41808: Issues in typology: Defining language families by phonemes

Leshem Choshen
[1]School of Computer Science and Engineering, [2] Department of Cognitive Sciences
The Hebrew University of Jerusalem
leshem.choshen@mail.huji.ac.il

1 Introduction

Language families and their classification are a base for various questions in many fields of linguistics. Questions are being asked about the differences between specific families and the differences between specific languages in them. We wish to raise some questions about the possibility of language classification using solely phonemic inventories. The identification of language families is not based on any one rule, but relies on countless factors such as history, lexicons, cognition and phonetics. We wish to compare various phonemic based automatic methods to help us better understand at least the phonetic aspects we rely upon or should rely upon when we make the decisions about language families. This paper conducts two main experiments evaluating automatic measures for language clusterings, finding phonemes to be more reliable than phonemic features as a space in which languages should be compared, and suggesting there are different features that are better still, metric learning proves to be a better way than any of the rule-based linguistically derived distance metrics. The second experiment included ordering different sets of features showing which features tend to signify two languages come from different families and which features tend to diverge more, without being of different origins.

2 Relevant background

As we use different computational tools and terms that some readers might not be familiar with, we wish to explain and define a few.

Edit distance between two phonemes is the minimal amount of actions needed to transform one phoneme to another. An action in that sense would be addition or deletion of a feature in the features that represent the phoneme. In terms of computation, edit distance is simply the count of features that represent one of the two phonemes but not the other.

Bipartite graph is a graph in which two sets of nodes exist which have edges connecting nodes between the sets and not between nodes in the same set. In figure 1 we see a bipartite graph in which the right nodes are one set and the left ones are the other set. We see edges crossing, but not edges inside those sets of nodes.

Bipartite graph matching - given a bipartite graph the goal is to look for the maximum matching. I.e. finding a set with the maximum number of edges such that each node is connected at most once. Intuitively, a perfect match would have all nodes connected, having each node from one side matching a node from the other. In figure 1 the edges in **bold** are such perfect match. There may be cases in which there is more than one possible choice of edges that would yield a maximal match, for breaking those ties, one could have weights on the edges. If there are weights on the edges, one would look for the maximal match with the minimal overall weight, thus choosing edges which satisfy a desired property represented by the weights. In our case, if a phoneme is a node, and an edge is the transformation of a phoneme in one language to a phoneme in another, such a property might be the edit distance

Metric learning is an area of supervised machine learning where the goal is to learn from examples a distance function that measures how similar or related two objects are. The result of Metric learning is

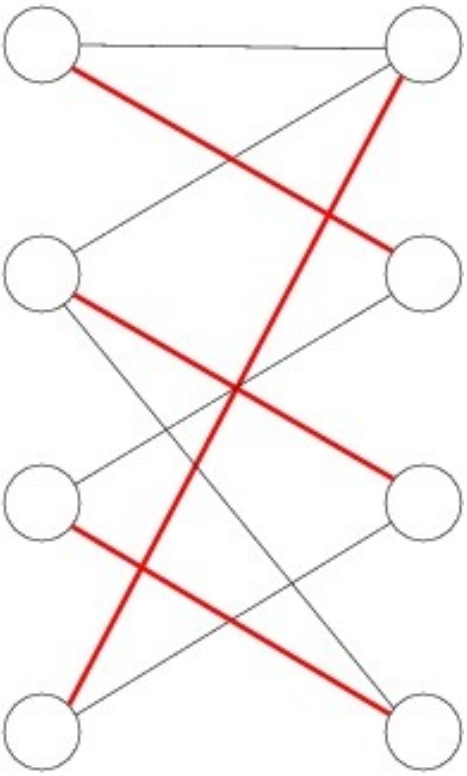


Figure 1: Bipartite graph match

usually a representation or, equivalently, a space, in which the euclidean distance is the learned distance. Technically speaking, the learned distance can be any pseudo-distance fulfilling the requirements of symmetry, non-negativity, and the triangle inequality, but indiscernibles may be not identical.

As far as linguistics background is concerned, we might note that language families are found in many linguistic studies (Aikhenvald and others, 1999) and sophisticated automatic tools were already used for various related tasks (Bouckaert et al., 2012). We do not know of other attempts to confront language proximity in terms of phonemic inventory and a comparison of automatic ways to do so. The use of automatic ways is important as it shows an general notion is relevant to many cases without the need of ad hoc solutions. Phonemic inventories have the advantage of being available in many languages and dialects and are also structured, allowing for meaningful comparison. This drives us to look for a better use of phoneme inventories in language comparison and in creating measures to assess language families using phonemic inventories. Despite those advantages, as an exact definition of language family is hard to get by, as phonemic inventories tend to change more than phonetic rules and syntax (Mohammadi,) and as there are many parameters with which to compare and data is relatively scarce, this is a challenging task.

3 Methods

Data Throughout this paper we use two datasets, the Kurdistan dataset created in class and the Eurasia database (Nikolaev and 2015,). We combine the two datasets using the fields Group and the broader gen subfield respectively to represent language families. Language families containing 4 or less languages were dropped from our experiments. [1]

The first part of the project was to try and automatically divide languages to families based on their phonological inventories. More specifically, Diana (Patnaik et al., 2016) divisive hierarchical cluster-

[1] All code is freely available and can be found in https://github.com/borgr/languageClustering

ing algorithm was used and the research question was which distance measure could be used to compare language inventories well. There is a hidden assumption in this effort that, in many cases, the phoneme inventory is sufficient to classify a language to its family. If the assumption doesn't hold, we would not be able to find any distance that gives more than random results. If on the other hand we do find a well behaving distance metric, we also prove, at least for the languages in our data, that phoneme inventories are a relevant factor for identifying language families. The proposed metrics that were created were based on three representation types, or three projections to n-dimensional spaces; Binary - a representation of the phonemes that existed in the language, bag of features - a positive integer representation counting the number of times each feature (e.g. âĂIJtapâĂİ) was found in the phonemes of the inventory and bag of n-grams - (n=2 was used to avoid too many features) containing the counts of phonemes that have the specific n-tuple features (e.g. having the phoneme d would add 1, among others, to the count of âĂIJalveolar + plosiveâĂİ). After projecting to these spaces, conventional metrics may be used (euclidean, cosine similarity etc.), choosing the right metric, even given a representation, is not an easy task by itself and many options exist. Thus, distances were chosen to cover different types of distance metrics(Cha, 2007; Choi et al., 2010), specifically for binary representations Hamming, Jaccard and Yole distances were used and for non binary cosine and euclidean. Another approach in the direction of choice over distances was metric learning. With manual choice of a metric we may always be in doubt that perhaps the interesting information is well represented in our current features, but we chose an inappropriate metric. For that reason half of the languages were randomly assigned for evaluation to ITML (Davis et al., 2007), a metric learning algorithm. Many other metric learning methods (Shental et al., 2002) were tested and dropped due to technicalities, mainly ones concerning the small number of instances of certain classes (language families) in the database. ITML algorithm uses information theoretic tools, to find a distance metric closest to the euclidean distance metric that satisfies the constraints that samples from the same class are closer to each other than samples that are not. Lastly, a distance of a different flavor was used, inspired by the work of Macklin-Cordes and Round (2015). It was not based on any vector space. Instead, to compute the distance between two phoneme inventories, an alignment of the most similar phonemes was done and the total edit distance of features was used. Alignment is done by reformulating the problem as finding a bipartite graph match with minimal weights, the edit distance is considered to be the weights between phonemes of the two compared languages. Using Kuhn Munkres algorithm we can efficiently find the best solution. Under this method underlies a relaxation of the way phonemes change, assuming most of the time phonemes that are not borrowed either gain or lose some features but do not split or converge. While that is evidently not true in all cases, it might be reasonable that this is not as big as a relaxation as a bag of features is. In this spirit, we check multi to multi alignment with edit distance. In this distance each phoneme is aligned to the closest phoneme in the other language. The distance is the overall distance between aligned phonemes. The idea behind such an alignment is that phonemes might split or be merged and this plays an important enough factor in the way languages evolve as to deem the latest relaxation's validity questionable. This assumes splits and merges of phonemes tend to be closer than unrelated letters that did not originate from the same proto language. This distance also assumes that creation of new phonemes and deletion of a phoneme altogether is rare enough in the same family and can be ignored.

A second part of the project was to assess which features are especially good for differentiating families of languages. For that I have used feature elimination tests using the 3 different sets of features spanning our spaces; letters, feature unigrams, feature bigrams. These tests give us a rank over the features telling us which feature is the most useful for classification. Technically, it is a repetitive fitting of a machine learning classifier, in our case logistic regression, again and again, each time removing the least important feature from the list of available features for classification. Linguistically, if a feature is ranked higher it means it is a better measure to determine languages are of different family, and if it is low, it is either rare or as frequent in one family as in

another, suggesting it might be a common thing that is added\{ }removed from the language by phonemic change or contact and not something that if added is stable across the generations and stays in many of the familyâĂŹs languages.

4 Disadvantages of the suggested methods and further work

- We obviously do not have a good representation of most of the language families, but we have enough for general conclusions.

- It is not unlikely that an ingenious distance metric exists that will be much better than any of the ones introduced here, that might be considered further work.

- Hierarchical clustering comparison was done mainly by visualization. Quantitative measures, such as validating indexes are a possible tool even though they add another significant time consuming part to the project and fit non-hierarchical clustering better.

- Our held-out data is per language, so perhaps a better approach for a train test method was to hold full language families out and to see what happens then, can metric learning methods generalize differences in language families they did not train with.

- Are the same features and distances the best when separating Eurasia languages from the languages of Kurdistan?

- As this was only an initial step subcategories that existed (for eurasia languages only) were not used to specify that languages inside a family are closer in terms of distance, but that will produce more data and surely make the metrics much more robust. Some metric learning algorithms get as input tuples of dist(X[a],X[b]) < dist(X[c],X[d]) where X data and a,b,c,d indices, for those we can use both groups and subgroups of languages as we know the relations. Other metrics only deal with equivalence classes as input and thus can not benefit from this information.

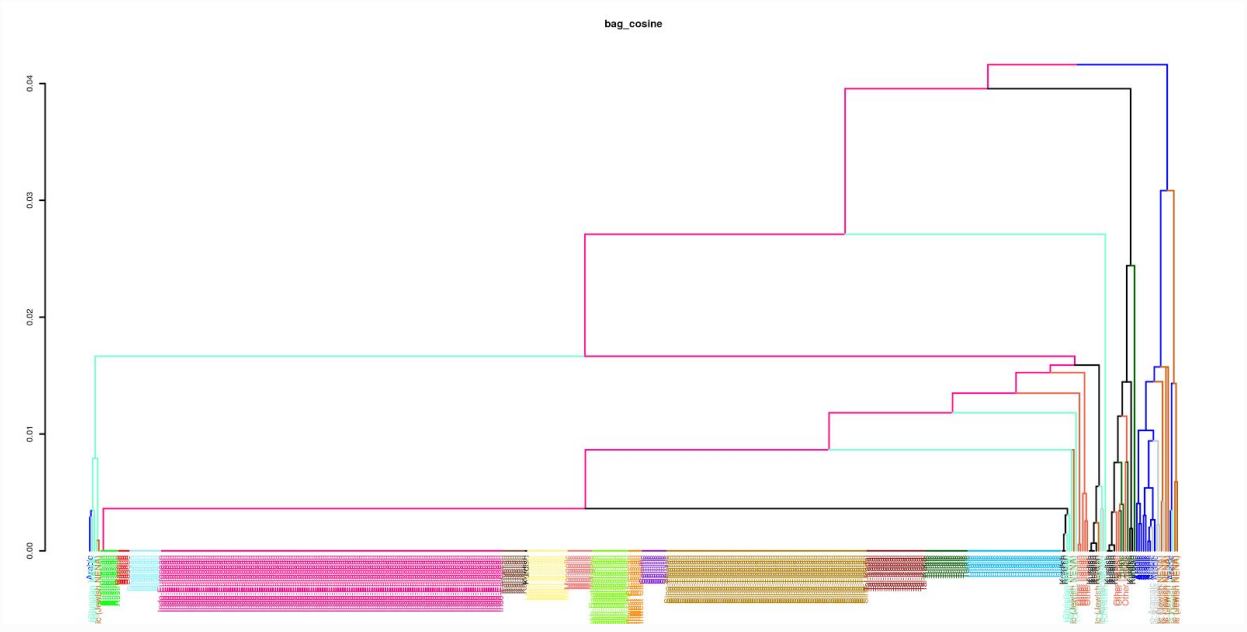
- In the scenario of metric learning we have a problematic split between validation and test, 50% of the languages in the test were the training data. Other distances do not train and thus have no such problems.

5 Analysis and findings

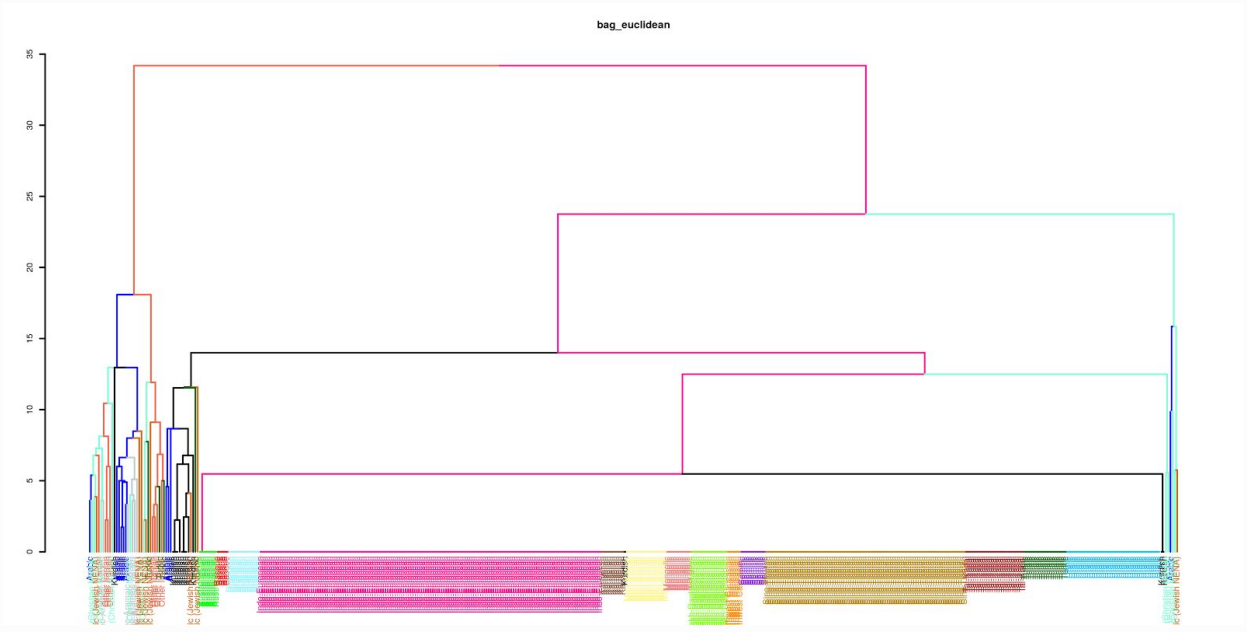
Clustering results show that alignment, bag of features and n-gram representations all fail to create reasonable metrics, with many of the phoneme inventories not necessarily from the same family having the exact same distances from each other and so being clustered together in large groups. Using binary representation and comparing with Hamming distance or Jaccard (not Yole), we see results which are quite promising. The plots in Appendix B show that the phonemes themselves have more information than the relaxations created for this experiment and that the information can be used as is. We may carefully deduce from the failure of the alignment distance that although phonemes may change gradually, some small features are not as easily changed, and to make a good comparison aspects of the way each feature change must be carefully studied and weighted. Results from learning to rank are not only a possible direction to look for metrics, they can also tell us how good are our extracted features, we base our features on the thought that the phonetic features are a fact and the only way to classify phonemes (and perhaps they are), but are they the right features to look at on this problem? The results we get suggested they might not be. With half the languages (and quarter of the relations) to train on, ITML is doing quite well, many of the families indeed get their own clusters. It seems it is doing even better than the binary phonemic inventory representation, as can be seen for example in the nice clustering it gives to Tai-kadai. This might be the place to address the fact that we still see the difficulty of the problem and witness the groups that are separated to smaller clusters and the languages that are attached to the wrong cluster of languages. Another interesting finding might be the matrix learned, we supply in

See Appendix C for the full rankings of the features. From the results over feature selection we can see that the glottal and pharyngeal as well as the phonemes 'ÊŤ', 'ÊŢ' are high in the list of fea-

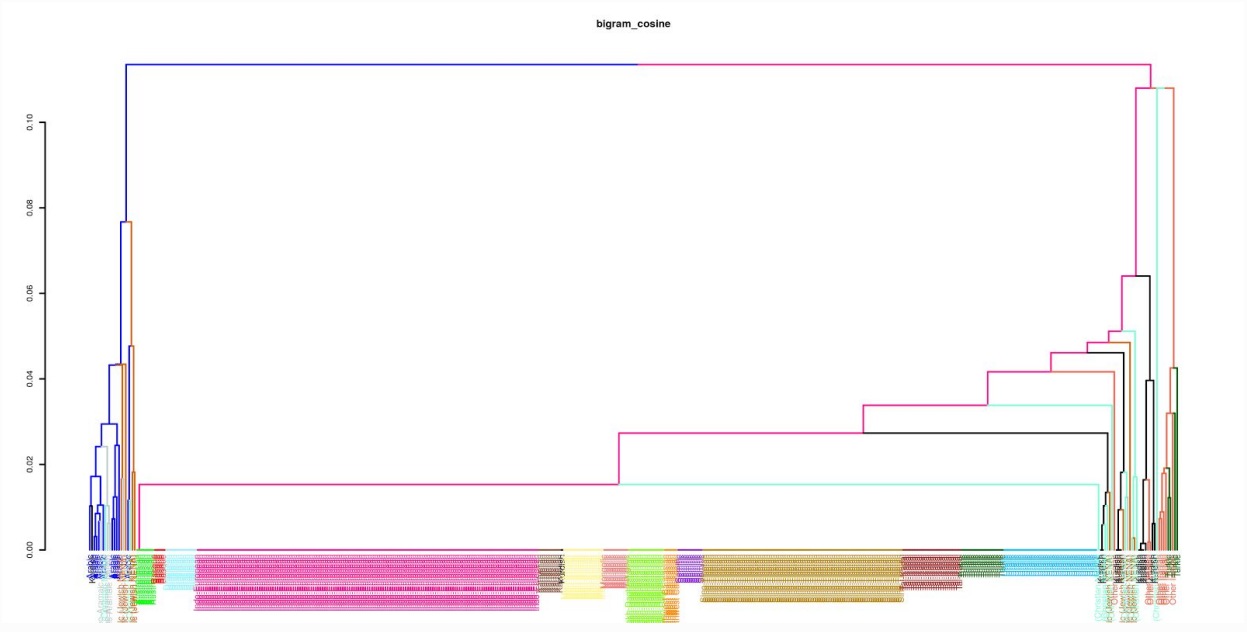
Cosine similarity with bag of features



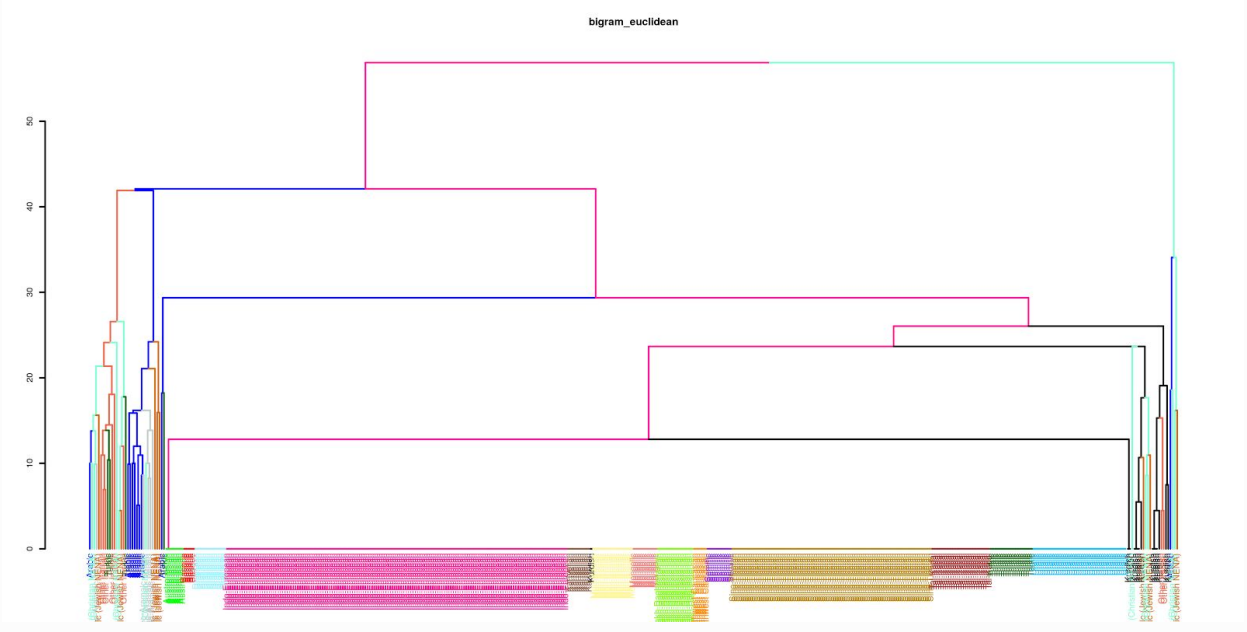
Euclidean distance with bag of features



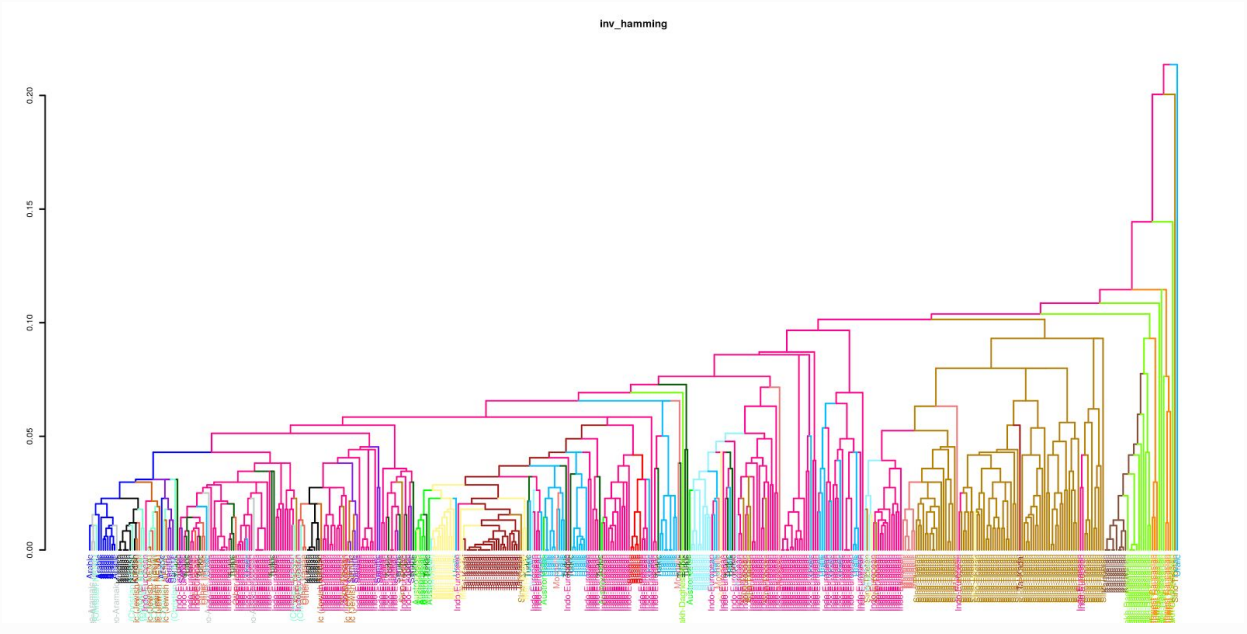
Cosine similarity with bigrams of features



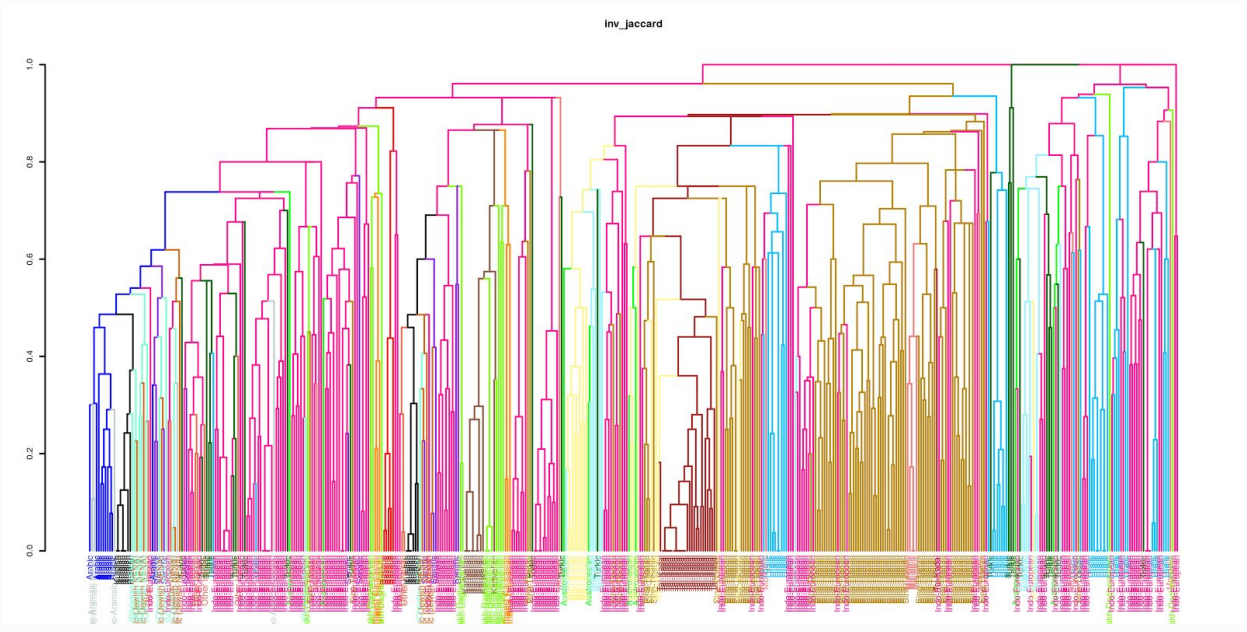
Euclidean distance with bigrams of features



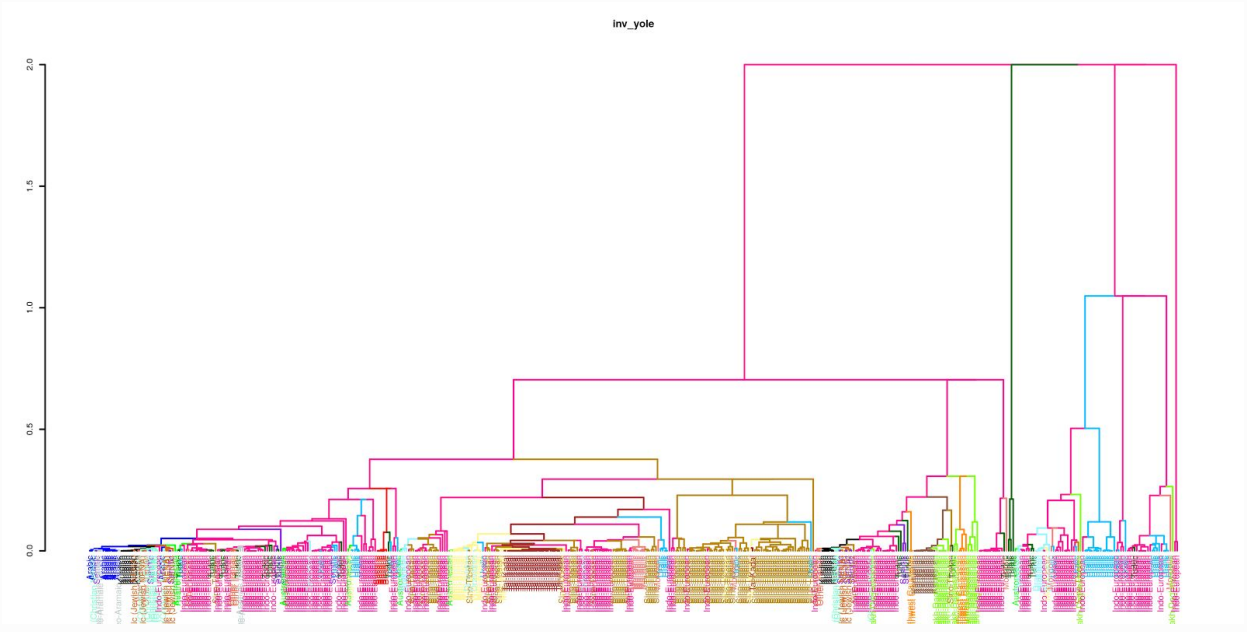
Hamming distance with binary phoneme inventory representation



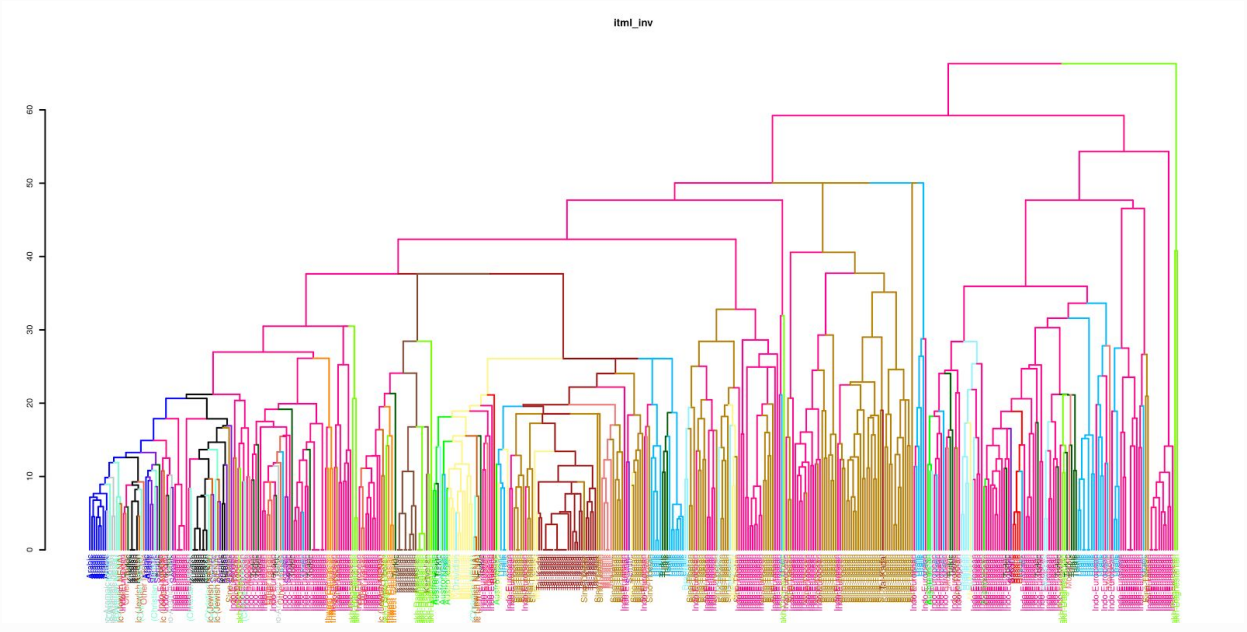
Jaccard distance with binary phoneme inventory representation



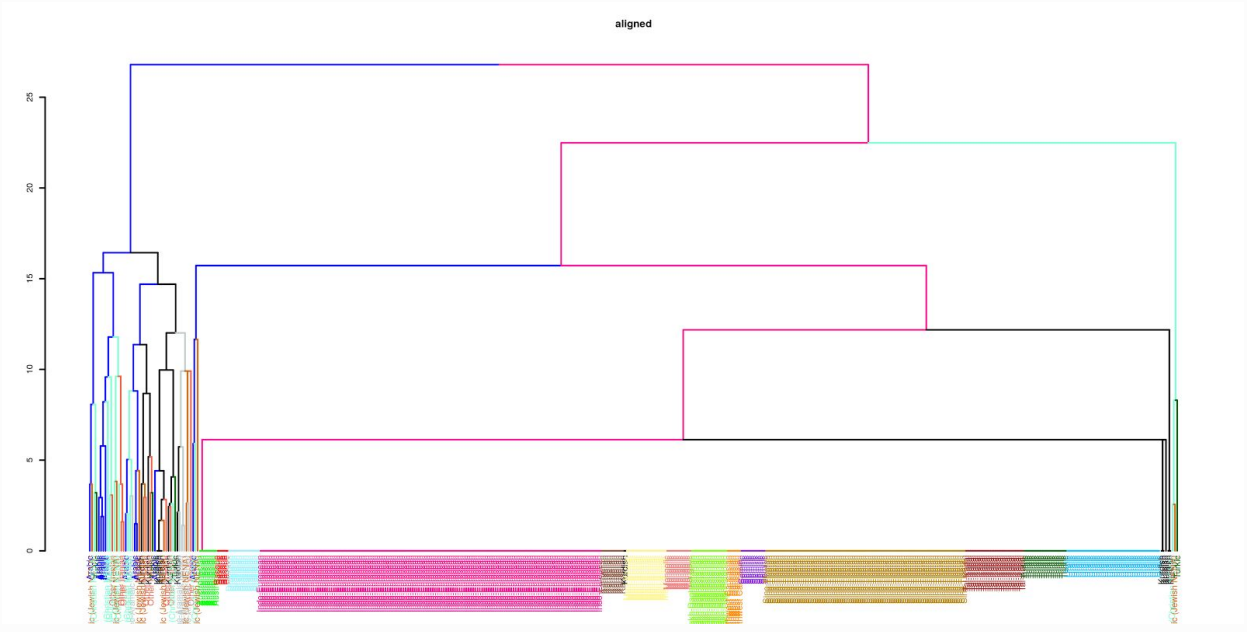
Yole distance with binary phoneme inventory representation



ITML model learned by half of the training data



Edit distance of features of aligned phonemes



tures, it is a sign that at least something expected is happening in the feature selection. Originally we were looking at the languages of Kurdistan, among those it is a significant sign of Arabic, which, to my knowledge is rare in other language families such as the Indo-Iranian. We also see various phonemes similar to âĂŸtsâĂŹ being very distinctive, perhaps suggesting they donâĂŹt tend to be adopted, while common and hardly changing âĂŸbâĂŹ is relatively low in the list. Most humbly I do not pretend to know enough to be able to analyze the results in this part well by myself, and they call for a linguist with more phonetic background to have a more insightful say about them.

A Executing

If the results are not found and you wish to calculate them by yourself note the following things:

- It might be faster to ask leshem.choshen@mail.huji.ac.il for the caches as the two levels of cache (db and distance matrix) make things much faster.
- The main python file âĂIJbinarize_representations.py contains mains of the two parts
- calculate_distances_main(inv_db, feature_db, base_db)
- choose_features_main(inv_db, feature_db, base_db)
- It may be the case you only wish to run one of those.
- Running of the distance main may require a lot of processing time, and for distance learning multithreading is done. Running everything from scratch may take a day of computing.
- Given that you have computed the distance matrices you want to plot (copied or ran the main), run the R file hierarchy_option2.R to create the plots

B Dendrograms

Features ordered by usefulness: ['tap,', 'glottal,', 'lateral_approximant,', 'pharyngeal,', 'labiodental,', 'postalveolar,', 'alveolar,', 'plosive,', 'post_aspirated,', 'lateral,', 'post_pharyngealised,', 'interdental,', 'velar,', 'voiced,', 'trill,', 'affricate,', 'voiceless,', 'fricative,', 'rounded,', 'bilabial,', 'approximant,', 'uvular,', 'non_lateral,', 'labial-velar,', 'front,', 'nasal,', 'close,', 'velarised,', 'palatal,', 'vowel,', 'consonant,', 'retroflex,', 'post_retracted,', 'post_velarised,']
calculated model
Bigrams of features ordered by usefulness: ['alveolar, plosive', 'fricative, alveolar', 'pharyngeal, voiced', 'post_pharyngealised, alveolar', 'alveolar, voiced', 'plosive, post_aspirated', 'voiceless, plosive', 'plosive, glottal', 'voiced, plosive', 'non_lateral, labiodental', 'post_pharyngealised, postalveolar', 'velar, fricative', 'approximant, post_pharyngealised', 'voiceless, bilabial', 'alveolar, fricative', 'postalveolar, post_pharyngealised', 'plosive, post_pharyngealised', 'consonant, pharyngeal', 'consonant, labiodental', 'plosive, consonant', 'alveolar, non_lateral', 'consonant, postalveolar', 'non_lateral, voiced', 'non_lateral, interdental', 'approximant, consonant', 'consonant, post_pharyngealised', 'post_pharyngealised, affricate', 'plosive, voiced', 'post_pharyngealised, approximant', 'glottal, plosive', 'voiced, pharyngeal', 'bilabial, voiceless', 'alveolar, post_pharyngealised', 'post_aspirated, plosive', 'labiodental, non_lateral', 'voiced, alveolar', 'affricate, postalveolar', 'affricate, post_pharyngealised', 'affricate, alveolar', 'voiced, fricative', 'plosive, alveolar', 'non_lateral, velar', 'consonant, lateral', 'labiodental, consonant', 'plosive, non_lateral', 'non_lateral, alveolar', 'interdental, non_lateral', 'voiceless, glottal', 'fricative, velar', 'bilabial, post_pharyngealised', 'consonant, voiced', 'non_lateral, postalveolar', 'post_pharyngealised, plosive', 'pharyngeal, non_lateral', 'approximant, lateral', 'postalveolar, fricative', 'consonant, post_aspirated', 'consonant, voiceless', 'fricative, voiced', 'fricative, interdental', 'post_pharyngealised, consonant', 'non_lateral, glottal', 'labiodental, fricative', 'non_lateral, pharyngeal', 'lateral, lateral_approximant', 'plosive, voiceless', 'velar, voiceless', 'uvular, non_lateral', 'alveolar, affricate', 'non_lateral, tap', 'postalveolar, affricate', 'post_pharyngealised, bilabial', 'voiced, non_lateral', 'consonant,

alveolar', 'labiodental, voiceless', 'fricative, postalveolar', 'lateral_approximant, consonant', 'glottal, consonant', 'pharyngeal, consonant', 'uvular, fricative', 'alveolar, tap', 'non_lateral, plosive', 'voiceless, velar', 'interdental, fricative', 'lateral_approximant, lateral', 'fricative, labiodental', 'post_aspirated, consonant', 'voiced, postalveolar', 'post_pharyngealised, voiced', 'alveolar, consonant', 'pharyngeal, fricative', 'glottal, non_lateral', 'plosive, velar', 'voiced, consonant', 'voiceless, labiodental', 'consonant, lateral_approximant', 'postalveolar, consonant', 'alveolar, voiceless', 'voiced, interdental', 'consonant, plosive', 'tap, consonant', 'fricative, uvular', 'lateral, voiced', 'fricative, pharyngeal', 'glottal, voiceless', 'post_pharyngealised, voiceless', 'interdental, voiced', 'velar, non_lateral', 'approximant, lateral_approximant', 'tap, voiced', 'voiced, labial-velar', 'voiceless, non_lateral', 'consonant, glottal', 'non_lateral, post_aspirated', 'consonant, approximant', 'postalveolar, voiced', 'bilabial, consonant', 'voiceless, alveolar', 'velar, plosive', 'consonant, interdental', 'voiced, lateral_approximant', 'voiced, labiodental', 'postalveolar, non_lateral', 'bilabial, voiced', 'voiced, tap', 'consonant, uvular', 'voiced, lateral', 'consonant, affricate', 'non_lateral, labial-velar', 'tap, alveolar', 'bilabial, non_lateral', 'labiodental, voiced', 'voiced, bilabial', 'lateral, consonant', 'voiceless, post_pharyngealised', 'interdental, consonant', 'voiced, velar', 'consonant, fricative', 'tap, non_lateral', 'lateral_approximant, voiced', 'voiced, post_pharyngealised', 'postalveolar, voiceless', 'post_aspirated, non_lateral', 'approximant, labial-velar', 'lateral_approximant, approximant', 'consonant, tap', 'consonant, bilabial', 'vowel, rounded', 'fricative, post_pharyngealised', 'non_lateral, voiceless', 'lateral, approximant', 'consonant, palatal', 'consonant, labial-velar', 'consonant, trill', 'voiceless, postalveolar', 'velar, voiced', 'post_pharyngealised, fricative', 'front, vowel', 'non_lateral, uvular', 'labial-velar, approximant', 'non_lateral, bilabial', 'affricate, non_lateral', 'trill, voiced', 'labial-velar, consonant', 'fricative, voiceless', 'fricative, non_lateral', 'rounded, vowel', 'voiceless, consonant', 'velar, consonant', 'labial-velar, voiced', 'non_lateral, affricate', 'uvular, consonant', 'voiced, trill', 'rounded, close', 'non_lateral, post_pharyngealised', 'consonant, non_lateral', 'alveolar, lateral_approximant', 'labial-velar, non_lateral', 'voiceless, fricative', 'voiceless, uvular', 'fricative, consonant', 'voiceless, post_aspirated',

'front, non_lateral', 'lateral_approximant, alveolar', 'trill, consonant', 'consonant, velar', 'non_lateral, palatal', 'post_pharyngealised, lateral_approximant', 'affricate, consonant', 'alveolar, approximant', 'close, non_lateral', 'post_aspirated, postalveolar', 'trill, alveolar', 'uvular, voiceless', 'post_pharyngealised, non_lateral', 'approximant, alveolar', 'post_aspirated, voiceless', 'postalveolar, post_aspirated', 'uvular, voiced', 'close, vowel', 'post_pharyngealised, lateral', 'trill, non_lateral', 'lateral, alveolar', 'voiceless, pharyngeal', 'non_lateral, fricative', 'post_aspirated, affricate', 'non_lateral, front', 'voiceless, affricate', 'lateral, post_pharyngealised', 'palatal, non_lateral', 'affricate, post_aspirated', 'pharyngeal, voiceless', 'alveolar, trill', 'alveolar, lateral', 'non_lateral, close', 'affricate, voiceless', 'velar, post_aspirated', 'nasal, consonant', 'voiced, uvular', 'lateral_approximant, post_pharyngealised', 'non_lateral, trill', 'nasal, voiced', 'post_aspirated, velar', 'vowel, front', 'voiced, nasal', 'palatal, consonant', 'rounded, front', 'approximant, velar', 'non_lateral, nasal', 'front, close', 'post_pharyngealised, nasal', 'non_lateral, retroflex', 'consonant, nasal', 'velar, approximant', 'retroflex, voiced', 'front, rounded', 'voiced, approximant', 'nasal, post_pharyngealised', 'uvular, plosive', 'non_lateral, consonant', 'voiced, retroflex', 'voiced, affricate', 'rounded, non_lateral', 'plosive, bilabial', 'retroflex, consonant', 'nasal, non_lateral', 'approximant, voiced', 'consonant, retroflex', 'close, rounded', 'affricate, voiced', 'retroflex, non_lateral', 'bilabial, plosive', 'close, front', 'plosive, uvular', 'approximant, retroflex', 'vowel, close', 'post_aspirated, voiced', 'approximant, non_lateral', 'velar, post_pharyngealised', 'palatal, voiced', 'retroflex, post_pharyngealised', 'non_lateral, vowel', 'nasal, velar', 'voiced, post_aspirated', 'non_lateral, approximant', 'retroflex, approximant', 'alveolar, post_aspirated', 'velar, nasal', 'non_lateral, rounded', 'post_pharyngealised, velar', 'post_pharyngealised, trill', 'post_pharyngealised, retroflex', 'post_aspirated, alveolar', 'voiced, palatal', 'vowel, non_lateral', 'trill, post_pharyngealised', 'post_pharyngealised, uvular', 'uvular, post_pharyngealised', 'velarised, alveolar', 'bilabial, nasal', 'velarised, approximant', 'nasal, bilabial', 'consonant, velarised', 'velarised, consonant', 'alveolar, velarised', 'velarised, voiced', 'velarised, lateral_approximant', 'glottal, fricative', 'lateral, velarised', 'fricative, glottal', 'voiced, velarised', 'approximant, velarised', 'tap,

post_pharyngealised', 'lateral_approximant, velarised', 'post_pharyngealised, tap', 'velarised, lateral', 'interdental, post_pharyngealised', 'post_pharyngealised, interdental', 'approximant, palatal', 'voiceless, approximant', 'bilabial, post_aspirated', 'approximant, labiodental', 'post_aspirated, bilabial', 'palatal, approximant', 'approximant, voiceless', 'labiodental, approximant', 'pharyngeal, post_pharyngealised', 'post_pharyngealised, pharyngeal', 'velar, post_retracted', 'interdental, voiceless', 'voiceless, interdental', 'consonant, post_retracted', 'post_retracted, non_lateral', 'alveolar, nasal', 'post_retracted, plosive', 'nasal, alveolar', 'plosive, post_retracted', 'voiceless, post_retracted', 'post_retracted, voiceless', 'post_retracted, velar', 'lateral_approximant, velar', 'velar, lateral_approximant', 'plosive, palatal', 'lateral, velar', 'post_retracted, consonant', 'velar, lateral', 'palatal, plosive', 'non_lateral, post_retracted', 'post_velarised, voiced', 'affricate, palatal', 'lateral_approximant, post_velarised', 'consonant, post_velarised', 'palatal, affricate', 'approximant, post_velarised', 'post_velarised, alveolar', 'post_velarised, lateral', 'alveolar, post_velarised', 'voiced, post_velarised', 'lateral, post_velarised', 'post_velarised, lateral_approximant', 'post_velarised, approximant', 'post_velarised, consonant', 'fricative, palatal', 'palatal, fricative', 'voiceless, palatal', 'palatal, voiceless', 'bilabial, fricative', 'palatal, nasal', 'nasal, palatal', 'fricative, bilabial', 'retroflex, fricative', 'fricative, retroflex', 'post_pharyngealised, labiodental', 'labiodental, post_pharyngealised', 'glottal, alveolar', 'labial-velar, post_pharyngealised', 'labial-velar, interdental', 'labial-velar, post_velarised', 'labial-velar, tap', 'interdental, uvular', 'palatal, post_pharyngealised', 'labial-velar, lateral', 'labial-velar, nasal', 'labial-velar, close', 'bilabial, post_retracted', 'bilabial, affricate', 'bilabial, palatal', 'glottal, labiodental', 'labial-velar, lateral_approximant', 'glottal, labial-velar', 'glottal, post_retracted', 'interdental, plosive', 'palatal, lateral_approximant', 'labial-velar, velarised', 'lateral_approximant, post_retracted', 'labial-velar, affricate', 'glottal, glottal', 'tap, post_retracted', 'labial-velar, palatal', 'tap, plosive', 'glottal, vowel', 'glottal, velar', 'voiceless, velarised', 'glottal, retroflex', 'glottal, bilabial', 'tap, velar', 'voiceless, close', 'palatal, interdental', 'tap, trill', 'glottal, postalveolar', 'tap, post_aspirated', 'glottal, uvular', 'glottal, front', 'glottal, close',

'tap, pharyngeal', 'glottal, post_velarised', 'tap, rounded', 'palatal, tap', 'glottal, trill', 'glottal, rounded', 'close, nasal', 'tap, labiodental', 'glottal, pharyngeal', 'glottal, interdental', 'glottal, post_aspirated', 'vowel, glottal', 'palatal, post_velarised', 'glottal, post_pharyngealised', 'alveolar, close', 'interdental, labiodental', 'interdental, alveolar', 'glottal, approximant', 'glottal, velarised', 'voiceless, retroflex', 'voiceless, trill', 'palatal, post_retracted', 'voiceless, front', 'glottal, nasal', 'glottal, lateral', 'post_aspirated, labial-velar', 'glottal, voiced', 'trill, lateral_approximant', 'glottal, tap', 'glottal, palatal', 'bilabial, labial-velar', 'affricate, approximant', 'post_retracted, velarised', 'front, labial-velar', 'trill, tap', 'glottal, lateral_approximant', 'trill, post_retracted', 'bilabial, labiodental', 'trill, front', 'bilabial, retroflex', 'trill, post_aspirated', 'bilabial, trill', 'trill, plosive', 'palatal, pharyngeal', 'bilabial, rounded', 'velarised, retroflex', 'trill, vowel', 'velarised, glottal', 'voiceless, rounded', 'glottal, affricate', 'bilabial, front', 'bilabial, velar', 'bilabial, glottal', 'pharyngeal, palatal', 'bilabial, bilabial', 'bilabial, vowel', 'bilabial, pharyngeal', 'trill, postalveolar', 'trill, labiodental', 'palatal, front', 'bilabial, uvular', 'bilabial, postalveolar', 'bilabial, lateral_approximant', 'tap, vowel', 'bilabial, alveolar', 'palatal, lateral', 'lateral, interdental', 'lateral_approximant, fricative', 'lateral_approximant, front', 'tap, postalveolar', 'velar, pharyngeal', 'velar, front', 'interdental, velar', 'bilabial, close', 'vowel, labial-velar', 'post_aspirated, glottal', 'nasal, interdental', 'trill, approximant', 'palatal, close', 'bilabial, lateral', 'bilabial, interdental', 'palatal, rounded', 'voiceless, vowel', 'bilabial, tap', 'nasal, uvular', 'bilabial, post_velarised', 'nasal, fricative', 'trill, voiceless', 'voiceless, voiceless', 'bilabial, velarised', 'bilabial, approximant', 'lateral, palatal', 'nasal, affricate', 'palatal, post_aspirated', 'voiced, glottal', 'plosive, nasal', 'rounded, post_retracted', 'palatal, trill', 'nasal, lateral', 'plosive, tap', 'nasal, tap', 'plosive, post_velarised', 'nasal, post_velarised', 'plosive, interdental', 'fricative, fricative', 'plosive, lateral_approximant', 'voiceless, labial-velar', 'vowel, bilabial', 'nasal, approximant', 'vowel, vowel', 'lateral, affricate', 'vowel, plosive', 'vowel, alveolar', 'trill, rounded', 'trill, pharyngeal', 'vowel, labiodental', 'vowel, postalveolar', 'voiced, vowel', 'palatal, bilabial', 'vowel, uvular', 'palatal, vowel', 'vowel, velar', 'trill, fricative', 'vowel, retroflex', 'interdental, vowel', 'vowel, trill', 'tap, glottal', 'vowel, post_aspirated', 'post_retracted,

approximant', 'tap, labial-velar', 'plosive, close', 'vowel, pharyngeal', 'vowel, voiceless', 'velar, velarised', 'trill, post_velarised', 'trill, interdental', 'tap, close', 'vowel, fricative', 'vowel, post_retracted', 'velar, affricate', 'tap, bilabial', 'velar, palatal', 'lateral, close', 'lateral, lateral', 'tap, voiceless', 'vowel, nasal', 'interdental, retroflex', 'vowel, lateral', 'fricative, front', 'vowel, voiced', 'tap, fricative', 'vowel, tap', 'tap, front', 'vowel, post_velarised', 'close, post_pharyngealised', 'vowel, interdental', 'post_aspirated, retroflex', 'vowel, lateral_approximant', 'vowel, post_pharyngealised', 'vowel, approximant', 'vowel, consonant', 'voiceless, lateral_approximant', 'plosive, lateral', 'velarised, labiodental', 'velarised, rounded', 'vowel, velarised', 'vowel, affricate', 'pharyngeal, interdental', 'vowel, palatal', 'pharyngeal, tap', 'plosive, labial-velar', 'pharyngeal, post_velarised', 'palatal, retroflex', 'pharyngeal, lateral_approximant', 'palatal, velar', 'pharyngeal, approximant', 'plosive, plosive', 'plosive, vowel', 'labiodental, front', 'front, glottal', 'fricative, post_retracted', 'palatal, uvular', 'fricative, close', 'plosive, labiodental', 'fricative, nasal', 'plosive, postalveolar', 'fricative, lateral', 'palatal, postalveolar', 'fricative, tap', 'palatal, labiodental', 'close, voiced', 'plosive, retroflex', 'close, lateral', 'plosive, trill', 'alveolar, pharyngeal', 'palatal, alveolar', 'pharyngeal, velarised', 'plosive, rounded', 'pharyngeal, affricate', 'plosive, pharyngeal', 'post_retracted, rounded', 'post_retracted, interdental', 'plosive, front', 'post_retracted, lateral_approximant', 'plosive, fricative', 'nasal, velarised', 'post_retracted, post_pharyngealised', 'alveolar, front', 'nasal, lateral_approximant', 'affricate, post_velarised', 'post_retracted, affricate', 'alveolar, velar', 'post_retracted, palatal', 'alveolar, retroflex', 'affricate, tap', 'alveolar, post_retracted', 'close, approximant', 'palatal, velarised', 'affricate, lateral_approximant', 'voiceless, nasal', 'voiceless, lateral', 'voiceless, voiced', 'plosive, affricate', 'voiceless, tap', 'voiceless, post_velarised', 'affricate, velarised', 'alveolar, labial-velar', 'affricate, affricate', 'palatal, glottal', 'alveolar, uvular', 'plosive, velarised', 'plosive, approximant', 'alveolar, bilabial', 'voiced, post_retracted', 'alveolar, vowel', 'front, bilabial', 'nasal, nasal', 'nasal, post_retracted', 'nasal, front', 'nasal, retroflex', 'velarised, vowel', 'alveolar, glottal', 'postalveolar, post_retracted', 'front, plosive', 'non_lateral, lateral_approximant', 'nasal, postalveolar',

'velarised, bilabial', 'palatal, labial-velar', 'nasal, labiodental', 'alveolar, alveolar', 'alveolar, postalveolar', 'alveolar, labiodental', 'alveolar, interdental', 'front, alveolar', 'affricate, front', 'front, labiodental', 'close, interdental', 'close, labial-velar', 'postalveolar, tap', 'close, glottal', 'postalveolar, post_velarised', 'close, bilabial', 'close, palatal', 'close, velarised', 'alveolar, rounded', 'postalveolar, interdental', 'postalveolar, lateral_approximant', 'front, postalveolar', 'close, plosive', 'front, uvular', 'close, alveolar', 'front, velar', 'close, labiodental', 'front, retroflex', 'close, postalveolar', 'close, uvular', 'uvular, trill', 'front, trill', 'close, velar', 'front, post_aspirated', 'close, retroflex', 'close, trill', 'close, post_aspirated', 'front, pharyngeal', 'affricate, interdental', 'front, voiceless', 'velarised, plosive', 'non_lateral, velarised', 'postalveolar, front', 'front, front', 'close, pharyngeal', 'front, fricative', 'close, voiceless', 'front, post_retracted', 'affricate, lateral', 'close, close', 'close, fricative', 'close, post_retracted', 'approximant, vowel', 'approximant, plosive', 'fricative, post_velarised', 'uvular, glottal', 'fricative, lateral_approximant', 'lateral_approximant, non_lateral', 'alveolar, palatal', 'lateral_approximant, close', 'labiodental, labial-velar', 'velarised, labial-velar', 'labiodental, glottal', 'fricative, approximant', 'labiodental, bilabial', 'fricative, velarised', 'labiodental, vowel', 'fricative, affricate', 'labiodental, plosive', 'approximant, affricate', 'labiodental, alveolar', 'post_retracted, labial-velar', 'labiodental, labiodental', 'post_retracted, glottal', 'labiodental, postalveolar', 'post_retracted, bilabial', 'labiodental, uvular', 'post_retracted, vowel', 'labiodental, velar', 'post_retracted, retroflex', 'labiodental, retroflex', 'post_retracted, alveolar', 'labiodental, trill', 'post_retracted, labiodental', 'labiodental, post_aspirated', 'post_retracted, postalveolar', 'labiodental, rounded', 'post_retracted, uvular', 'labiodental, pharyngeal', 'uvular, bilabial', 'uvular, vowel', 'velar, rounded', 'approximant, postalveolar', 'approximant, uvular', 'post_retracted, trill', 'post_retracted, post_aspirated', 'labiodental, post_retracted', 'uvular, postalveolar', 'consonant, vowel', 'post_retracted, pharyngeal', 'labiodental, close', 'post_retracted, post_velarised', 'labiodental, nasal', 'post_retracted, front', 'labiodental, lateral', 'uvular, uvular', 'post_retracted, fricative', 'post_retracted, post_retracted', 'labiodental, tap', 'post_retracted, tap', 'labiodental, post_velarised',

'post_retracted, close', 'labiodental, interdental', 'post_retracted, nasal', 'labiodental, lateral_approximant', 'uvular, velar', 'post_retracted, lateral', 'uvular, retroflex', 'post_retracted, voiced', 'lateral_approximant, nasal', 'tap, nasal', 'labiodental, velarised', 'tap, lateral', 'labiodental, affricate', 'affricate, fricative', 'labiodental, palatal', 'tap, tap', 'postalveolar, labial-velar', 'tap, post_velarised', 'postalveolar, glottal', 'tap, interdental', 'postalveolar, bilabial', 'tap, lateral_approximant', 'affricate, nasal', 'velarised, trill', 'postalveolar, vowel', 'postalveolar, plosive', 'postalveolar, alveolar', 'tap, velarised', 'tap, palatal', 'velarised, post_aspirated', 'postalveolar, labiodental', 'postalveolar, postalveolar', 'postalveolar, uvular', 'post_velarised, front', 'postalveolar, velar', 'post_velarised, fricative', 'postalveolar, retroflex', 'post_velarised, post_retracted', 'postalveolar, trill', 'post_velarised, close', 'post_velarised, nasal', 'post_aspirated, vowel', 'postalveolar, pharyngeal', 'postalveolar, rounded', 'uvular, post_aspirated', 'uvular, pharyngeal', 'front, nasal', 'close, tap', 'close, post_velarised', 'uvular, front', 'front, lateral', 'approximant, approximant', 'front, voiced', 'close, lateral_approximant', 'approximant, interdental', 'post_pharyngealised, palatal', 'non_lateral, lateral', 'postalveolar, close', 'non_lateral, non_lateral', 'postalveolar, nasal', 'post_pharyngealised, post_velarised', 'postalveolar, lateral', 'uvular, nasal', 'uvular, lateral', 'front, tap', 'close, affricate', 'front, post_velarised', 'approximant, tap', 'front, interdental', 'approximant, nasal', 'front, lateral_approximant', 'approximant, close', 'front, post_pharyngealised', 'approximant, fricative', 'approximant, post_retracted', 'non_lateral, post_velarised', 'consonant, rounded', 'postalveolar, approximant', 'consonant, front', 'postalveolar, velarised', 'consonant, close', 'interdental, trill', 'postalveolar, palatal', 'interdental, post_aspirated', 'uvular, labial-velar', 'interdental, rounded', 'interdental, pharyngeal', 'velarised, postalveolar', 'uvular, rounded', 'velarised, uvular', 'velarised, velar', 'post_velarised, voiceless', 'nasal, labial-velar', 'nasal, glottal', 'post_velarised, labial-velar', 'uvular, labiodental', 'uvular, alveolar', 'uvular, tap', 'nasal, vowel', 'uvular, post_velarised', 'nasal, plosive', 'uvular, interdental', 'post_velarised, postalveolar', 'uvular, lateral_approximant', 'post_velarised, retroflex', 'post_velarised, pharyngeal', 'post_velarised, rounded', 'affricate, pharyngeal', 'velar, vowel', 'velar,

labial-velar', 'velar, bilabial', 'velar, glottal', 'post_aspirated, velarised', 'interdental, tap', 'interdental, post_velarised', 'interdental, interdental', 'interdental, lateral_approximant', 'affricate, rounded', 'post_velarised, non_lateral', 'velar, alveolar', 'affricate, close', 'uvular, post_retracted', 'retroflex, labial-velar', 'lateral_approximant, plosive', 'affricate, post_retracted', 'uvular, close', 'retroflex, glottal', 'voiced, front', 'retroflex, bilabial', 'retroflex, vowel', 'post_aspirated, palatal', 'interdental, postalveolar', 'rounded, labial-velar', 'interdental, bilabial', 'rounded, glottal', 'interdental, glottal', 'rounded, bilabial', 'voiced, close', 'voiced, voiced', 'lateral_approximant, labiodental', 'lateral_approximant, uvular', 'approximant, bilabial', 'close, consonant', 'fricative, rounded', 'uvular, approximant', 'fricative, post_aspirated', 'uvular, velarised', 'fricative, trill', 'uvular, palatal', 'uvular, affricate', 'velar, labiodental', 'rounded, fricative', 'velar, velar', 'velar, postalveolar', 'velar, uvular', 'lateral_approximant, retroflex', 'rounded, plosive', 'lateral_approximant, trill', 'rounded, alveolar', 'lateral_approximant, post_aspirated', 'rounded, labiodental', 'lateral_approximant, rounded', 'rounded, postalveolar', 'lateral_approximant, pharyngeal', 'rounded, uvular', 'lateral_approximant, voiceless', 'rounded, retroflex', 'rounded, voiceless', 'rounded, trill', 'rounded, pharyngeal', 'rounded, post_aspirated', 'rounded, rounded', 'velar, trill', 'velar, retroflex', 'retroflex, plosive', 'retroflex, alveolar', 'front, consonant', 'voiced, voiceless', 'front, approximant', 'approximant, front', 'front, velarised', 'voiced, rounded', 'front, affricate', 'approximant, pharyngeal', 'front, palatal', 'approximant, rounded', 'fricative, labial-velar', 'approximant, trill', 'approximant, post_aspirated', 'approximant, glottal', 'trill, bilabial', 'consonant, consonant', 'nasal, trill', 'velar, close', 'nasal, post_aspirated', 'post_velarised, post_pharyngealised', 'fricative, vowel', 'nasal, rounded', 'nasal, pharyngeal', 'nasal, voiceless', 'affricate, trill', 'velar, tap', 'lateral_approximant, tap', 'velarised, pharyngeal', 'velar, post_velarised', 'velar, interdental', 'rounded, nasal', 'nasal, close', 'rounded, lateral', 'post_velarised, interdental', 'rounded, voiced', 'post_velarised, post_velarised', 'rounded, tap', 'post_velarised, tap', 'rounded, post_velarised', 'lateral_approximant, interdental', 'lateral_approximant, lateral_approximant', 'velarised, voiceless', 'rounded, velar', 'velarised, front', 'trill,

uvular', 'velarised, fricative', 'trill, velar', 'velarised, post_retracted', 'trill, trill', 'trill, retroflex', 'rounded, interdental', 'lateral_approximant, affricate', 'rounded, lateral_approximant', 'lateral_approximant, palatal', 'rounded, post_pharyngealised', 'post_pharyngealised, labial-velar', 'post_pharyngealised, glottal', 'affricate, retroflex', 'retroflex, labiodental', 'lateral, labial-velar', 'retroflex, postalveolar', 'lateral, glottal', 'retroflex, uvular', 'lateral, bilabial', 'retroflex, velar', 'lateral, vowel', 'retroflex, retroflex', 'lateral, plosive', 'retroflex, trill', 'lateral, front', 'retroflex, post_aspirated', 'lateral, labiodental', 'retroflex, rounded', 'lateral, postalveolar', 'retroflex, pharyngeal', 'lateral, uvular', 'retroflex, voiceless', 'retroflex, front', 'rounded, consonant', 'lateral, post_aspirated', 'lateral, retroflex', 'retroflex, post_retracted', 'rounded, approximant', 'lateral, trill', 'interdental, labial-velar', 'lateral, rounded', 'retroflex, close', 'lateral, pharyngeal', 'retroflex, nasal', 'retroflex, lateral', 'rounded, velarised', 'lateral, voiceless', 'post_velarised, palatal', 'lateral, fricative', 'retroflex, tap', 'lateral, post_retracted', 'retroflex, post_velarised', 'lateral, non_lateral', 'retroflex, interdental', 'retroflex, lateral_approximant', 'rounded, affricate', 'post_velarised, affricate', 'rounded, palatal', 'lateral, nasal', 'post_velarised, velarised', 'lateral, tap', 'post_pharyngealised, vowel', 'velarised, non_lateral', 'retroflex, velarised', 'velarised, close', 'retroflex, affricate', 'velarised, nasal', 'retroflex, palatal', 'trill, glottal', 'trill, labial-velar', 'post_pharyngealised, velarised', 'post_pharyngealised, post_pharyngealised', 'trill, close', 'interdental, close', 'trill, nasal', 'interdental, nasal', 'trill, lateral', 'trill, velarised', 'interdental, lateral', 'trill, affricate', 'trill, palatal', 'pharyngeal, labial-velar', 'affricate, uvular', 'pharyngeal, glottal', 'tap, approximant', 'pharyngeal, bilabial', 'lateral_approximant, bilabial', 'pharyngeal, vowel', 'tap, affricate', 'lateral_approximant, glottal', 'velarised, tap', 'post_aspirated, trill', 'post_aspirated, post_aspirated', 'pharyngeal, plosive', 'post_velarised, uvular', 'pharyngeal, alveolar', 'post_velarised, velar', 'pharyngeal, labiodental', 'lateral_approximant, vowel', 'pharyngeal, postalveolar', 'post_velarised, trill', 'post_velarised, post_aspirated', 'velarised, post_velarised', 'post_aspirated, rounded', 'post_aspirated, pharyngeal', 'pharyngeal, uvular', 'post_velarised, glottal', 'post_velarised, bilabial', 'post_aspirated, front', 'pharyngeal, velar', 'post_velarised,

vowel', 'pharyngeal, retroflex', 'post_velarised, plosive', 'pharyngeal, trill', 'lateral_approximant, postalveolar', 'post_velarised, labiodental', 'post_aspirated, approximant', 'pharyngeal, post_aspirated', 'post_pharyngealised, post_aspirated', 'pharyngeal, rounded', 'pharyngeal, pharyngeal', 'fricative, plosive', 'post_pharyngealised, rounded', 'affricate, velar', 'tap, uvular', 'interdental, post_retracted', 'post_aspirated, labiodental', 'pharyngeal, front', 'tap, retroflex', 'interdental, front', 'post_aspirated, uvular', 'post_pharyngealised, front', 'velarised, interdental', 'affricate, vowel', 'post_aspirated, fricative', 'post_aspirated, post_retracted', 'post_pharyngealised, post_retracted', 'pharyngeal, post_retracted', 'affricate, labiodental', 'interdental, approximant', 'post_aspirated, close', 'interdental, velarised', 'post_aspirated, nasal', 'interdental, affricate', 'post_aspirated, lateral', 'interdental, palatal', 'lateral_approximant, labial-velar', 'post_aspirated, post_velarised', 'post_aspirated, tap', 'post_pharyngealised, close', 'velarised, post_pharyngealised', 'affricate, plosive', 'post_aspirated, interdental', 'post_aspirated, post_pharyngealised', 'post_aspirated, lateral_approximant', 'affricate, bilabial', 'velarised, velarised', 'pharyngeal, close', 'velarised, affricate', 'pharyngeal, nasal', 'velarised, palatal', 'pharyngeal, lateral', 'affricate, labial-velar', 'affricate, glottal', 'palatal, palatal', 'labial-velar, post_retracted', 'labial-velar, fricative', 'labial-velar, front', 'labial-velar, voiceless', 'labial-velar, pharyngeal', 'labial-velar, rounded', 'labial-velar, post_aspirated', 'labial-velar, trill', 'labial-velar, retroflex', 'labial-velar, velar', 'labial-velar, uvular', 'labial-velar, postalveolar', 'labial-velar, labiodental', 'labial-velar, alveolar', 'labial-velar, plosive', 'labial-velar, vowel', 'labial-velar, bilabial', 'labial-velar, glottal', 'labial-velar, labial-velar']

calculated model

Phonemes as features ordered by usefulness: ['ŋ', 'm', 'kʰ', 'tsˈˈ', 'ʔ', 'ʕ', 'tsʰ', 'ɖ', 'k', 'q', 'fˈ', 'ɲ', 't̪', 'ts', 'pʰ', 'j', 'n', 'g', 'tʰ', 'tɞ', 'sˤ', 'dʒ', 'dz', 'p', 'ʈ', 'ɾ', 'ʃ', 'tʃʰ', 'sʲ', 'd̪', 'tʃˈˈ', 'c', 'χ', 'ɽ', 'r', 'h', 'r̪', 'w', 'v', 'x', 's', 't', 'ħ', 'ʁ', 'ɞ', 'ʌ', 'd', 'lˌ', 'ɬ', 'lˤ', 'ʒ', 'nʲ', 'ð', 'l̪', 'sˌ', 'bʲ', 'ɟ', 'dˤ', 'ɳ', 'χʷ', 'z', 'ɦ', 'kˈˈ', 'ɣ', 'b', 'θ', 'tʃ', 'tˤ', 'lʲ', 'ʌ̹', 't̪ʰ', 'g', 'n̪', 'ɢ', 'ʂ', 't̪ʰ', 'rˤ', 'kʲ', 'sˌ', 'ʊ', 'rˌ', 'l', 'cç', 'ɬ̪', 'pˈˈ', 'ɬ', 'sʰ', 't̪s̪ʰ', 'ç', 'pˤ', 'qʰ', 'gˠ', 'rʲ', 'tˈˈ', 'd̪z̪', 'ɬ̪', 'ʈʂ', 'ʐ', 'qː', 't̪ˈˈ', 'ɲ̟', 'tʃˤ', 'gʱ', 'ɯ',

'ʜ', 'tsˌ', 'ⁿd', 'dʑ', 'gʲ', 'tʰ', 'zʲ', 'nˌ', 'sˌ', 'nˌ', 'zˤ', 'ɖʐ', 'dʒʰ', 'vʲ', 'ðˤ', 'y', 'kʷ', 'ɗ', 'bˌ', 'ɬʲ', 'β', 'sː', 'tˌsˌ', 'rˌ', 'ɻʲ', 'tɞʰ', 'ɭ', 'qʼ', 'hʲ', 'rˌ', 'tˌ', 'ɐː', 'cʰ', 'ʀ', 'ʁʷ', 'βˌ', 'cˌ', 'mʲ', '(lʲ)', 'fʲ', 'ðˤ', 'zˌˤ', 'bɦ', 'qχ', 'dzˌ', 'gˤ', 'ɓ', 'ɻ', 'ʟ', 'ɴ', 'tʲ', 'ʔ', 'j̃', 'qˤ', 'kʷʰ', 'ʒ', 'lˌ', '(gʲ)', 'ˈd', 'sˌˤ', 'dˌz', 'ɻ', 'kˤ', 'ðˌ', 'zˌ', 'tʃʲ', 'ⁿg', 'zˌ', 'lˠ', 'sˤ', 'pʲ', 'cçː', 'ɲ', 'mˌ', 'ɬˀ', 'dˌɦ', 'dˌ', 'pʲʰ', '(dʲ)', 'dʰ', 'ˈb', 'kx', 'ɻ', 'tˌ', 'lˌ', 'ɻ', 'mɣ', 'ɦ', 'dˌˤ', 'gʷ', 'zˤ', 'ʃʲ', 'ɾˌ', 'ʕ', 'tʂ', 'xˌ', 'dˌʰ', 'tˌsˌ', 'dˌ', 'ⁿdz', '(sʲ)', 'tˌʰ', 'ˈɣ', 'dɦ', 'ʃː', 'tç', 'tˌs', 'tʰʲ', 'tʲʰ', 'ŋˌ', 'ɾˌ', 'ɖɦ', 'fɣ', 'dˌˌ', 'tˤ', 'tˌʲ', 'ɢˌ', 'sˌˌ', 'ˈg', 'nˌ', 'qχ', 'ɬ', 'fˀ', '(bʲ)', 'tsˌ', 'ɣʲ', 'gʰ', 'ɾˌ', 'g˜', 'ʍ', 'ɬˌ', 'tˌ', 'dɦʷ', 'zˌ', 'ⁿb', 'sɣ', 'xʲ', 'ɽɦ', 'mˤ', 'ˈj', '(nʲ)', 'kʰʲ', 'qʷ', 'rˌ', 'pf', 'dɻ', 'ŋ', 'ɥ', 'tː', 'nˌˤ', 'xː', 'ˈm', 'ɬˌ', 'bˤ', 'nˌɣ', 'tʂʰ', 'dʲ', 'gʲ', 'ɖʰ', 'dzʰ', 'ɻ', 'ⁿdʐ', 'ɾˤ', 'xʷ', 'pː', 'k˜', 'ˈŋ', 'k', 'dˌɣ', 'lˌ', 'ⁿdʑ', 'mˌ', 'wˌ', 'ɻˤ', 'kʷ', 'ʁˌ', 'ɦ', 'çˌ', 'ˈn', 'lˌɣ', 'b˜', 'tˌsˌ', 'dˌ', 'rɦ', 'ɾ̃', 'dː', 'dˈ', 'qʷ', 'ˈw', 'dzˌ', 'θˌ', 'ʒʲ', 'bɣ', 'ɾʲ', 'ɮ', 'ɢʁ', 'ⁿdʒ', 'ɦˤ', 'pʰ˜', 'ˈɲ', 'dˌʒˌ', 'ʰt', 'qʼ', 'ɬː', 'lɦ', 'zˌ', 'ʃˌ', 'kʲ', 'kʼʷ', 'vˌ', 'ɥ', 'mʰ', 'ʔː', 'gʷ', 'ɲˌ', 'tˌsˌˌ', 'ɠɣ', 'tˌˌ', 'gː', 'lˌ', 'χʲ', 'ɖˌʐˌ', 'nˤ', 'ŝ', 'pʰʲ', 'tʃʷ', 'ʃˌ', 'ŋˌ', 'ɟɦ', 'gʷ', 'ɣ', 'ʰp', 'tˌˌ', 'nˌˌ', 'bʰ', 'kʼʷ', 'rˌʲ', 'tˌˤ', 'kʰ˜', 'ɾɣ', 'tsʲ', 'mɦ', 'ˈt', 'vˤ', 'tʷʰ', 'dˌˌzˌˌ', 'ʒ́', 'ʃ́', 'sˈ', 'bː', 'qʼʷ', 'ʈʂʷ', 'tˌː', 'ðˌˌ', 'pˆ', 'nˌ', 'ʐˌ', 'kʰʷ', 'θˌ', 'tʃː', 'tɬˈ', 'tˌsˌʰ', 'ɻˌ', 'ðˈ', 'ⁿtʰ', 'ɾˌ', 'nˌɦ', 'ⁿɟ', 'ɟˌ', 'ʰk', 'kʲʰ', 'ʔʷ', 'ʈʲ', 'rː', 'ðˌˌ', 'pˈ', 'cˈ', 'dˌʲ', 'dˌʐˌ', 'ɮˈ', 'lʰ', 'dˌɦ', 'ʰts', 'ⁿpʰ', 'tʃˈ', 'ʂˌ', 'ʂʲ', 'ɖˌʐˌ', 'ŋɦ', 'tɬ', 'nˌˌ', 'kː', 'nˌʲ', 'dˌˌ', 'ʃˌ', 'tˈ', 'ʐʷ', 'ts', 'ⁿkʰ', 'ɦ̃', 'dˌzˌ', 'dzɦ', 'ɟʰ', 'ɲː', 'ʔʷ', 'ʈˌʂˌ', 'ʃʷ', 'tˌˌ', 'nː', 'tˢ', 'ʂʷ', 'ɬˈ', 'ʰtʰ', 'ⁿdˌzˌ', 'ɦʷ', 'tˌʲ', 'wˌʲ', 'tɞˈ', 'ŋː', 'ʒˌ', 'tsːʰ', 'ɖʲ', 'ⁿdˌ', 'zˌʲ', 'ðʲ', 'χː', 'ʰtɞʰ', 'Ɥˤ', 'sˈ', 'lː', 'dˌʲ', 'tˌɣ', 'sˌˌ', 'qʲˈ', 'ɾʰ', 'tsʷ', 'ʰtʃʰ', 'tsː', 'tɬːʰ', 'ŝʷ', 'wˌ', 'xk', 'θˈ', 'dˌ', 'ⁿɖ', 'dʷ', 'dˌʲ', 'ʰtsʰ', 'dˌˌ', 'ɻʲ', 'ɟˌɻˌ', 'tˌsˌʲ', 'tʃːʰ', 'mː', 'tŝʷˈ', 'ŋʲ', 'dˌʒ́', 'ɮʲ', 'ɞˈ', '(ʑ)', 'ʐʲ', 'pːʰ', 'nˌʲ', 'tˌʰʲ', 'tʷˈ', 'tˌ', 'ʰqʰ', 'ʰtʃ', 'ɦ̃ʲ', 'ʈʂˈ', 'ɳʲ', 'pɣ', 'ŋˌ', 'fː', 'ʰʈʂʰ', 'qʼʷ', 'rʰ', 'tŝʷʰ', 'lˌʲ', 'lˌʲ', 'ŋʷ', 'tˌʰ', 'ⁿtɞʰ', 'wʲ', 'ʰpʰ', 'tˌsˌː', '(ɞ)', 'ʁʲ', 'ɾˌɣ', 'kˌ', 'ɸ', 'ʷr', 'ʰsʰ', 'pʸʰ', 'nʰ', '(ʒ)', 'tʂʰ', 'ʰtʃː', 'dˌʰ', 'vɣ', 'ʰkʰ', 'ɬɣ', 'sʷ', 'tɬˈ', 'tˌˠʰ', 'ʂˌ', 'ʰkː', '(z)', 'ɾˌʲ', 'qːʷ', 'tʼʷ', 'ɾˌʲ', 'ʰpː', 'dzʲ', 'ɖʑ', '(dzˌ)', 'xːʷ', '(ʃ)', 'jː', 'dˌ□ʲ', 'qˤˈ', 'ʈʂʰ', 'β', 'dʒʲ', 'kːʷ', 'ʰtsː', 'ⁿɢ', 'ʁˤ', 'çː', 'χːʷ', 'qʰʷ', 'tˌˢʲ', 'tʷ', 'ⁿtˌsˌʰ', 'ʰtː', 'tsˌ', 'dʒː', 'tʃʷʰ', 'tˌɬˈ', 'ʈʂ', 'ðː', 'tˌsˌˈ', 'mʷ', 'tˌɬˈ', 'ʑ', 'kxːˈ', 'dʑʰ', 'ⁿʈʂʰ', 'tŝʷ', 'qχːʰ', 'ʒʷ', 'ʃˈ', 'ʇɾ', 'ʂʰ', 'ʔˈ', 'tɬːˈ', 'qʷʰ', 'ⁿɖʑ', 'dˌʒˌ', 'ʃːʲ', 'pˈʷ', 'ŋʰ', 'xʰ', 'ⁿʈɾʰ', 'ŝˈʷ', 'dˌzˌʲ', 'tˌʃˌ', 'ʇɾʰ', 'ʔˈʷ', 'vː', 'bˌˌɦ', 'ʒːʲ', 'ɬː', 'ˈ ʑ', 'ɖɾ', 'rˌʲ', 'ⁿɖɾ', 'tˌʃˌ', 'tʃːʲ', 'ɖˌˌɦ', 'ɖː', 'lˌʲ', 'dzː', 'dˌzˌ', 'tˌˌ', 'dˌˌʒˌˌɦ', 'ʊʲ', 'kˈː', 'ⁿg', 'ɳˌ', 'dˌzˌː', 'tˌʃˌʷ', 'tsˈː', 'gˌˌɦ', 'tˌsˌʰʷʲ', 'ⁿtsʰ', 'pʷ', 'tˌˌsˌˌ', 'sˌː', 'tʃˈʷ', 'qˈːʷ', 'ɻˌ', 'tˌsˌʰʲ', 'tˌsˌ', 'lʸ', 'ɢʷ', 'dˌʒˌ', 'tˌʰʷʲ', 'm ', 'pʰʷ', 'qˈː', 'tʃˌʰ', 'qˈʷʲ', 'nʸ', 'dˌʒˌʷ', 'ɞʰ', 'tʃˈː', 'qʰʷʲ', 'zˌˌ', 'tʃˌ', 'tɬˈː', 'tˌˈʷʲ', 'ⁿtʃʰ', 'kˈːʷ', 'sˌˌ', 'qʲ', 'lˌː', 'ɳɦ', 'tɬː', 'dˌˌ', 'qʷʲ', 'gːʷ', 'nɦ', 'sˌʷʲ', 'ʔ̃', 'qʰʲ', 'ɠ', 'χʷʲ',

'pʼʲ', 'ʃ', 'kʷʲ', 'tˌʼʷ', 'kʼʲ', 'ɬˌ', 'dʐ', 'zˌʷ', 'kʼʷʲ', 'tˌʰʷ', 'nˌˌ', 'tˌsˌʷʲ', 'tˌsˌʼʲ', 'tˌʷ', 'χˤʷ', 'tˌsˌʼʷ', 'sʲʰ', 'jʰ', 'tˌʷʲ', 'χːʼ', 'tɕʰ', 'ʁːʼ', 'sˌʷ', 'nˌʰ', 'qχʰ', 'dʑ', 'sːʼ', 'mˌʲ', 'tˌʼʲ', 'ʕʷ', 'sˌ', 'ʔˤ', 'tˌsˌʷ', 'ʡʰ', 'χˤ', 'ʃːʼ', 'qʼʲ', 'ɬʷ', 'cːʰ', 'ʀˌ', 'sˌʲ', 'zʰ', 'tɕʼ', 'tʃːʼ', 'kːʷʰ', 'ʜʷ', 'tˌsˌʰʷ', 'zˌ', 'kːʰ', 'tɬʰ', 'qχːʼ', 'tˌsˌʼʷʲ', 'kʰʷʲ', 'hʷ', 'tsːʼ', 'tʃʰʷ', 'ʀ..ʱ', 'kxːʰ', 'tɬʼʷ', 'tʃːʷ', 'tɬʰʷ', 'd..ʱ', 'sˌːʼ', 'tsʰʷ', 'tˌsˌːʼ', 'tʰʷ', 'ɣʷ', 'tʃʼˤ', 'ʃʰ', 'dˌ..ʱ', 'ʃːʷ', 'ʒˤ', 'tˌʼˤ', 'tsʰˤ', 'ʔˤ', 'tˌʰˤ', 'ɬɻˌ', 'pʰˤ', 'lˌ', 'ʀˌ', 'tʃʰˤ', 'rˌˌ', 'zʷ', 'lˌʱ', 'tˌɞˌ', 'tsʼʷ', 'dˌʑˌ', 'dʒʱ', 'dˌʑˌʱ', '□', 'tˌɞˌʰ', 'vˌ', 'ʒˌ', 'ⁿʈʂʰ', 'gːʼ', 'ʰpʲ', 'ŋʲːʼ', 'tɕːʼ', 'ⁿcʰ', 'sʲːʼ', 'zːʼ', 'βʲːʼ', 'βʲ', 'bʲːʼ', 'wʰ', 'lˌʲːʼ', 'mʲːʼ', 'nʲːʼ', 'ʰtʲ', 'bˌːʼ', 'ⁿʈʰ', 'ʰtɕ', 'tsʲːʼ', 'mˌʲːʼ', 'pʲːʼ', 'zʲːʼ', 'lʲːʼ', 'ʰtsʲ', 'kʲːʼ', 'ʼʃ', 'rˌːʼ', 'nˌːʼ', 'ʼʼs', 'ɻːʼ', 'ʼʼl', 'fʲːʼ', 'gʲːʼ', 'rʲːʼ', 'βːʼ', 'dˌːʼ', 'nˌʲ', 'ʂːʼ', 'xʲːʼ', 'nˌʲːʼ', 'rˌʲːʼ', 'ʰkʲ', 'dˌzˌʲ', 'mˌːʼ', 'cçʰ', 'çʰ', 'ʷɟ', 'ʷʑ', 'ʱg', 'ʱl', 'ⁿxʰ', 'ʰʷsʰ', 'ʷm', 'ʱɟ', 'ʷz', 'ʰʷs', 'ʱɖ', 'ʰʷtɕ', 'ʱdz', 'ʰʷʈ', 'ʰʷɕʰ', 'ⁿsʰ', 'ʱɲ', 'ʱdʑ', 'ʱn', 'ʱm', 'ʰc', 'ʰlˌ', 'ʷd', 'ʰʷʂʰ', 'ʱʐ', 'ʱŋ', 'ʷl', 'ʱd', 'ʱb', 'ʰʷk', 'ʷʐ', 'ʰʷts', 'ʱz', 'ʰʈ', 'ʰʷt', 'ʱj', 'ʊˌ', 'nˌˌ', 'tˌɞˌ', 'ⁿqʰ', 'ɬˌˌ', 'dʒˤ', 'ʁˌ', 'ⁿq', 'ⁿʈʂ', 'ⁿts', 'ⁿp', 'ⁿt', 'ⁿk', 'ⁿtʃ', 'sˌ', 'tˌɞˌ', 'ʔˤ', 'fˌ', 'qˌ', 'ⁿc']

C Feature ranking

References

Alexandra Y Aikhenvald et al. 1999. The arawak language family. *The amazonian languages*, pages 65–106.

Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J Greenhill, Alexander V Alekseyenko, Alexei J Drummond, Russell D Gray, Marc A Suchard, and Quentin D Atkinson. 2012. Mapping the origins and expansion of the indo-european language family. *Science*, 337(6097):957–960.

Sung-Hyuk Cha. 2007. Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):1.

Seung-Seok Choi, Sung-Hyuk Cha, and Charles C Tappert. 2010. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48.

Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. 2007. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM.

Jayden L Macklin-Cordes and Erich R Round. 2015. *High-Definition Phonotactics Reflect Linguistic Pasts*. Universitätsbibliothek Tübingen.

Hiwa Asadpour1&Maryam Mohammadi. A comparative study of phonological system of kurdish varieties. *Journal of Language and Cultural Education*, 2:3.

Dmitry; Andrey Nikulin; Nikolaev and Anton Kukhto. 2015. The database of eurasian phonological inventoriess. (available online at http://eurasianphonology.info ; accessed on august 25, 2017).

Ashish Kumar Patnaik, Prasanta Kumar Bhuyan, and KV Krishna Rao. 2016. Divisive analysis (diana) of hierarchical clustering and gps data for level of service criteria of urban streets. *Alexandria Engineering Journal*, 55(1):407–418.

Noam Shental, Tomer Hertz, Daphna Weinshall, and Misha Pavel. 2002. Adjustment learning and relevant component analysis. In *European Conference on Computer Vision*, pages 776–790. Springer.