Preliminary Identification of Language Groups and Loan Words in Central Asia

René van der Ark,* Philippe Mennecier,† John Nerbonne,* & Franz Manni†
*University of Groningen †Musée de l'Homme
Groningen, The Netherlands Paris, France
r.van.der.ark@student.rug.nl

Abstract

We use Levenshtein distance to classify language groups as opposed to dialects. If successful, this technique could be usefully applied in the preliminary analysis of linguists' field notes. We expect this classification task to be easier than dialect classification. We also suggest using Levenshtein distance for identifying loan words in different language groups of a same region.

Keywords

Levenshtein, Dialectology, Loan Words

1 Introduction

This paper aims to extend the applicability of a range of computational techniques that have been implemented and successfully applied to problems in dialectology. The dialectological work has made prominent use of Levenshtein distance (also known as edit distance) to analyze the relations among the various varieties of a range of European languages, including Irish, Dutch, Norwegian, Sardinian, German, American English, and Bulgarian [2, 11].

This paper focuses on an area in Central Asia including the former Soviet states Uzbekistan, Kirgizstan and Tadjikistan, and distinguishes itself from the other dialectological research by analyzing data from more than one language group. Levenshtein distance works well to distinguish dialects, but we investigate here whether it work equally well to distinguish separate language groups. If Levenshtein distance can be used successfully in such cases, it could be useful in the preliminary analysis of linguists' field notes. Since we expect this classification task to be easier than dialect classification, we expect this application of Levenshtein distance to be successful.

Besides researching a novel application of the Levenshtein measure, the present paper will examine the problem of identifying loan words in the same linguists' field notes. Three language groups (Turkish, Tadjik and Yagnobi) exist alongside each other and are somewhat diversely spread across the region, so we expect to find loanwords in all three language groups. We expect Levenshtein analysis to identify some words in pairs of varieties that are much more similar to one another than the overall similarity of the varieties would lead us to expect.

Finally, we examine data based on both the Swadesh list of 100 common concepts and the list based on 200 common concepts. There has been criticism of the use of the larger list, but we wished to check empirically whether the indications of the two would coincide [5].

2 Data Description

The data were collected by Philippe Mennecier for *Musée de l'Homme* in a larger project aimed at investigating the association of genetic and linguistic indicators of relatedness in Central Asia. Figure 1 shows the location of the area under investigation. There are three main language groups present: Turkic, Tadjik and Yagnobi. Within the Turkic group Kirgiz, Uzbek and Kazakh subgroups are represented in the data. Philippe Mennecier indicated which (sub-)group the speech of each of respondents appeared to belong to. His tentative classification provides a good means for testing the software developed at the University of Groningen mentioned in the next section. Besides identifying the varieties (languages and dialects) of each speaker, Mennecier also indicated the lexical identity of the different pronunciations of the concepts from the Swadesh list, classifying every pronunciation with a single letter. These annotations include a conjecture about which pronunciations are loanwords, and can therefore be used to test the effectiveness of automatic loanword detection.

The data consist of a set of 180 spoken words for each of the 78 respondents, distributed across 23 test-sites in the region. The words were recorded acoustically and transcribed in IPA which was subsequently recoded as X-SAMPA. While Tadjik samples hail from Tadjikistan, the Turkic samples come from sites spread across Uzbekistan and Kirgizstan. The words are mostly members of the Swadesh-group of common basic words, 84 of them belonging to the Swadesh-100 group and 163 of them belonging to the Swadesh-200 group. In Table 1 we present an excerpt from two data lists. We present both the X-SAMPA and the IPA transcriptions.

3 Methods of Analysis

We compare pronunciations using the Levenshtein algorithm, also known as edit distance. When calculating edit distance between a pair of words in two

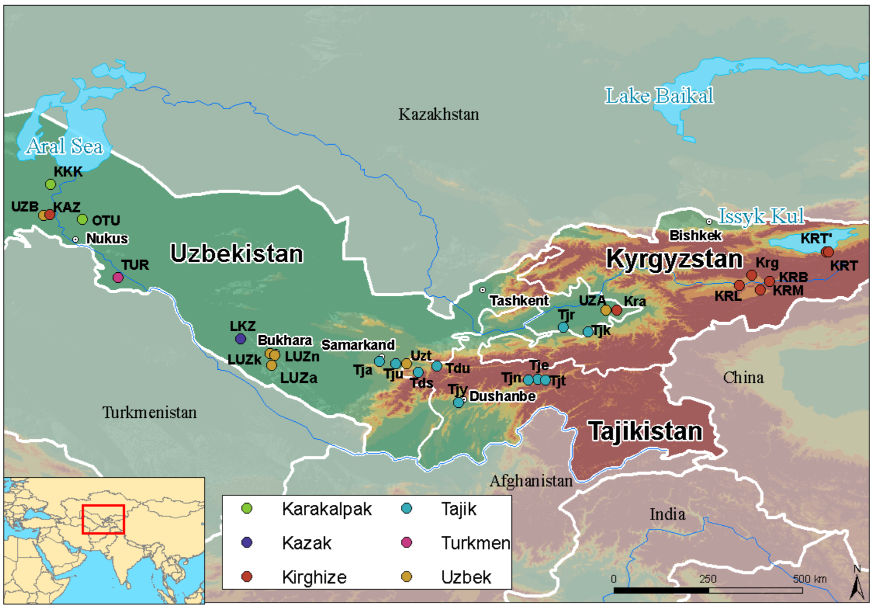

Fig. 1: *Map of the area investigated. Note that the linguistic data of some locations in Kirgizstan are not yet included in the current dataset.*

English	X-SAMPA	IPA	X-SAMPA	IPA
one	bIr	bɪr	i:	iː
two	jIk@	jɪkə	dU:	dʊː
three	yt_S	yt͡ʃ	saraj	saraj
four	t8rt	tɵrt	safOr	safɔr
five	b_jIS	bʲɪʃ	panZ	panʒ
big	}lk@n	ʉlkən	kalOn	kalɔn
long	uzaq	uzaq	darOz	darɔz
wide	k_jeN	kʲeŋ	kuSOd	kuʃɔd
thick	s_jemIz_O	sʲemɪz̥	Gafs	ɣafs
heavy	awIr	awɪr	vazmin	vazmin

Table 1: *Some example data in X-SAMPA and IPA*

different varieties—whether they be dialects or as in the present case, languages—we seek the minimal set of operations that can be used to transform one pronunciation into another. The operations can be insertions, deletions, substitutions or swaps and each is associated with a cost. Although we have experimented with elaborate cost schemes, we have in general found simple schemes to function effectively when the purpose is to characterize the overall similarity among varieties. In this research, therefore, a standard cost scheme is adopted for Levenshtein measurement in which all operations cost a single unit (1.0). Heeringa (2004) presents the application of Levenshtein distance in great detail [2]. We ensure (roughly) that only vowels substitute for vowels, and consonants for consonants, and the distance scores are normalized according to word-length (see Heeringa et al. [3] for details).

As in previous dialectometric research conducted by the Groningen group, the software package L04, developed by Peter Kleiweg, is adopted for analysis (`http://www.let.rug.nl/~kleiweg/L04`). This package contains several methods to analyze phonological and lexical data statistically, building on the Levenshtein measure. The focus is on the comparison of pronunciation data such as IPA transcriptions.

For the purpose of testing effectiveness in distinguishing language groups, L04 is used to calculate a distance between each pair of words, and then an aggregate distance score for each pair of sites (the mean of word distances). We collect the aggregate site distances into a site × site matrix which is further analyzed using multi-dimensional scaling and clustering (both further explained below). The results of these methods can also be used to automatically generate regional dialect maps. In the present paper, we focus on the degree to which the aggregate distances between sites agrees with Mennecier's pre-classification. We evaluate this by comparing the mean distance between varieties of the same language (family) with the mean distance between varieties from different language (families).

Note that for each concept in the Swadesh list, and for each pair of sites we obtain a pronunciation distance—the distance between the pronunciation at the one site from the pronunciation at the other. We use these single word distances to detect likely loan words, under the leading hypothesis that words from unrelated language families that are very similar are likely to be related as loan words (meaning that the word was borrowed from the one language into another or from a third language into each of languages under comparison). We can evaluate this hypothesis by comparing pronunciation distances of the words with Mennecier's conjecture about whether the words are related as loans (noted above). We quantify the success of recognizing loan words using precision and recall. Finally we investigate whether the words related by loan as a set differ from other words (whether the mean pronunciation differences differ significantly), and we investigate analyzing the two sets as a mix of distributions, using the EM algorithm [10], implemented in the 'mixdist'-package in R (`http://www.r-project.org/`).

3.1 Analysis of Aggregate Distances

3.1.1 Multidimensional Scaling

Multidimensional scaling (MDS) is a technique to reduce a site × site distance matrix to a matrix in which each site is represented not by the distances to each of the other sites, but rather by a set of coordinates in limited set of dimensions. In this research both two-dimensional and three-dimensional scaling are applied. MDS provides good means of visualizing differences between dialects insightfully, using two- or three-dimensional scatterplots. Three dimensions can also be used to automatically generate a color-scheme which can be used in a regional dialect map. Although there are several types of MDS, we use exclusively the classical, so-called metric variant. For more detailed description of the algorithms for applying MDS, we refer to Heeringa (pp. 156–163) [2].

We apply MDS in this paper to verify that the different language groups are indeed separated well using the pronunciation difference measure based on Levenshtein distance. We compare the MDS results with the language groups Mennecier identified, in particular zooming in on the Turkic subgroups. The effect of using only words from the Swadesh 100- and 200-group are also compared to the full set of words.

3.1.2 Clustering

Hierarchical agglomerative clustering searches a site *times* site distance matrix for pairs of sites that are minimally distant. These sites are then "fused" to obtain a smaller distance matrix, in which the distances from the fused sites to others are determined by averaging the distances in the input matrix (we omit complications introduced by alternatives to averaging). The process is repeated until all the sites are fused into a DENDROGRAM—or, tree—in which fused items are joined under a single node. Branch length in dendrograms corresponds to the distance between items at the moment of fusion. So, whereas MDS locates varieties within an idealized, low-dimensional space, clustering identifies groups among different varieties. For this paper UPGMA (Unweighted Pair Group Method using Arithmetic averages) clustering was used, following Heeringa (pp. 146-156) [2].

We shall employ clustering to verify that distinct language groups are indeed distinguished via pronunciation distance.

3.2 Identifying Loanwords

3.2.1 Pairwise analysis

To identify loanwords the data need to be analyzed at the word level. For this purpose, words were compared pairwise. For 78 test-sites and 180 words this generated 531.531 pairs. If we ignore pairs for which a score is not available—cases where a pronunciation was not recorded—we are left with 500.485 pairs.

Based on the regional pre-classification the entire set of word pairs was divided into two subsets, the first subset containing word-pair scores from the same language group, and the second containing scores for word pairs in different groups. Thus contained scores of word pair differences for word pairs from the following pairs of languages (language families): Tadjik-Tadjik, Yagnobi-Yagnobi, and Turkic-Turkic. The second subset contains the scores for word pairs from different language groups, which results in scores for word pairs from the following pairs: Tadjik-Yagnobi, Tadjik-Turkic, and Turkic-Yagnobi. The 'same-family-set' contains 221.357 scores and the 'different-family-set' contains 279.128 scores.

The same is also done for the Turkic subgroups, resulting in a full set of 127.020 scores, a 'same-family-set' of 33.351 scores and a 'different-family-set' of 93.669 scores.

3.2.2 Testing Pre-classified Loanwords

For the task of automatic loanword detection we focus here on the (admittedly easier) problem of detecting loan words within the 'different-family-set'. It is expected that zero- and low edit-distances in this set indicate loanwords. We test this assumption by calculating precision and recall measures, using the classification provided by Mennecier. If, for instance, a pronunciation in site X is classified as being from family A and a pronunciation in site Y is also classified as being from family A, then it has been pre-classified as a loanword. Recall is the percentage of word pairs pre-classified as being from the same family which are indeed recognized as loanword (i.e. by having a low score for edit-distance). Precision is the percentage of the pairs identified as loanwords on the basis of low edit distance scores which were also pre-classified as loanwords.

3.2.3 Mix of Distributions

We are also interested in looking at the distribution of the entire set of pairs as a mix of distributions of the two subsets noted above. Before doing this, we wish to verify that the two subsets are indeed statistically distinct. A simple t-test will be used for this purpose. If these subsets are indeed distinct, the next step is to analyze the full set of pairs as a mixed distribution of these two subsets. To address this issue we apply the 'mixdist'-package for R (`http://www.r-project.org/`), written by dr. Peter MacDonald (McMaster University, Ontario) [4].

We are also interested in identifying loans within related languages. For this reason we also examine the set of Turkic languages and analyze them as a mix of distributions.

4 Results and Evaluation

4.1 Preliminary remarks

Distance matrices were generated for the entire data collection as well as for the following two subsets: all

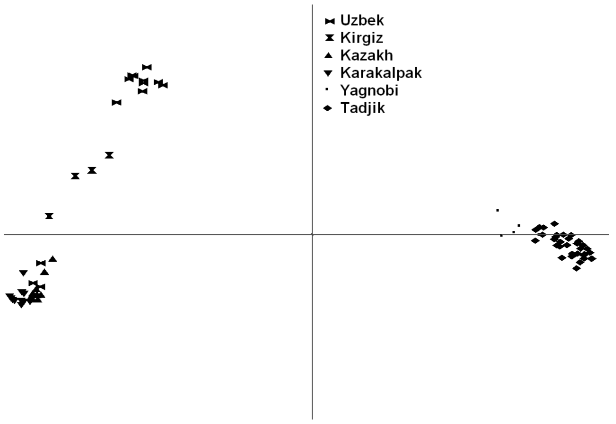

Fig. 2: *Scatterplot of the two-dimensional reduction of distances among all the sample sites (classical, metric MDS).*

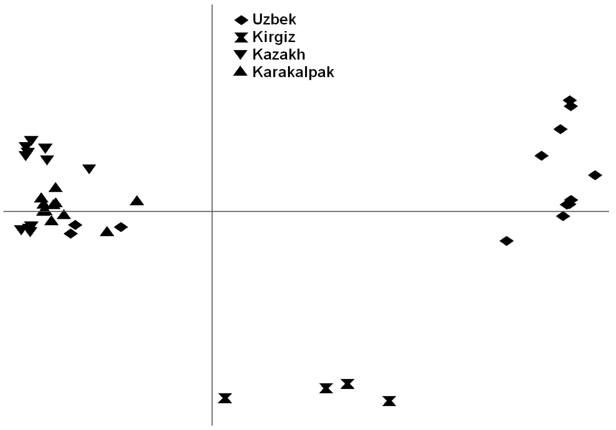

Fig. 3: *Scatterplot of (classical) two-dimensional MDS for the sample restricted to the Turkic languages. Notice how the Hitoj (represented by the diamonds on the left, among the triangles and inverted triangles) appears in the Kazakh/Karakalpak cloud, even though Hitoj respondents were conjectured to speaking a variety of Uzbek.*

the Turkic sites on the one hand and all the Tadjik and Yagnobi sites on the other. Consistency measures (Cronbach's α) of the data range from 0.955 (Tadjik and Yagnobi sites) to 0.993 (full set), confirming the strong signal in the data. When eliminating words in the Swadesh 100 group the scores range from 0.921 (Tadjik and Yagnobi) to 0.989 (all sites). Because the varieties of different languages are somewhat interspersed, dialect maps are limited in utility. For that reason we concentrate on visualizations using dendrograms and MDS scatterplots, to which we now turn.

4.1.1 Multidimensional Scaling

Two-dimensional MDS analysis for the full dataset correlated with the input data correlated very highly ($r = 0.963$), indicating that the data is represented very well in two dimensions. Figure 2 presents the MDS results for the entire set of data in the 2-dimensional reduction. Although we have experimented with different variants of MDS (e.g., distinguished by whether the approximation algorithm seeks to minimize linear error or squared error), we restrict our attention here to classical MDS in which squared error is minimized.

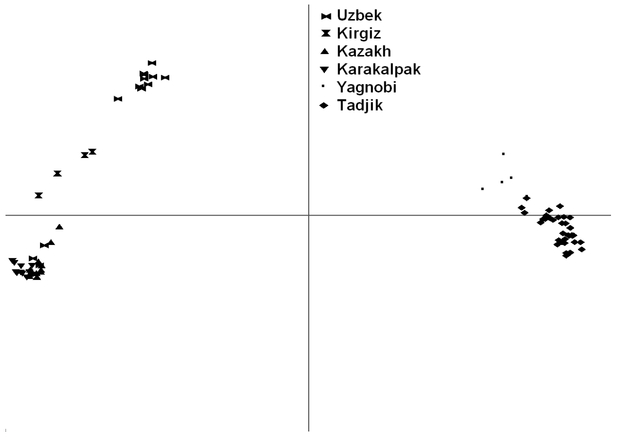

Fig. 4: *Scatterplots of 2-dimensional MDS for the 100-word Swadesh list. If one compares this plot to Fig. 2, the overall similarity is striking, but the location of the Yagnobi group has shifted in this reduction based on the 100-wd. Swadesh list.*

Figure 3 focuses on the Turkic varieties. Especially interesting here is the position of three test persons of (Hitoj), represented by diamonds, meaning that they were pre-classified as belonging to the Uzbek dialect. But they are closer to the Kazakh/Karakalpak area in this standard MDS analysis. One explanation might be that the site is situated in a region where more Kazakh and Karakalpak sites are located (see Figure 1), suggesting that we may be seeing the influence of contact. This is a subject for further analysis.

We also compare MDS analyses to gauge the impact of the tow different Swadesh lists. As mentioned above, we have worked with both a full set of words, but we can compare the results based on this full 200-wd. set to results based on subsets containing only words from the Swadesh-100 list [5, 9]. The MDS results for the 100-element list may be seen in Figure 4. This plot should be compared to the earlier Fig. 2. The comparison displays almost no differences. Nonetheless, the Yagnobi sites *do* appear to stand further away from the Tadjik sites when the analysis is based on the Swadesh-100 subset. Since other scholars have argued that the more restricted list is more likely to reflect older linguistic relations [5, 9], we speculate that the difference in the relative position of Yagnobi may be a reflection of contact.

It remains to be explored which of the word-sets provides the most reliable results. Generally speaking, larger sets of data tend to be more reliable, arguing for the larger word-sets, but as we noted above the 100-wd. provides a very strong signal. We conjecture that the detection of loan words should provide further insight in this issue. If the Yagnobi 200-wd. set indeed contains loan words, this could explain its shifted position.

4.1.2 Clustering

Fig. 5 shows the result of applying UPGMA clustering (explicated above) to the full set of data for all data collection sites. As one may verify, the major distinction is between the Turkic languages in the top half of the diagram and the Indo-Iranian languages in the bottom half. The major split among the Turkic languages is between Uzbek varieties (in the lower part of the top section) and the others (Karakalpak, Kazakh, and Kirgiz), but the dendrogram confirms the exceptional status of the (Uzbek) Hitoj varieties. Among the Indo-Iranian varieties (in the lower half of the dendrogram), we see the split between Tadjik and and Yagnobi which we also noted in the MDS analysis.

We expect little insight from projecting the classifications to geography due to the mobility of the peoples in this area, so we refrain from present maps here. We turn therefore to the problem of detecting loanwords.

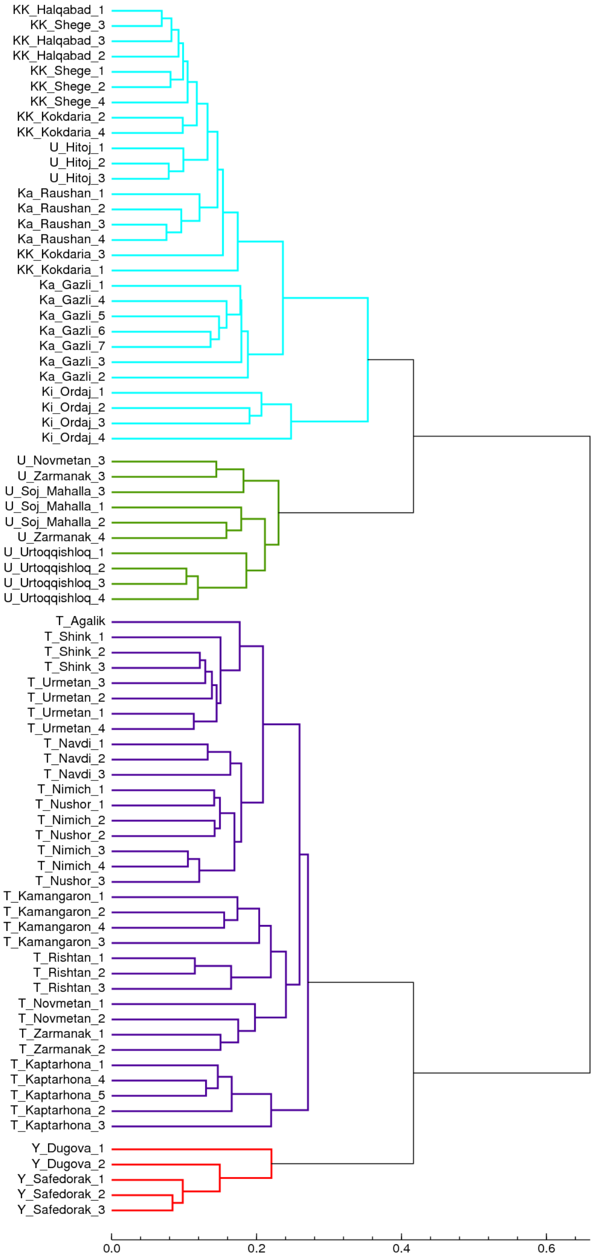

Fig. 5: *UPGMA dendrogram based on all words.*

4.2 Loanword Identification

As mentioned above, we are also interested in the problem of detecting loanwords automatically or semi-automatically. We proceed from the assumption that the pronunciation variants of a word involving a loan may be recognized by a very small (near zero) pronunciation distance. This will be easier if the loan effect is to be detected in unrelated languages.

We note that there is a substantial literature on the detection of translation equivalents, or “cognates” for the purpose of translation [12, 7, 1], many of which are in fact loanwords. We do not take this up here for two reasons. First, we would like to push the approach of relying only on surface similarity, and second, the techniques often rely on having large samples of the languages in which “cognates” are sought and/or bilingual corpora which also provide semantic clues. We note that a detailed comparison with at least some of these techniques will be sensible, in particular Kondrak’s work [7, 6] but we postpone it to future work.

We have attempted to detect loanwords both between the three major groups (Turkic, Tadjik and Yagnobi) and between the Turkic subgroups—as might be expected, the former more successfully than the latter.

4.2.1 Testing Pre-classified Loanwords

To test whether the assumption of (near-)zero edit-distance for loan-related words in different language groups, a histogram of the ’different-family-set’ of distances has been created which can be seen in figure 6. This histogram suggests a normal distribution of pronunciation differences, except that there are very many zero-scores and near-zero scores. We assume that these are words related by loan.

To further examine this assumption, a recall/precision analysis has been adapted to suit our data. We compare the near-zero pronunciation distances as our hypothesis about loans with the hand annotations supplied by Mennecier. Let’s first compare this use of precision and recall to the use of those measures in information retrieval, where documents are automatically ranked according to their relevance to a given query, and precision and recall quantify how

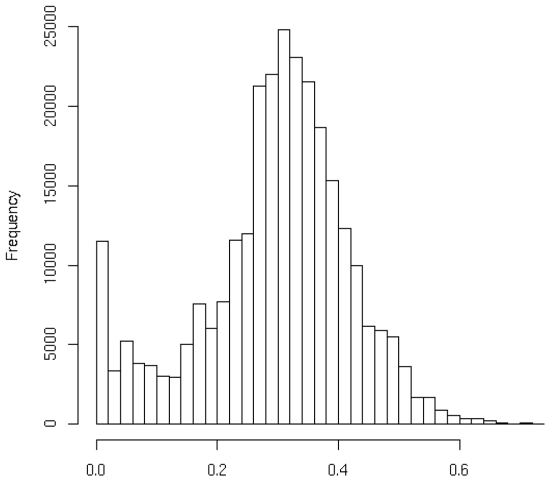

Fig. 6: *Histogram of frequencies of Levenshtein-distance scores for word pairs from different language families. Note the anomalous "bump" near zero disturbing the otherwise normal distribution. We hypothesize that these are the words related by loan.*

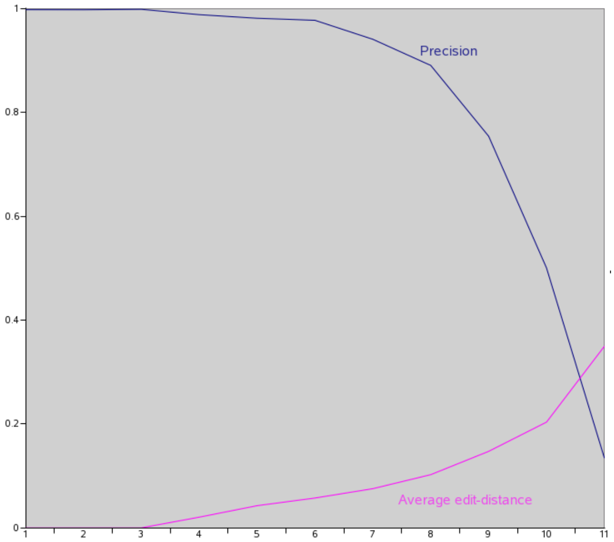

Fig. 7: *Averaged 11-point recall/precision curve for all language groups, combined with average edit distance score per 10th percentile of recall.*

well this overlaps with hand annotations provided by expert users. Our automatic technique is the near-zero score of pronunciation distance, and we shall make use of AVERAGED 11-PT. PRECISION-RECALL CURVE [8].

To sketch this curve, the word pairs are measured for their pronunciation distance using the Levenshtein measure, and then sorted in order of increasing distance, zero scores first. RECALL is then the percentage of words which were hand-picked to be related by loan and that have been "retrieved" at a certain proportion of the sorted result set. PRECISION at a given recall-percentile is the percentage of pairs pre-classified as loans that are found in the data sorted in increasing order of pronunciation distance. We divide the entire set to detected into eleven sections, and record the relation between recall and precision in each of the sections. It is worth emphasizing that the point at which recall is 100% need not contain the entire data set, and in fact normally does not. It is the point at which all the true loan relations have been detected.

In Figure 7 the x-axis is calibrated according to the 11 percentile bins of recall effectiveness. The falling line (which begins at 1.0) indicates the precision of detection calculated on the basis of the words sorted according to increasing edit distance. The lower, rising line is the average Levenshtein-distance for all the pairs within a given recall-percentile bin. The choice of a threshold for detecting loan words should be based on the corresponding precision-score at the same recall-percentile. A reasonable assumption might be to use a score before the sharp drop in precision, for instance at the 70th percentile, where precision is 0.754, and the average normalized edit-distance is 0.147. If we used this as a threshold, we would of course accept nearly 25% error. If this is regarded as too "noisy", we might instead back off to a threshold at the 50th percentile, where precision is over 0.95, and error less than 0.05.

We can repeat this analysis within the Turkic subgroup, expecting that detection will be more difficult since there will be more scores near zero due to the cognacy of the words under comparison. In the interest of space, we do not do present those results in detail here.

4.2.2 Mix of Distributions

We also investigate whether well-known techniques for detecting the mixture of two distributions might help in interpreting the data. This section is speculative. Since these require that one specify the sorts of component distributions involved, and since the basic distribution appeared to be Gaussian, we first check whether the word distributions are distributed normally. We restrict our attention here to the full set of pairs from all of the languages, without distinguishing whether the languages are known to be genealogically related or not.

We examine the hypothesis that the distribution is normal by sketching a normal quantile plot (see Fig. 8), which indeed suggests that a normal distribution is being mixed with a uniform distribution around zero. We likewise see a "ceiling" effect for the most different word pairs, suggesting a further deviation from the normal distribution. We add that the same analysis applied to the 'same-family-set' suggests a more complicated situation whose analysis we postpone to a later paper.

We continue to the analysis of the distribution of full set of edit-distances as a mixture of two Gaussian distributions. This may not be unproblematically applicable, but it may nonetheless prove insightful. The result can be observed in Figure 9. The histogram shows a drop in frequency after the (near-zero) region where we expect to find the words related by loan. The left component (lower curve on the left) is the normal curve inferred by the analysis, which we should prefer to interpret as the words related by loan. The lower

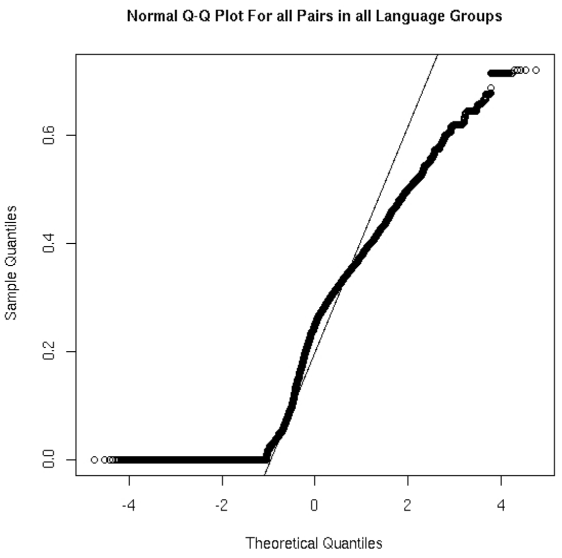

Fig. 8: *Normal quantile plot for full set of pairwise edit-distances, including both the 'different-family-set' and the same-family-set'.*

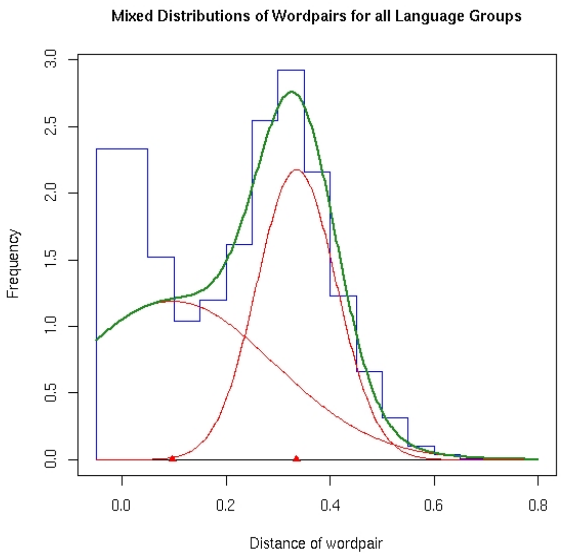

Fig. 9: *Mixed distributions plot for all language groups. Although the two components that are inferred seem reasonable, they also overlap a good deal.*

curve on right should then reflect the normal sort of variation one finds in sets of words not related by loan. The disappointing feature of the mixture is the large intersection of the two component curves. The size of the area under both component curves reflects the number of words that might be due to either process—either loan or non-loan variation.

We conclude only that this sort of analysis is worth further pursuit, especially in conjunction with the examination of the 'different-family-set' and including the use of mixture analyses that attempt to infer non-Gaussian components.

5 Conclusions and Discussion

This paper aimed to extend the applicability of a range of computational techniques that have proven successful in dialectology. We focused on an area in Central Asia including the former Soviet states Uzbekistan, Kirgizstan and Tadjikistan, and includes both Turkic and Indo-Iranian languages. As expected, the pronunciation distance techniques perform well in the preliminary classification of varieties even when the dataset includes unrelated varieties. We conclude therefore that the technique could be useful in the preliminary analysis of linguists' field notes.

Besides researching a novel application of the Levenshtein measure, the paper also examined the problem of detecting loan words in the same linguists' field notes. Three language groups (Turkish, Tadjik and Yagnobi) exist alongside each other and are somewhat diversely spread across the region, so we expected to find loanwords in all three language groups. Levenshtein analysis was able to identify words related by loan, but not perfectly: there is a recall/precision tradeoff that needs to be kept in mind in pairs of varieties that are much more similar to one another than the overall similarity of the varieties would lead us to expect. In a speculative section we examined the possibility of determining two components of the overall distribution of pronunciation distances, one due to loans, and the other due to "normal" variation. Two components can indeed be identified, but they overlap a good deal so that they do not provide much more help in distinguish loanwords from others.

Finally, we examine data based on both the Swadesh list of 100 common concepts and the list based on 200 common concepts. There has been criticism of the use of the larger list, but we wished to check empirically whether the indications of the two would coincide [5].

References

[1] O. M. Frunză. *Automatic Identification of Cognates, False Friends, and Partial Cognates.* PhD thesis, University of Ottawa, 2006.

[2] W. Heeringa. *Measuring Dialect Pronunciation Differences using Levenshtein Distance.* PhD thesis, University of Groningen, 2004.

[3] W. Heeringa, P. Kleiweg, C. Gooskens, and J. Nerbonne. Evaluation of string distance algorithms for dialectology. In J. Nerbonne and E. Hinrichs, editors, *Linguistic Distances*, pages 51–62, Shroudsburg, PA, 2006. ACL. Proc. of a workshop held at the joint meeting of ACL and COLING, Sydney, July, 2006.

[4] J. Ju. Combined algorithms for constrained estimation of finite mixture distributions with grouped and conditional data. Master's thesis, McMaster University, Ontario, Canada, 2002.

[5] B. Kessler. *The Significance of Word Lists.* CSLI Press, Stanford, 2001.

[6] G. Kondrak and T. Sherif. Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification. In *Proceedings of the Workshop on Linguistic Distances*, pages 43–50, Shroudsburg, PA, 2006. ACL. Proc. of a workshop held at the joint meeting of ACL and COLING, Sydney, July, 2006.

[7] W. Mackay and G. Kondrak. Comparing word similarity and identifying cognates with pair Hidden Markov Models. In *Proceedings of the 9th Conference on Natural Language Learning (CoNLL)*, pages 40–47, Shroudsburg, PA, 2005. ACL.

[8] C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. CUP, Cambridge, 2008.

[9] A. McMahon and R. McMahon. *Language Classification by Numbers*. Oxford University Press, Oxford, 2005.

[10] T. Mitchell. *Machine Learning*. McGraw-Hill, New York, 9th pr edition, 1997.

[11] P. Osenova, W. Heeringa, and J. Nerbonne. A quantative analysis of Bulgarian dialect pronunciation. *Zeitschrift für slavische Philologie*, accpt'd, 2007.

[12] J. Tiedemann. Extraction of translation equivalents from parallel corpora. In *Proceedings of the 11th Nordic Conference on Computational Linguistics (NODALI98)*, University of Copenhagen, 1998. Center for Sprogteknologi.