ / yʧ / ʉʧ / uʃta / ʉʧ / ʉʧ / uʧ / uʧ / |
| Indo-Iranian | jak / jak / jak / jakta / jak / jak / jak / jak / jak / jak / jak / jaktə / jaktɐ / jak / jak / jak / jak / jakta / jak / jak / jak / jak / jak / jak / jaktə / jak / jakta / jak / jak / jak / jaktɒ / jak / jakta / jak / iː / iː / iː / iː / iː / | sʲɛ / se / se / setta / sʲe / se / sɛ / se / sʲe / sʲɛ / se / sʲetɐ / sʲetɐ / se / se / se / se / sʲeta / se / se / se / se / sʲe / se / sʲeta / seː / seˑta / sʲeː / se / se / seta / se / sʲeta / sʲe / saraj / traj / tʲiraj / traj / tɪraj / |

## Measuring pronunciation similarity

We compare three different algorithms that have been used to gauge pronunciation similarity. In fact all the algorithms produce dissimilarity measures, but by looking at the pairs of words that are dissimilar to only a small degree, we obtain the best candidate loanwords, just as we wish. The first two methods were developed within dialectology, and the third within historical linguistics.

The first method we examine is Heeringa’s (2004) modification of the edit-distance or Levenshtein algorithm. The Levenshtein algorithm has been used frequently as a measure of spelling and pronunciation similarity (Nerbonne 2003) and functions by calculating the least costly set of operations needed to transform one string into another, where the operations are normally restricted to insertion, deletion, substitution and sometimes a transpose operation that might model metathesis. Heeringa (2004) modified the algorithm in order to ensure that substituting a sound for a similar one is less costly. He experimented with modifications based on feature systems, but the best performance was obtained from a version in which the similarity of two sounds was determined by measuring how close their spectrographic representations were. He used a demonstration recording of the IPA and measured the distance between the curves at a large number (of combinations) of points of time and frequencies. He used a logarithmic correction of the distance in keeping with psychoacoustic practice. Heeringa was also able to assign costs to insertions and deletions by measuring the distance between silence and the sound being inserted or deleted. We refer to this manner of determining pronunciation similarity (and dissimilarity) as the SPECTROGRAM METHOD.

Proceeding from the same edit distance algorithm, Wieling et al. (2012) exploited one of the most useful properties of the procedure, namely, that, in determining the difference between two strings, the algorithm automatically induces an alignment in which corresponding elements can be identified. For example, the pronunciation of the German *Durst* ‘thirst’ is [tɔʃt] in Vielbrunn and [tʊəʃ] in Aachen (in *Kleine Deutsche Lautatlas*, see Nerbonne & Siedle 2005), and the algorithm produces the following as an alignment:

| | | | | |
|---|---|---|---|---|
| t | ɔ | | ʃ | t |
| t | ʊ | ə | ʃ | |

Wieling et al. used 0/1 substitution costs to align 200 word pairs at 20,000 pairs of sites and collected the frequencies with which sounds (including the “null sound” in insertions and deletions) appeared in alignment in a large contingency table. They then recalculated the substitution costs, assigning lower costs to all those sound pairs that were frequently aligned. They then iterated the alignment procedure and recalculation of substitution costs until no further alignment changes were noted. This method thus assigns low distances to word pairs with frequent sound correspondences, which is like the importance assigned in historical linguistics to regular sound correspondence. Because the recalculation was performed using POINTWISE MUTUAL INFORMATION, an information-theoretic measure, we refer to this technique as the PMI-BASED METHOD.

Finally, List (2012) developed an alternative measure of pronunciation difference especially focused on application in historical linguistics. We present this technique based on List et al. (2018), as well. List proceeds from sound classes, e.g., bilabial obstruents {[p], [b]}, which often correspond in historical linguistics. After assigning all sounds to their classes, List aligns the sounds based on their classes, which is why his technique is known as SOUND CLASS ALIGNMENT (SCA). List derived his original classes from Dogopolsky (1964), whose set of ten List expanded to 28. While we shall note several differences between SCA and the other two methods (Spectrogram-based and PMI-based) below, one difference can be noted immediately. While both edit-distance measures discriminate sensitively, the SCA ignores at least initially all the differences of sounds in the same classes, e.g. [p] and [b]. The edit-distance-based measures are more sensitive, eschewing the equivalence classes of sounds in the SCA approach, which leads to fine distinctions being ignored by the SCA. Of course, ignoring fine distinction might be an advantage in detecting loan words, since loan words are often forced into the phonology of the borrowing languages, which may lead to substantial differences. The SCA procedure also assigns weights (contributions to distance) for pairs of classes, and here a second difference emerges. While the edit-distance measures are symmetric, SCA aims to model historical development, and therefore assigns different weights, depending on whether one is measuring the likelihood of segment $s_1$ developing into $s_2$, or vice versa. Once SCA creates a basic alignment, the overall similarity may be further adjusted depending on the prosodic contours of the words, and finally, also on the exact phonetic realization of the segments (not just the classes). These later steps in processing incorporate levels of sensitivity which suggest that the less discriminating classes may not be so important. The role of the classes is limited to determining alignment, and distance further depends on prosodic contour and the detailed phonetic nature of the phonetic segments.

Software implementing the different measures was provided by Martijn Wieling for the PMI-base method, by Wilbert Heeringa for the spectrogram-based method, and through LingPy for the SCA method. LingPy[v] is a package distributed by Johann-Mattis List.

From previous work (Wieling et al. 2012) we expect the PMI method to be superior to the spectrogram method, but we wish to test the methods on the novel problem of detecting loan words. It is more difficult to predict how well the SCA will fare. In particular, neither edit-distance-based measure (the PMI and spectrogram-based methods) attempts to incorporate asymmetric substitution costs nor to account for the importance of the prosody of the word. SCA uses both asymmetric costs and takes prosodic differences into account) It is also difficult to make predictions with respect to this point because SCA reintroduces segmental sensitivity at a later stage in processing.

## General setup

Recall our basic principle: if two words from unrelated languages meaning roughly the same (are both elicited by the same Swadesh concept) and are similar in pronunciation, then the chances are that one has been borrowed from the other's language (or that they have both been borrowed from a third). We follow Kondrak in assuming that some sort of borrowing is likely (Kondrak, Marcu & Knight 2003; Kondrak & Sherif 2006). To detect this for a given concept in the Swadesh list, we measure the pronunciation difference between every pair of realizations, one from a Turkic language and the other from an Indo-Iranian one. Our hypothesis is that those pairs showing the most similar pronunciations involve a loan.

But how similar do two pronunciations have to be in order to be regarded as loans? We know of no way to answer this question analytically, so we opted for an empirical approach. We had hoped to see a clear break between the distributions of pronunciation distances of pairs of borrowed words and those of words where no borrowing is involved, but no such break emerged from the data. We therefore tested a large number of thresholds empirically and opted for the optimal one. If we keep in mind the prospect of using our approach on new language families, then exhaustively searching for an optimal threshold is impractical. To gauge the likely success of our approach in this situation, we also apply a cross-validation technique.

To gauge the optimal result, i.e. the one using the optimal threshold, we first need to explain how we evaluate a given threshold. As is customary in computational linguistics (CL) when evaluating an automatic process where a human-annotated set is available, we compared the algorithmic results to the "gold truth" of the human annotator (Black, Lafferty and Roukos 1992). CL converged fairly quickly on a scheme borrowed from information retrieval in which both PRECISION and RECALL play a role. In this sort of evaluation, one analyzes a substantial amount of representative material for which the correct analyses have been noted by human experts, in this case loan words. We refer to the automatic classifications as positive in case the procedure deems them a loan, and negative in case it does not. We then distinguish:

1) the genuine loan words correctly classified (true positives, *tp*);
2) the genuine loans incorrectly classified (false negatives, *fn*);
3) the non-loan words incorrectly classified as loans (false positives, *fp*); and finally
4) the non-loan words correctly classified (true negatives, *tn*).

Precision is then the fraction of classifications that are correct (recognized by human experts), *tp/(tp+fp)*, and recall is the fraction of the humanly recognized loans that the process detects, *tp/(tp+fn)*.

Obviously we would like to see both scores as high as possible, i.e. as close to 1 as possible, but note that it is trivial to score very well on one score if one disregards the other. Procedures that uniformly classify everything positively will score perfectly on recall. To overcome this difficulty, we examine a combination of the two scores, the so-called F-score (or F1-score), which is the harmonic mean between the two:

$$F1 = 2 \cdot \frac{precision \ \times recall}{precision + recall}$$

Fig. 1 shows how precision, recall and the F-score range over 200 different thresholds of pronunciation difference used in the experiments with the PMI method. The blue line shows precision, which is naturally quite high at low thresholds, falling steeply from 0.02 on; the green line with dots and dashes shows recall, which is near zero at low thresholds, but climbs steadily; and the dashed red line traces the F-score, which conveniently shows a single peak, the one used in the experiments.

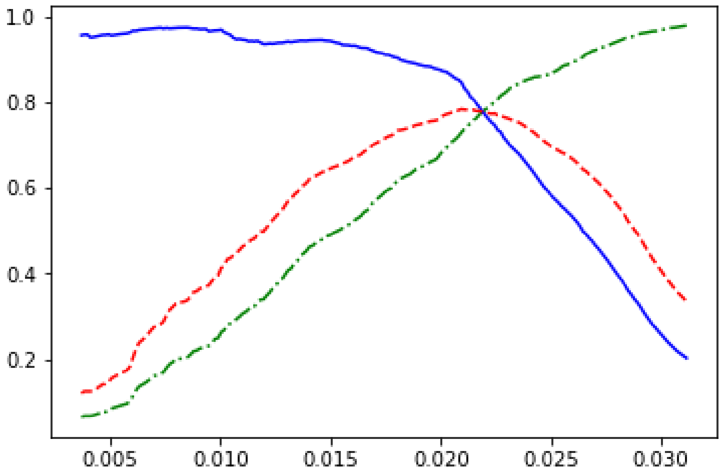


**Figure 1. Precision (solid blue), recall (dash-dotted green), and F score (dashed red) for a range of 200 putative thresholds examined with the PMI method. The x-axis is pronunciation difference and the y-axis the fractional value of precision, recall and F-score. We settled on the threshold where the F score peaks, around 0.022.**

Similar sets of curves were determined for the other two techniques, the spectrogram-based method and the SCA method. Another advantage of determining the threshold in this fashion is that it obviates the need to calibrate the three different scales used by the three pronunciation difference measures.

An alternative way of determining the threshold empirically is to use the entire distribution, examined in quartiles. Tukey (1977) suggested that all the points in a distribution that are more than 1.5 interquartile ranges below the first quartile be regarded as outliers (similarly all the points 1.5 interquartile ranges above the third, but that will not interest us here). This definition is widely used in statistics packages, in particular in the well-known box-and-whisker diagrams, where the low outliers are those below the bottom whisker. Because the concept can be easily understood based on introductory texts, we will not present it in any further detail here.

## Evaluation and results

As expected, all of the algorithms were able to identify a large number of loan words correctly when presented with a list of lexicalizations of the Swadesh concepts in unrelated languages. This is the “gold standard” we are ultimately interested in. But there were also interesting differences.

We first present the distribution of pronunciation difference scores together with the optimal threshold as determined by examining 200 potential thresholds. We also show the border below which outliers (in Tukey’s sense) are found (see Table 4). Note, in particular, the “bump” on the left in the SCA distribution, which shows that this technique assigns low pronunciation difference scores to a rather large number of word pairs.

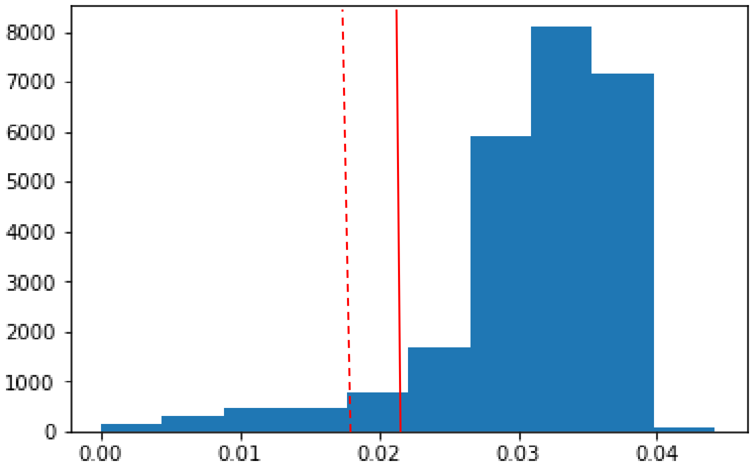
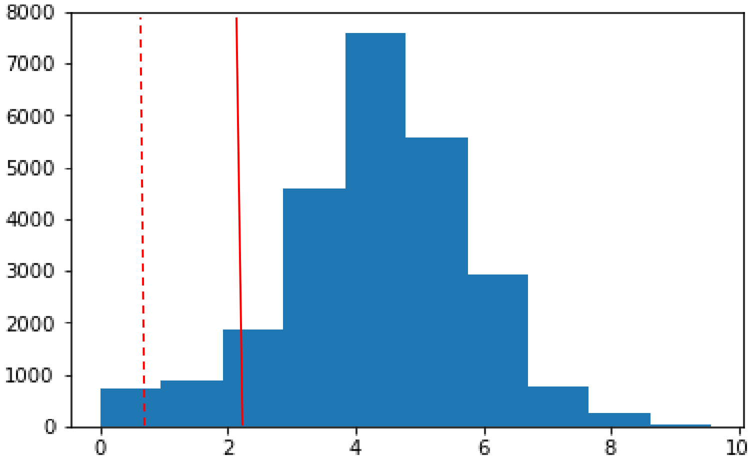
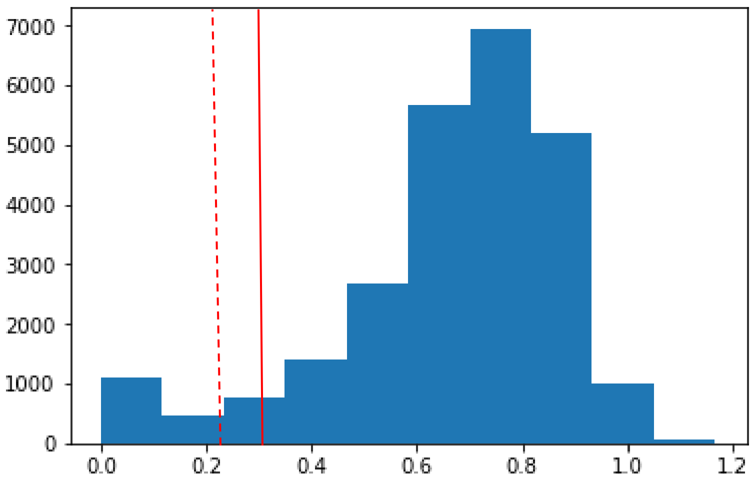
Based on the thresholds found we also evaluated the algorithms on the basis of the same precision, recall and F scores introduced above. Which algorithms detect loan words most effectively? Table 3 summarizes the success of the algorithms in detecting loan words, showing that the SCA method is clearly superior in this task.

Table 3 The performance of the algorithms based on edit distance using pointwise mutual information (PMI) and spectrograms as well as that of the SCA-based method in detecting loan words. The SCA methods is clearly superior.

| | PMI | Spectrogram | SCA |
|---|---|---|---|
| Precision | 0.84 | 0.75 | 0.85 |
| Recall | 0.74 | 0.74 | 0.88 |
| F score | 0.78 | 0.74 | 0.85 |

We should note that, although we have focused our study on maximizing the F score of recognition, there are perspectives which might, e.g., emphasize recall over precision. If the intention is to hand over a list of candidate loan words to a human expert for further review, then perhaps one might argue that it is better to err on the side of high recall and accept candidate lists of lower precision. As far as we can tell, however, this would not change our conclusion that the SCA method is superior in detecting loan words, since it is very clearly superior in recall.

Table 4 The distribution of pronunciation differences based on the three techniques examined. In addition, the outlier boundary (see text) is presented with a dashed line and the optimal threshold with respect to F score (see text) with a solid line. The x-axis is the pronunciation distance assigned by the algorithm, and the y-axis is the absolute frequency of the word pairs. The “bump” on the left of the SCA curve is promising.

| | Distribution of distance measures<br>Solid line – best threshold<br>Dashed line – low outlier |
|---|---|
| PMI (Frequency)-Based<br>Edit Distance |  |
| Spectrogram-Based<br>Edit Distance |  |
| Sound Class<br>Algorithm |  |

In addition to the comparison based on setting an optimal threshold, we likewise wished to test the algorithms performance on unseen data. This promises to provide a better view of how well the algorithms might work in a genuine research situation, where one could never know exhaustively what threshold to set.

In the cross-validation setup, we divide the entire data set into evenly sized subsets, using all but one of these to set the threshold, and then testing the procedures on the remaining data. In our case we applied 10-fold cross validation, which meant that we divided the data into ten subsets, first using nine of them (and therefore 90% of the data) to set the threshold parameter optimally. We note here that we did not continue the computationally expensive system of checking 200 candidate thresholds, but instead reduced the number of candidates to ten to keep running times manageable. We then tested the procedure on the remaining 10%. We repeated this ten times in order to avoid conclusions that might be based on a fortunate division into 90%/10%, and we report only the mean accuracy over all ten repetitions.

Table 5 Algorithm performance based on 10-fold cross validation.

| | PMI | Spectrogram | SCA |
|---|---|---|---|
| Precision | 0.88 | 0.73 | 0.88 |
| Recall | 0.66 | 0.75 | 0.80 |
| F score | 0.77 | 0.74 | 0.81 |

The size of the differences between SCA and the other two methods is lower in Table 5 than in Table 3, but SCA remains clearly superior.

We do not present the results that would be obtained using Tukey’s outlier heuristic, but these were never superior to the results obtained by searching for an optimal threshold. We provide some examples of loanwords detected by SCA but not by the other methods. We discuss this further in the concluding section.

Table 6 Loanwords detected by SCA but not by the other two measures of pronunciation difference.

| Concept | Turkic | Indo-Iranian |
|---|---|---|
| breast | kʊkrɛk | quqrak |
| correct | tuːra | tɔɣrɪ |
| fruit | miwe | mʲeva |
| sea | tʲeŋiz̥ | dzingɨz |
| tree | tʲerɛk | daraxt |

## Examples of detected loans

A substantial number of correctly detected loans involve the insertion or deletion of segments, presumably to satisfy phonotactic constraints. Examples are given in the table below (Table 6).

Table 7 Some loans detected that involve the insertion of deletion of a segment.

| Concept | Turkic | Indo-Iranian |
|---|---|---|
| correct | tuːra | tøɣri |
| forest | wʊrmon | urmʊn |
| meat | gwʉʃt | guʃt' |
| lake | kwøl' | kul |
| old | kwøn$^j$e | kʉjna |

A second large number of detected loans depended on certain sounds being aligned, for example the vowels, [a], [e], [ɔ], and [ɒ], as in 'fruit', Turkic [miβ**e**], Indo-Iranian [mev**a**] (where the corresponding vowels are printed in bold), or 'breast', Turkic [køkr$^j$**e**k], Indo-Iranian [kukr**a**k]. Another vowel set often exchanged was [i] ,[ɨ], [ɪ], and [ə], as in 'smooth' Turkic [sill**ə**q], Indo-Iranian [siłɫ**ɪ**q'] or 'fruit', Turkic [m$^j$**ɪ**va], Indo-Iranian [m**i**va]. Finally the following also often corresponded: [ə], [u], [ɜ], [ʉ],[ø], [ʊ], and [ɔ]; see examples such as 'egg', Turkic [tux**u**m] but also [tq**ʊ**m], Indo-Iranian [tx**ə**m]. Note that schwa [ə] appears in both of the last two lists.

We present some frequent consonant correspondences in Table 8. The sounds illustrating the correspondence are printed in bold. Although we have not tried to quantify their frequency, we suspect that it is the reason for SCA's superiority in this task.

Table 8 Some regular consonantal correspondences and illustrations from Turkic and Indo-Iranian.

| Consonant group | Swadesh concept | Turkic | Indo-Iranian |
|---|---|---|---|
| [k], [q], [g] | leaf<br>breast | para**k**<br>køkrɛ**k** | bar**g**<br>kukra**q** |
| [q], [x], [χ] | blood<br>back | **q**an<br>ɔr**q**a | **x**un<br>ar**χ**a |
| [t], [d] | tree<br>sea | **t**$^j$erɛk<br>'**t**$^j$eŋɪz̥ | **d**arax<br>**d**eŋiz̥ |
| [β], [v], [w], [f] | animal<br><br>to dig<br>fruit | aj**β**an<br>haj**β**an<br>ka**β**lɛm<br>mi**w**e | haj**v**ɔn<br>haj**w**on<br>kɔ**f**tan<br>m$^j$e**v**a |
| [ʧ],[ʃ], [ʤ], [j], [s] | dust<br>to live<br>star<br>bird<br>star | **ʃ**aŋ<br>**ʤ**aʃa<br>**ʤ**ɨldɨz̥<br>qʉ**s**<br>**ʃ**ʉldɨz̥ | **ʧ**aŋ<br>**j**aʃam<br>**j**ɨldɨz̥<br>ku**ʃ**<br>**j**ɨldɨz̥ |

## Discussion and prospects

We turn to the conclusions we draw and some further research this work suggests.

## Conclusions and a speculation

We can answer affirmatively that computational measures of pronunciation similarity can identify pairs of words in which a loan relationship is likely. Still, we also need to admit that F scores of between 0.75-0.80 also mean that further work will be required. Further, we conclude that SCA is superior to the other edit-distance based methods we examined, and that it is the best algorithm available for identifying loans.

This leads to the interesting question of why SCA performs so much better, which is also more difficult to answer. SCA aligns based on sound *classes*, a very rough basis for similarity, but it also incorporates asymmetric substitution weights, which is lacking in the other approaches. Given that loanword detection is a symmetric task in our operationalization, it is unlikely that using asymmetric substitution costs is the key advantage. SCA also takes prosody into account, and it adds detailed segmental information in obtaining the final measures of pronunciation difference (distances). Either of these, but also the reduction to sound classes might be the key to SCA's success. In choosing among these, we shall hazard to speculate about the reason for SCA's superiority in the loan word recognition task.

It is common to note that sounds (and sound sequences) unavailable in a borrowing language are replaced quite regularly in loan words. For example Spanish sequences involving stops followed by glides are commonly pronounced by native English speakers by stops followed by vowels. So *Buenos días* 'good day'[ b$^w$e.nos. dias] is often rendered by these native English speakers as [bu.en.os.di.as]. But this replacement is regular and can also be heard in English speakers' pronunciations of other words with stop-glide combinations, such as *igual* 'same, equal' [i.g$^w$al], pronounced as [i.gu.al] or *agua* 'water' [a.g$^w$a] as [a.gu.a], etc. It is plausible to assume that some of these replacements involve sounds in the same classes in the SCA algorithms, so that their alignment is likely to succeed. Table 6 (above) shows examples of where the SCA algorithm alone was able to detect the loan word, and these examples confirm the suspicion that attributing very low costs to potentially very different sounds can be an advantage. If, in addition, the segmental differences do not add much to the sequence distance greatly, this regularity of replacement will also explain the point of accumulation on the left side of the SCA frequency curve (Table 4). We conjecture then that SCA is more successful because it assigns very little (perhaps no) weight to such regular correspondences. In any case, our empirical study confirms that the overall measure detects loanwords more accurately. Naturally SCA assigns zero weight initially to the elements of various sound classes by design, and we do not mean to suggest that the observation is novel.

## Future work

It is easy to imagine alternative technical approaches, since no algorithm was able to detect some of the loans. Table 9 indicates where there is room for improvement.

Table 9 Examples of loanword pairs undetected by any algorithm.

| Concept | Turkic | Indo-Iranian |
|---|---|---|
| fire | ɔlɔv | aɣaw |
| old | kwønʲɪ | kʉhna |
| egg | tqʊm | tuxm̩ |
| meat | etʲ | jota |
| dust | ʃaq | ʧank |

One promising avenue for further research would seek to incorporate more information in the comparison. Since the approach in this paper compares word pairs one pair at a time, we fail to exploit all the information in the data set. We might therefore try to compare not just one word pair at a time, but rather examine an aggregate measure of pronunciation difference. We might then compare the *mean* difference of the putative loan word to all the words in the sample, both those in its own language family and those in its putative source family. To be more precise, we might represent a given pronunciation j of a concept c in a Turkic language as $t_j$, suppressing the reference to c, which will be the same in any comparison of word pronunciations. Given a set of pronunciations of a given concept c from Turkic languages $\{t_1, t_2, \ldots, t_n\}$ and a second from Indo-Iranian $\{i_1, i_2, \ldots, i_m\}$, we compare $t_j$ both to all the other pronunciations in Turkic $t_{j'}$ as well as to all the pronunciations in Indo-Iranian $i_{k.}$ In this way we obtain the mean distance of $t_j$ to all other Turkic pronunciations of the same concept, as well as the mean distance to all Indo-Iranian pronunciations, i.e.

$$D_{own} = \overline{d(\,t_j, i_{j'})}, \forall j' \neq j \text{ and } D_{other} = \overline{d(\,t_j, i_k)}\ \forall k$$

It is clear that we should combine these somehow, but also that we should be particularly interested in $t_j$’s, for which $D_{own}$ is large and $D_{other}$ small.

## Prospects

In the rapidly evolving societies of our globalized times, social interaction – including interaction among speakers of different languages – is increasing apace, which will likely result in loanwords becoming more frequent as well. This should delight students of contact linguistics, as it should provide the larger amounts of data needed to discern the patterns of new loanwords, including what sorts of concepts are involved, the influence of the degree and nature of the contact on the likelihood of success, and the semantic and phonetic deformation that is involved. Mufwene has been among the first linguists to frame this phenomenon in terms of the *population structure* of the speakers involved (Collins & Mufwene 2005), bringing to contact linguistics a wider social and cultural dimension that profits from the analogy to a discipline proceeding through quantitative, mathematical models: ecology. Given the complexity of the contact situation, including the languages involved, but also the familial, material, social, economic, and cultural relations among the speakers, it is only sensible to explore these situations using quantitative techniques.

## Acknowledgements

We are grateful to Philippe Mennecier for the use of the Central Asian data, and to Wilbert Heeringa and Martijn Wieling for the use of their software for spectrogram-based and PMI-based edit distance measures. Wilbert Heeringa generously criticized an earlier draft of this paper quite extensively, and the paper has improved on the basis of his suggestions. We thank also an anonymous referee, who was unusually helpful. John Nerbonne is grateful to the DFG Science Center "Words, Genes, Bones and Tools" at the University of Tübingen for support while working on this paper.

## Supplementary data

Resources related to this work, such as data and code, are available at https://github.com/jayliqinzhang/computational-loanword-detection.

## References

Black, E., Lafferty, J., & Roukos, S. 1992. Development and evaluation of a broad-coverage probabilistic grammar of English-language computer manuals. In: *Proc. 30th ACL Mtg.* Association for Computational Linguistics: Shroudsburg, PA. 185-192. Avail. at https://www.aclweb.org/anthology/

Collins, M. & Mufwene, S.S. 2005. What We Mean When We Say 'Creole': An Interview with Salikoko S. Mufwene. *Callaloo* 28 (2): 425-462. Avail. at https://www.jstor.org/stable/3805668

Dellert. J. 2017. *Information-Theoretic Causal Inference of Lexical Flow*. PhD dissertation, Eberhard-Karls-Universität Tubingen.

Dellert, J. 2018. Combining Information-Weighted Sequence Alignment and Sound Correspondence Models for Improved Cognate Detection. *Proc. 27th COLING*. 3123-3133. Avail. at https://www.aclweb.org/anthology/

Delz, M. 2013. A theoretical approach to automatic loanword detection. Master thesis, Eberhard-Karls-Universität Tübingen.

Dolgopolsky, A. B. (1964). Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točky zrenija [A probabilistic hypothesis concerning the oldest relationships among the language families of Northern Eurasia]. *Voprosy jazykoznanija*, *2*, 53-63.

Heeringa, W. 2004. Measuring Dialect Pronunciation Differences using Levenshtein Distance. PhD Dissertation, University of Groningen.

International Phonetic Association (IPA) 1999. *Handbook of the International Phonetic Association*. Cambridge University Press: London.

Jobling, M. A., Hurles, M.E. & Tyler-Smith, C. 2004. *Human Evolutionary Genetics: Origins, Peoples and Diseases*. Garland: New York.

Köllner, M. & Dellert, J. 2016. Ancestral state reconstruction and loanword detection. Universitätsbibliothek Tübingen.

Kondrak, G., Marcu, D. & Knight, K. 2003. Cognates can improve statistical translation models. In *Proc. 2003 North American ACL. Vol. 2*. Association for Computational Linguistics: Shroudsburg, PA. 46-48. Avail. at https://www.aclweb.org/anthology/

Kondrak, G. & Sherif, T. 2006. Evaluation of several phonetic similarity algorithms on the task of cognate identification. In J.Nerbonne & E.Hinrichs (eds.) *Proceedings of the Workshop on Linguistic Distances*. Association for Computational Linguistics: Shroudsburg, PA. 43-50. Avail. at https://www.aclweb.org/anthology/

List, J.-M. 2012. SCA: phonetic alignment based on sound classes. In: *New Directions in Logic, Language and Computation*. LNCS 7415: 32-51. Springer: Berlin Heidelberg.

List, J.-M., Walworth, M., Greenhill, S. J., Tresoldi, T., & Forkel, R. 2018. Sequence comparison in computational historical linguistics. *Journal of Language Evolution*, *3*(2): 130-144.

Manni, F. 2017. Linguistic probes into human history. PhD Dissertation, University of Groningen.

Manni, F., Heeringa, W., & Nerbonne, J. 2006. To what extent are surnames words? Comparing geographic patterns of surname and dialect variation in the Netherlands. *Literary and Linguistic Computing 21*(4): 507-527.

Mayr, E. 2002. Interview with Ernst Mayr. *BioEssays* 24(10): 960-973. Avail. at http://www.stephenjaygould.org/library/bioessays_mayr.pdf

Mennecier, P., Nerbonne, J., Heyer, E., & Manni, F. 2016. A Central Asian Language Survey. Collecting data, measuring relatedness and detecting loans. *Language Dynamics and Change* 6(1):57-98.

Mufwene, S. S. 2001. *The ecology of language evolution*. Cambridge University Press: London.

Mufwene, S. S. 2005. Language evolution: The population genetics way. In G. Hauska (ed.) *Gene, Sprachen, und ihre Evolution*, Schriftenreihe der Universität Regensburg 29. University of Regensburg. 30-52.

Mufwene, S. S. 2008. *Language evolution: Contact, competition and change*. Bloomsbury Publishing: London.

Nerbonne, J. 2003. Linguistic variation and computation. In *Proc. 10th European ACL* . 3-10. Association for Computational Linguistics, Shroudsburg, PA. Avail. at https://www.aclweb.org/anthology/

Nerbonne, J. 2009. Data−driven dialectology. *Language and Linguistics Compass 3*(1): 175-198.

Nerbonne, J., & Siedle, Ch. 2005. Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik*, 72(2): 129-147.

Shackleton, R. G., Jr. 2010. Quantitative assessment of English-American speech relationships. *Groningen Dissertations in Linguistics. 81*. CLCG: Groningen

Sperber, D. 1996. *La contagion des idées*. Odile Jacob: Paris.

Swadesh, M. 1952. "Lexicostatistic Dating of Prehistoric Ethnic Contacts." *Proceedings of the American Philosophical Society*, 96: 452–463.

Tukey, J. W. 1977. *Exploratory data analysis.* Addison-Wesley: Reading, MA.

Wieling, M., Leinonen, T., & Nerbonne, J. 2007. Inducing sound segment differences using pair hidden Markov models. In *Proc. 9th Meeting, ACL Special Interest Group in Comp. Morphology and Phonology*. 48-56. Association for Computational Linguistics. Shroudsburg, PA. Avail. at https://www.aclweb.org/anthology/

Wieling, M., Margaretha, E. & Nerbonne, J. 2012. Inducing a Measure of Phonetic Similarity from Pronunciation Variation. *Journal of Phonetics* 40(2): 307-314.

Zhang, L. 2016. A More Sensitive Edit-Distance for Measuring Pronunciation Distances and Detecting Loanwords. MSc thesis, Groningen and Malta.

[i] It's become good practice to identify the contributions of individuals in multi-authored papers. Liqin Zhang wrote his master's thesis on this topic and was responsible for the code (except where otherwise noted), the evaluation, and the first description of the work (his Master's thesis in Language and Communication Technology in Groningen and Malta – Zhang 2016). All the graphs came from his thesis. John Nerbonne designed the project, supervised the thesis, and wrote the first version of this paper, while Ray Fabri was second supervisor. Franz Manni designed the linguistic and genetic research in Central Asia.

[ii] Indeed, evolutionary theorists have often emphasized that this view is too simplistic. Ernst Mayr emphasizes how sophisticated species have sometimes been lost and how many more primitive forms of life exist than sophisticated ones (Mayr 2002).

[iii] See http://www.blackwellpublishing.com/ridley/a-z/Founder_effect.asp for an explanation with animation.

[iv] See Mennecier et al. (2016), Language Dynamics and Change 6(1), Supplementary Materials, Table S2. http://booksandjournals.brillonline.com/content/journals/10.1163/22105832-00601015 (May 24, 2018). or Github: https://github.com/jayliqinzhang/computational-loanword-detection

[v] LingPy is available at http://lingpy.org, and is introduced at http://github.com/lingpy/lingpy-tutorial.