Principal components of sound systems: An exercise in multivariate statistical typology[*]

Vsevolod Kapatsinski
Indiana University

ABSRACT: Phoneme inventories of the world's languages as depicted by the UPSID database (Maddieson and Precoda 1990) are analyzed using multivariate statistical techniques of principal components analysis and k-means and hierarchical clustering. The first two meaningful principal components, representing dimensions that account for the most variance in sound systems but are not caused by differences in typological frequencies of phonemes, are found to separate languages into three large clusters, distinguished by glottal articulations present in the stop inventory and the sonority of other types of sounds present in the language. Clustering analyses, which automatically categorize sound systems and phonemes, are shown to reveal both areal groupings of languages, for instance, categorizing together genetically unrelated languages of India, and groupings of phonemes that are often interpretable in featural terms, especially when clustering analyses are conducted within phoneme categories defined by manner of articulation / sonority.

1. Introduction

A traditional goal of linguistic typology is to estimate the likelihoods that two languages are related, either due to being descendants of a common ancestor or being spoken by interacting populations of speakers. Relatedness is estimated based on how similar the compared languages are in terms of the relevant features. The goal of the present paper is to compare languages in terms of the similarities of their phoneme inventories and to compare phonemes in terms of similarities in the sets of languages in which they occur. To this end, I will use the technique of principal components analysis (PCA), which reduces a set of correlated variables to a smaller set of orthogonal variables. The resulting orthogonal variables are termed the "principal components" and can be seen as dimensions defining a space in which the objects of study can be situated. Since some objects of study will be closer to each other in the resulting space than others, PCA can serve as the basis for typology (and has been used for this purpose in linguistics by Szmrecsanyi and Konrad, to appear; closely related methods also used in linguistic typology are factor analysis, Biber 1988, Silnitsky 2003, and multidimensional scaling, Croft and Poole, to appear).

In the present paper, I intend to compare phoneme inventories of languages found in the UPSID database (Maddieson and Precoda, 1990). An obvious simple way to measure how similar two phoneme inventories are would be to calculate the proportion of phonemes that are shared by the two inventories and normalize by the sizes of the inventories. One problem with this approach is that phonemes are not equal in terms of how likely they are to occur in both of the compared inventories just by chance. A common phoneme, like /m/, which occurs in 94% of languages, is very likely to occur in both languages. Therefore, if two languages both contain /m/, this is not as surprising as when the two languages contain a very rare segment, like /r^{j}/, and thus should not contribute as much to our estimate of between-language relatedness. Two inventories that

[*] Many thanks to Kenneth de Jong and Stuart Davis for helpful comments on this manuscript and especially to Henning Reetz for providing me with a version of UPSID in a tractable matrix format.

contain only frequent segments are likely to have many of the same phonemes but we would not want to conclude that the languages are related. By contrast, if two phoneme inventories both contain a phoneme that is very rare in the languages of the world in general, this may suggest a genetic or areal connection between the languages. PCA automatically takes into account differences in typological frequency between phonemes and thus is preferable to a simple overlap measure.

Another advantage of PCA is the elimination of correlations between dimensions. To take an extreme example, the presence of a phoneme Y in a language may perfectly predict that the phoneme Z is also present in the language. If this is the case, then if we know that two languages share Y, we also know that they share Z. Thus, if two languages that share Y also share Z, this fact is not surprising and should not increase our estimate of relatedness between the two languages. In a PCA analysis, the perfectly correlated X and Y would be mapped onto a single dimension, correctly capturing this intuition.

The orthogonal variables generated by PCA can be characterized by the phonemes with whose presence or absence they correlate. The closer two languages are to each other in the resulting space, the more similar they are in terms of the phonemes they feature. The resulting measure of similarity between phoneme inventories takes into account how likely the component phonemes are to co-occur in inventories by chance. Therefore, it provides a better estimate of relatedness between languages than simpler measures of overlap.

While situating languages in a similarity/estimated-relatedness space is one goal of linguistic typology, another traditional goal is the identification of natural groupings of phonemes and languages. This can be accomplished by applying clustering techniques, grouping points together based on proximity in the space defined by principal components. In particular, k-means clustering can be used to group points into a limited number (k) of categories based on optimizing the distance between the k categories in the space defined by principal components. In the present paper, k-means clustering will be used to determine whether languages or phonemes separate naturally into multiple groups. In addition, hierarchical clustering will be used to group languages and phonemes into binary-branching hierarchies.

2. Methods

The UPSID database (Maddieson and Precoda 1990, available at http://www.linguistics.ucla.edu/faciliti/sales/software.htm and http://web.phonetik.uni-frankfurt.de/upsid_segment_freq.html) was converted into a phoneme-language coincidence matrix. UPSID contains a quota sample of phonological segment inventories collected from grammatical descriptions of the languages. The present version of UPSID uses an extended sample described by Maddieson and Precoda (1990) and Maddieson (1991). As indicated by Maddieson (1991:196),

> The languages are chosen to represent a properly structured quota sample of the genetic diversity of extant languages. One and only one language is included from each cluster of related languages judged to be separated from its nearest relative to a degree similar to the separation of North and West Germanic (taken to be equivalent to about 1500 years of separate development).

The only modification made to the database for the purposes of the present study was to collapse 'dental', 'alveolar' and 'unspecified dental/alveolar' segments, yielding 771 distinct phonemes. This decision was made because the distinction between specified and unspecified coronals is likely to be due to differences in description methods, and not to inherent characteristics of the segments involved.

If a phoneme Y occurred in a language X, the cell in the X^{th} row of the Y^{th} column in the matrix would have a value of '1', while if Y did not occur in X, the cell would have a value of zero. A snippet of the coincidence matrix is shown in Table 1.

Table 1. A small part of the coincidence matrix analyzed in the present study

	rʲ	θ	b	ɣ
Russian	1	0	1	0
English	0	1	1	0

In addition to the full segments-by-languages matrix, submatrices that contained subsets of the segments occurring in the complete matrix were derived including 1) plosives, 2) fricatives and affricates, 3) sonorant consonants, and 4) monophthong vowels. There are several reasons for splitting the matrix in this way.

First, the principal dimensions along which vowels vary are likely to be different from the principal dimensions of variation for the consonants. Likewise, there are certain phonological features that apply only to fricatives and affricates (stridency) or only diphthong vowels (direction of change). Clicks are restricted in the range of features they can have and in their areal distribution. Since the same features do not apply to all segments, interpreting dimensions derived from principal components analysis in featural terms is much easier if the segments included in the analysis belong to the same major class. This is also true because the same feature may behave differently in different classes of sounds. For instance, the voicing feature behaves differently with sonorant and non-sonorant sounds since voiceless sonorants are marked relative to voiced ones but voiceless obstruents are unmarked relative to voiced ones.

In addition, certain principles of optimal sound systems are more easily observable within a restricted class of sounds than in the entire database. For instance, it is easy to test whether languages favor maximal perceptual dispersion within the monophthong vowel system (Liljencrantz and Lindblom 1972) since the perceptual space of monophthong vowels is well understood and easily described with a small number of dimensions. Thus, one can easily observe that the cardinal vowels are more perceptually distinct than centralized vowels. The same principle is rather difficult to test in the full database since, for instance, vowels and obstruents are so distinct that they would rarely be confused with each other. This lack of relevant confusability data makes it difficult to determine the degree of perceptual similarity between dissimilar segments.

Finally, a language tends to have segments of varying degrees of sonority. Therefore, analysis of a phoneme-language coincidence matrix is expected not to reveal sonority as one of the principal

components and segments are not expected to cluster by sonority, which would indicate that some languages tend to have sonorant segments while others have non-sonorant ones.

The R software environment (http://www.r-project.org/) was used for all statistical analyses. Two types of analyses were conducted: one analysis type used the matrix as described above while the other used the transpose of that matrix, in which the rows of the original matrix became columns and vice versa. The matrix and its transpose were then submitted to a Principal Components Analysis with centering and scaling, which yields a matrix of coordinates for either languages or segments in the new space and a matrix of loadings for either segments or languages respectively. Hierarchical cluster analysis (Altmann 1971, Altmann and Lehfedt 1973, 1980, Cysouw 2007, Szmrecsanyi and Kortmann to appear) was performed on the matrices of coordinates, since this analysis does not discard information and weighs the different dimensions in accordance with how much variance they account for. Plots of languages in principal components space and k-means clustering, on the other hand, are based on the loadings matrices, since these tend to produce a more even distribution of points along the axes as opposed to a large clump and a few outliers arranged along the principal component axes observed in a coordinate matrix. Since the loadings matrices are used for determining locations of points in this analysis, interpretations of the principal components in the resulting space are necessarily done with the matrices of coordinates.

Ward's clustering method and a Euclidean distance measure were used for hierarchical cluster analysis. Other distance measures (Minkowski, city-block) tended to yield similar clusters, except for Canberra distance, which produced a much more even distribution of languages by cluster but removed the tendency for genetically and areally related languages to be located close to each other in the hierarchy. Other clustering methods tended to perform worse than Ward's, producing extensive chaining, and hence maintaining high uncertainty about the cluster assignments of individual languages. For k-means clustering, the initial number of clusters to try was determined by visual inspection of the graph. The number of clusters was then adjusted if successive runs of k-means did not produce the same set of clusters at least 9/10 times.

Finally, in order to obtain the geographical distribution of languages, the World Atlas of Language Structures (WALS, Bibiko 2005, Haspelmath et al. 2005) was used to plot the languages in a particular clusters obtained with k means clustering on a world map. If a language from UPSID was not found in WALS, it was not plotted.

3. Principal Components of the space of phoneme inventories

The results of PCA produce a set of orthogonal dimensions, the first of which in our case perfectly tracks phoneme frequency, i.e., the number of languages a phoneme occurs in. The distribution of languages along principal component (PC) 1 is unimodal and does not deviate significantly from a normal distribution with the same mean and standard deviation (Kolmogorov-Smirnov Test D=.03, p=.96), indicating that phoneme inventories do not split into distinct into groups based on how typologically common the phonemes they contain are. Rather, the distribution of phoneme frequencies across languages is approximately random.

The second and third principal components present a very different picture. Figure 1 shows the distribution of languages in the space defined by the two principal components. The figure shows that languages form three distinct clusters in the space and are reliably separated into the clusters by k-means clustering. The distribution of languages is non-normal ($D=.48$, $p<.0001$ according to the Kolmogorov-Smirnov test), indicating that this distribution is unlikely to come about by chance. Since some cluster assignments may be due to the assumptions that k-means clustering makes about category boundaries (convex, elliptical), a full listing of coordinates in the PC2-PC3 space and resulting cluster assignments for UPSID languages is presented in Appendix 1.

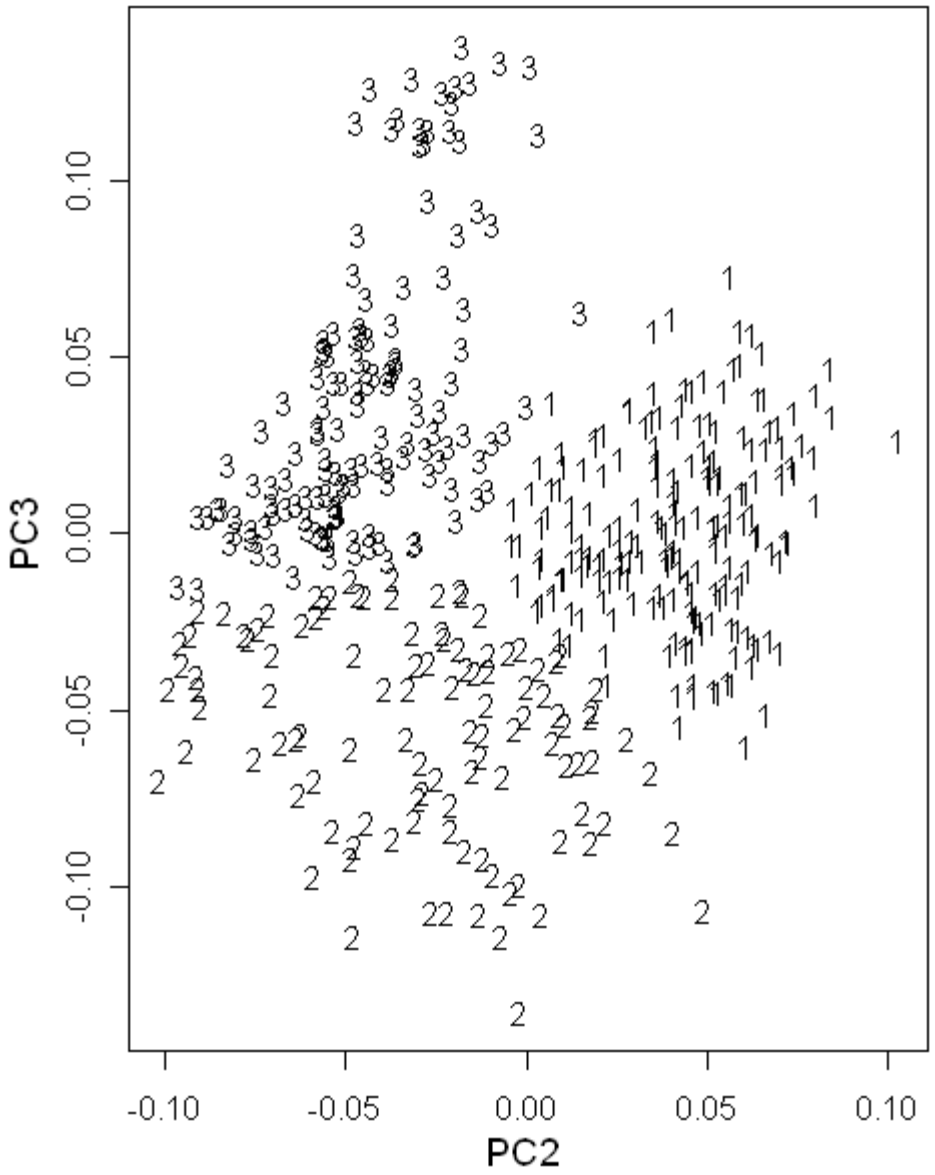


Figure 1. The distribution of languages in the space defined by principal components 2 and 3 with cluster assignments derived by k-means clustering shown by numbers.

Although Figure 1 shows that there are three distinct types of phoneme inventories found in languages of the world, it does not show what the three inventory types are. In order to interpret

this finding, we need to determine the phonemes whose occurrence or non-occurrence correlates with principal components 2 and 3. The phonemes, in UPSID notation, are situated in the PC2-PC3 space in Figure 2.

Cluster 1 membership is strongly correlated with the presence of the basic voiced stops /b/ (r=.65), /d/ (r=.66), and /g/ (r=.61), and less strongly but still significantly (p<.01/771) with the presence of voiced stops with other places of articulation: /ɟ/ (r=.30), and /ɖ/ (r=.21), fricatives /f/ (r=.41), /v/ (r=.30), and /z/ (r=0.28), doubly articulated stops /gb/ (r=.35) and /kp/ (r=.33), the affricate /dʒ/ (r=.30), and the velar nasal /ŋ/ (r=.21). All but two of these segments can be characterized as voiced obstruents. Languages in Cluster 1 tend to lack obstruents with a glottal articulation (ejectives, the glottal stop and aspirated obstruents), significantly correlating with the absence of /t'/, /k'/, /p'/, /q'/, /tʃ'/, /ts'/, /ʔ/, /t^h/, /ts^h/, /k^h/, /p^h/, /$k^{w'}$/ and /k^w/. Cluster 2 membership is correlated with the presence of glottal and to a lesser extent labial or lateral secondary articulation on obstruents, correlating most strongly with the presence of ejectives (.54<r<.63), /h/, and the glottal stop. Interestingly, primary and secondary glottal articulations appear to pattern together, both being associated with languages belonging to Cluster 2. Membership in the cluster also correlates with the absence of /ŋ/ and plain voiceless stops. Cluster 3 membership correlates strongly with the absence of plain voiced stops and many other obstruents and more weakly with the presence of the voiceless palatoalveolar stop /t̠/ The full list of significant correlates of the three clusters is shown in Appendix 2.

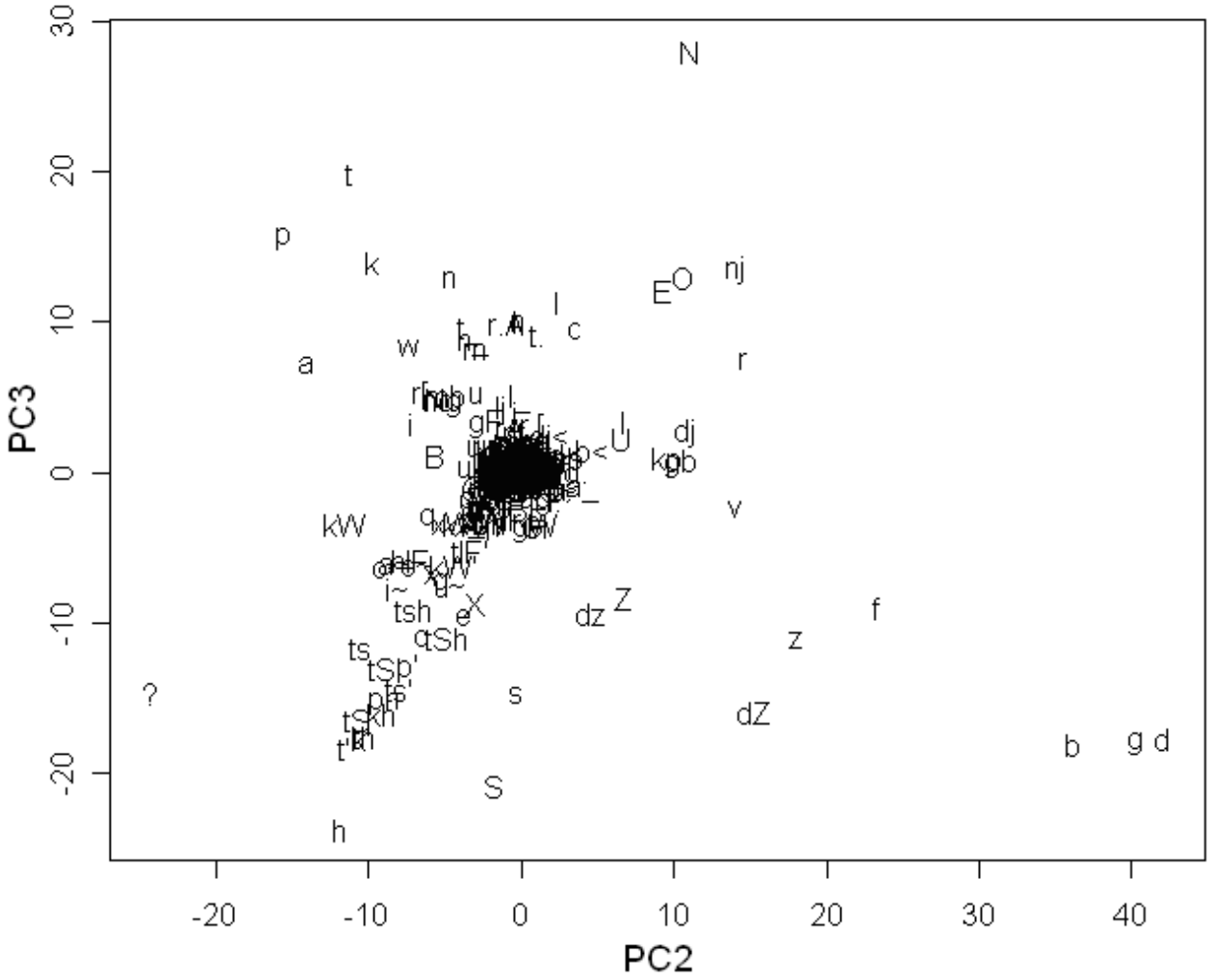


Figure 2. Segments whose presence or absence determines whether a language belongs to the top cluster or to the bottom cluster in Figure 1.

The languages in each cluster were then submitted to the Interactive Reference Tool (Bibiko 2005) of the World Atlas of Language Structures (Haspelmath et al. 2005) to obtain the geographical distributions of languages belonging to the three clusters. Figures 3-5 show the results. Languages in Cluster 1 include almost all African languages, almost all languages of Europe (except Spanish and Breton), all languages of India, and almost all languages of the Middle East. Cluster 2 includes almost all Caucasian and the majority of American languages. Languages in Cluster 3 include the great majority of Australian languages (which form the distinct subcluster at the very top of Cluster 3), and most languages of China and Southeast Asia. Languages of the Americas are unlikely to belong to Cluster 1. Languages of the Pacific islands appear to be split between Clusters 1 and 3 with very few belonging to Cluster 2.

In terms of genetic affiliation, Indo-European languages are associated with Cluster 1 (17/22 belong to this cluster, $\chi^2(1)=8.25$, $p<.005$), as are Nilo-Saharan languages (15/19, $\chi^2(1)=10.54$, $p<.005$), and Niger-Congo languages (40/44, $\chi^2(1)=44.01$, $p<.001$). All six Dravidian languages and the eleven Altaic languages in the database belong to Cluster 1. Cluster 2 includes all five

Mayan languages in the sample, the four Na-Dene languages, and 5/6 Salishan languages. Since languages of the Americas are not classified into large language families, no large family is associated with Cluster 2. Cluster 3 includes all eight Arawakan languages and all four Macro-Ge languages in the sample. It is also significantly associated with Australian languages (17/22 belong to this cluster, $\chi^2(1)=30.29$, $p<.001$). There are also some language families that are split quite evenly between two clusters but tend not to occur in the third one. Austronesian and Trans-New Guinea languages tend not to belong to Cluster 2 (only 1/20 Austronesian languages belongs to this cluster, $\chi^2(1)=6.33$, $p<.025$, and 0/17 Trans New Guinea languages do, $\chi^2(1)=8.44$, $p<.005$). Afro-Asiatic languages tend not to belong to Cluster 3 (0/23 Afro-Asiatic languages belong to this cluster, $\chi^2(1)=14.28$, $p<.001$). Interestingly, mainland Asian language families (Sino-Tibetan, Tai-Kadai) are quite evenly divided between clusters but there are clear areal groupings seen on the maps in Figures 3-5, suggesting borrowing of phoneme characteristics between unrelated languages.

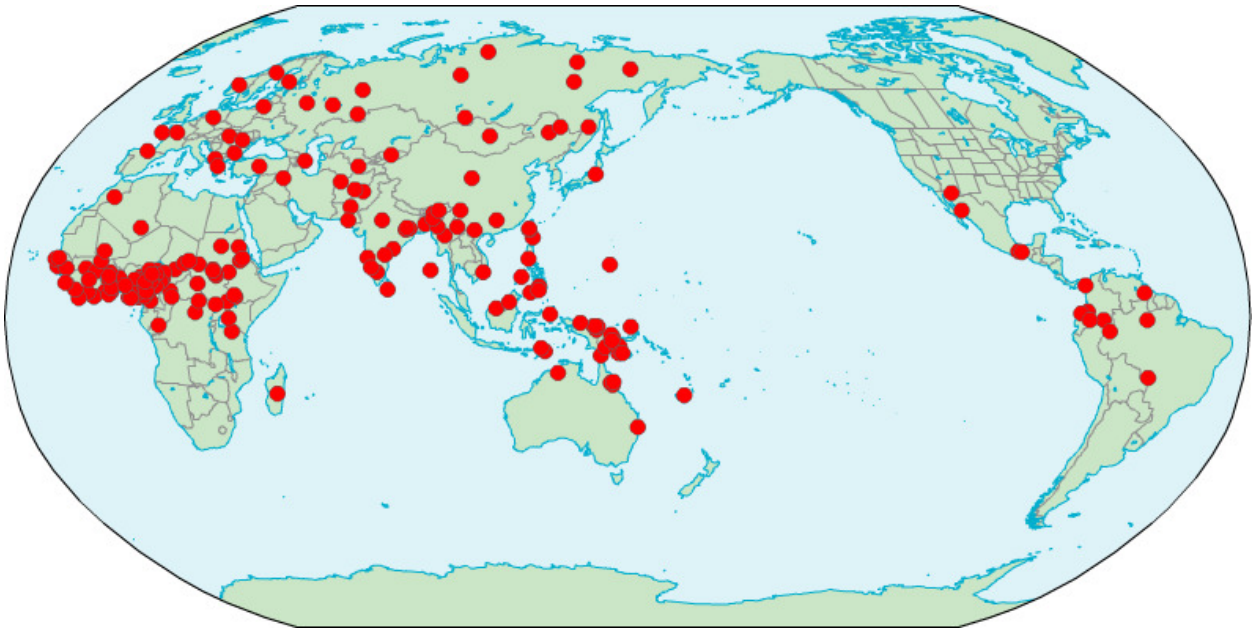


Figure 3. The geographical distribution of languages in Cluster 1.

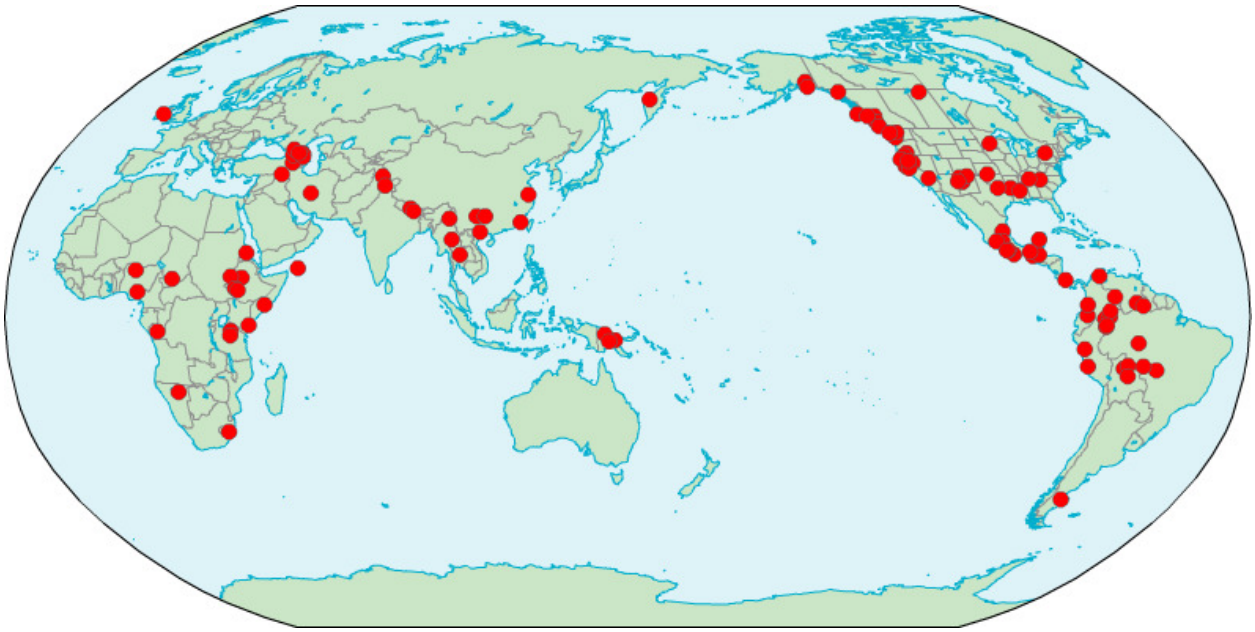


Figure 4. The geographical distribution of languages in Cluster 2.

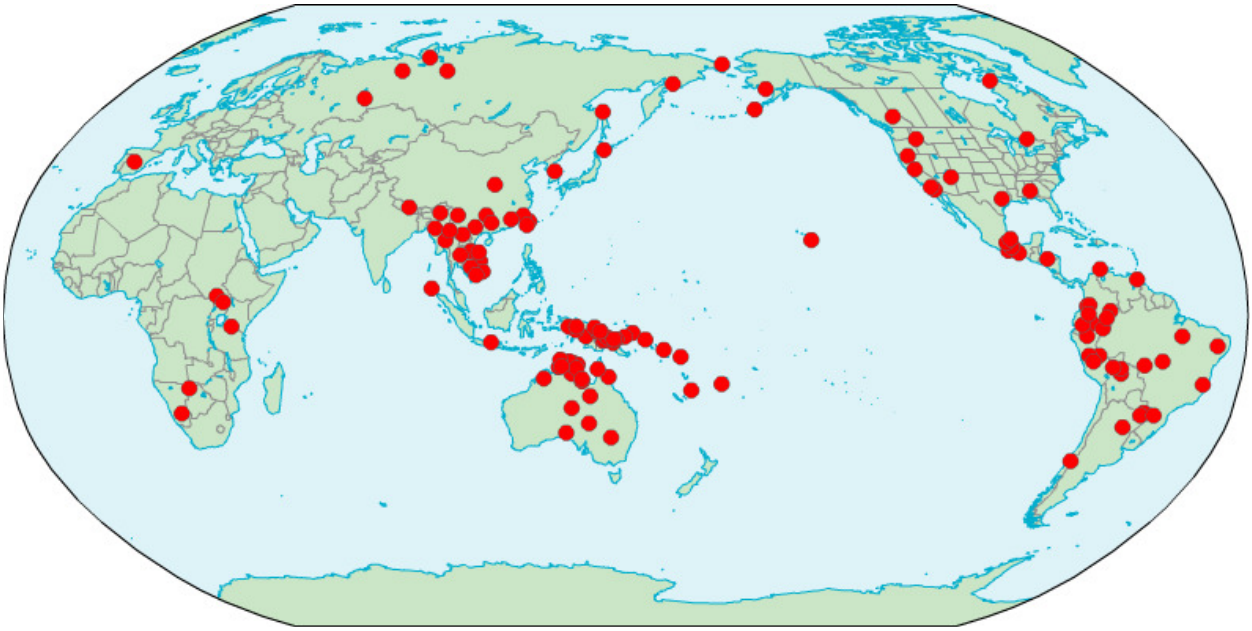


Figure 5. The geographical distribution of languages in Cluster 3.

To summarize the results of the present section, the languages of the world form distinct clusters based mostly on the glottal articulations present in the obstruent inventory. The clusters track areal and genetic boundaries to a significant extent, suggesting that differences in obstruent inventories provide information about language relatedness. Combinations of principal components other than PC2-PC3 do not appear to reliably separate languages into groups and therefore fall outside the scope of this paper.

4. Stop consonant systems

Figure 6 shows the first five principal components of stop systems. As can be seen in the figure, there are three distinct clusters of languages in the PC1-PC2 space. Clear clustering is also seen in the PC3-PC4 space and along PC5. Thus stop inventories of the world's languages separate into distinct subclasses.

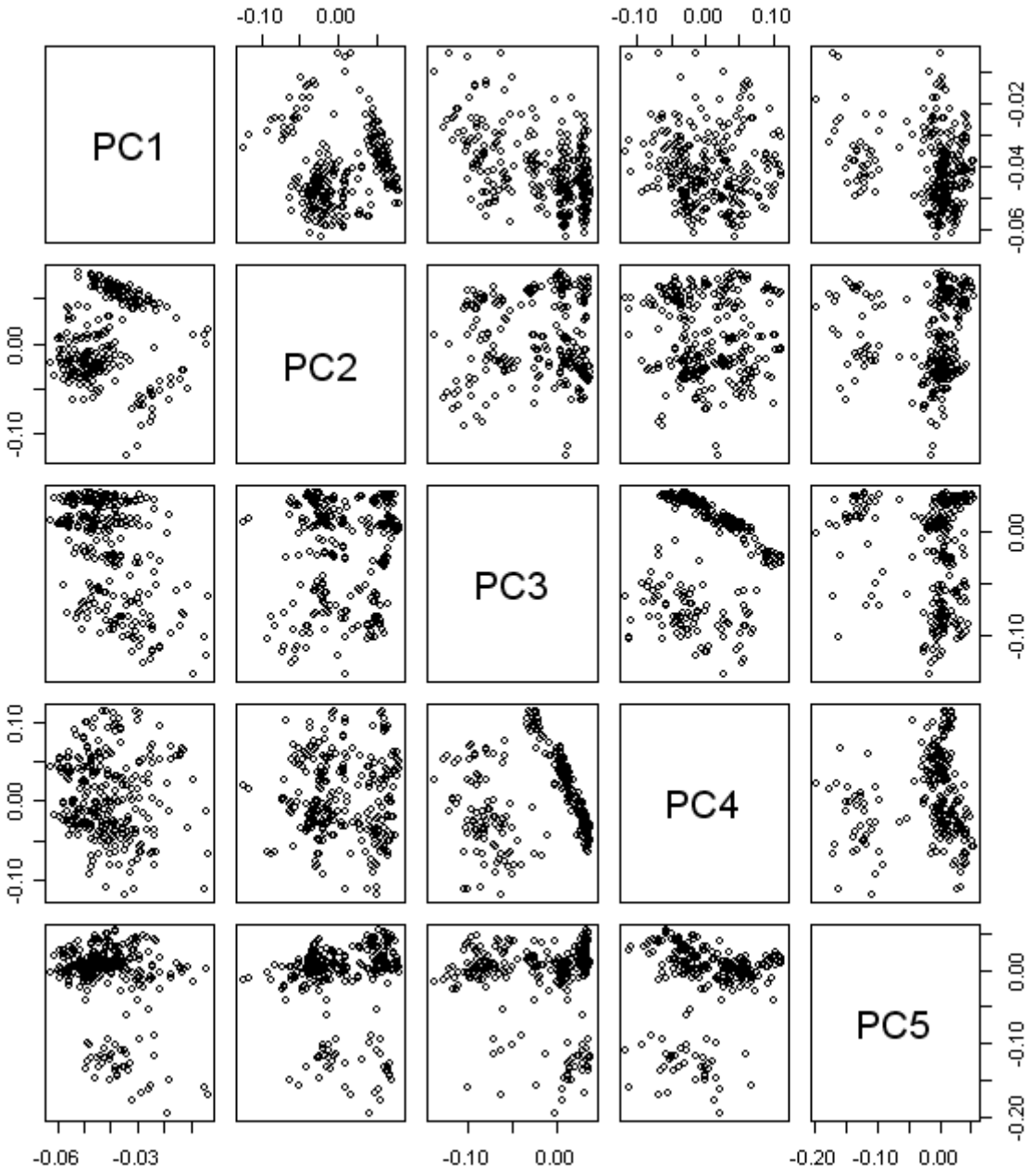


Figure 6. The first five principal components of stop systems display clear clustering of languages.

High scores on the first principal component are assigned to languages lacking one of the basic stops (/p/, /t/, /k/, /b/, /d/, /g/, -.62 < r < -.31) and having aspirated, ejective, or labialized stops. PC2 is correlated very strongly with the presence/absence of the voiced stops /b/, /d/, /g/ (r= -.90): languages with high scores on PC2 lack the basic voiced stops. PC2 has a positive correlation with voiceless basic stops, indicating that languages low on PC2 tend to lack /p/, /t/, or /k/. As Figure 7 shows, the clustering seen in PC1-PC2 space is based almost entirely on PC2, i.e., whether the language is missing voiced basic stops or voiceless ones. The languages that lack some voiced stop(s) are centered around the Pacific Ocean, as shown in Figure 7, but also include Spanish. This cluster corresponds quite closely to clusters 2 and 3 in the analysis of complete sound systems. The much smaller group of languages that lack basic voiceless stops (45 vs. 165 that favor voiceless ones) is very geographically dispersed, including mostly North African languages (although the majority of North African languages do not belong to this cluster), as well as the 5 Nakh-Daghestanian languages in the Caucasus, Norwegian and Irish in Europe, Persian, Khalkha, and Kota in Asia, four Australian languages (Bandjalang, Dyirbal, Yidiny, and Mbabaram), five languages of Papua-New Guinea, and 8 American languages. In the case of some of these languages, like Norwegian and Alamblak, the apparent favoring of voiced stops is because the voiceless stops are aspirated. In the case of others, like Bandjalang or Jomang, there are no voiceless stops in the language while the voiced ones are well represented.

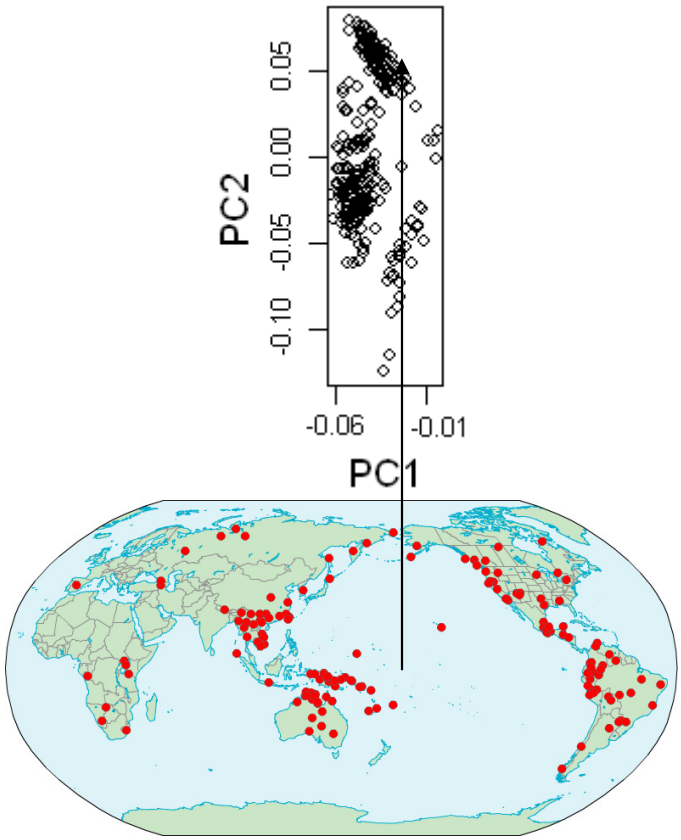


Figure 7. Languages that lack basic voiced (top cluster, which is also shown on the map) or voiceless stops (bottom cluster).

Figure 8 shows the associations between the third and fourth principal components and the presence or absence of specific stop phonemes. High scores on PC3 are strongly correlated with the absence of aspirated stops /p^h/, /k^h/, /t^h/ (r=-.9) and to a lesser extent with the absence of ejective stops or the glottal stop (r=-.4) and the presence of plain voiceless stops (r=.4). Thus a language with a high score on PC3 (on the right in Figure 9) should have few aspirated and ejective stops and many plain voiceless ones. PC4 distinguishes between different types of glottal articulation. High scores on PC4 are associated with the presence of a glottal stop and/or ejectives and the absence of aspirated stops.

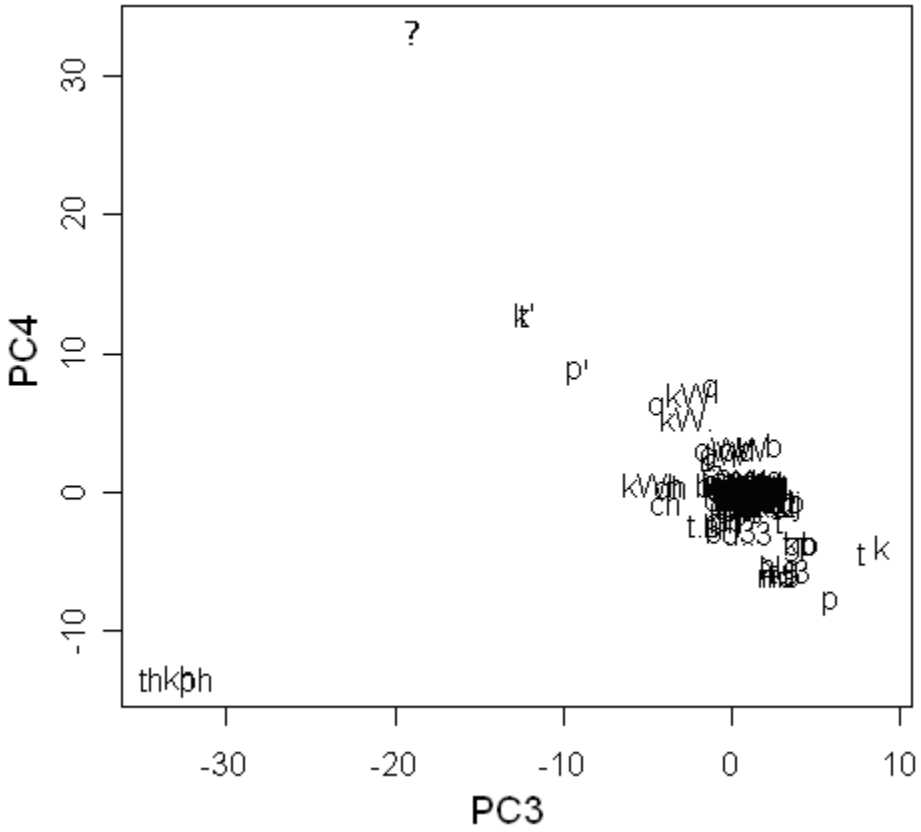


Figure 8. The third and fourth principal components of stop systems.

The distribution of stop systems along the third and fourth principal components and the corresponding geographical distribution are shown in Figure 9. The upper right in Figure 9 is associated with the absence of contrastively aspirated stops, while the bottom left is associated with their presence. Bottom right is associated with the absence of ejectives whereas the top left is associated with their presence. The clusters of languages defined by PC3 and PC4 show some geographical patterning: North American languages are unlikely to occur on the lower left while Southeast Asian languages are concentrated there. European and, to a lesser extent, Australian languages are concentrated on the lower right (cluster 5, not shown).

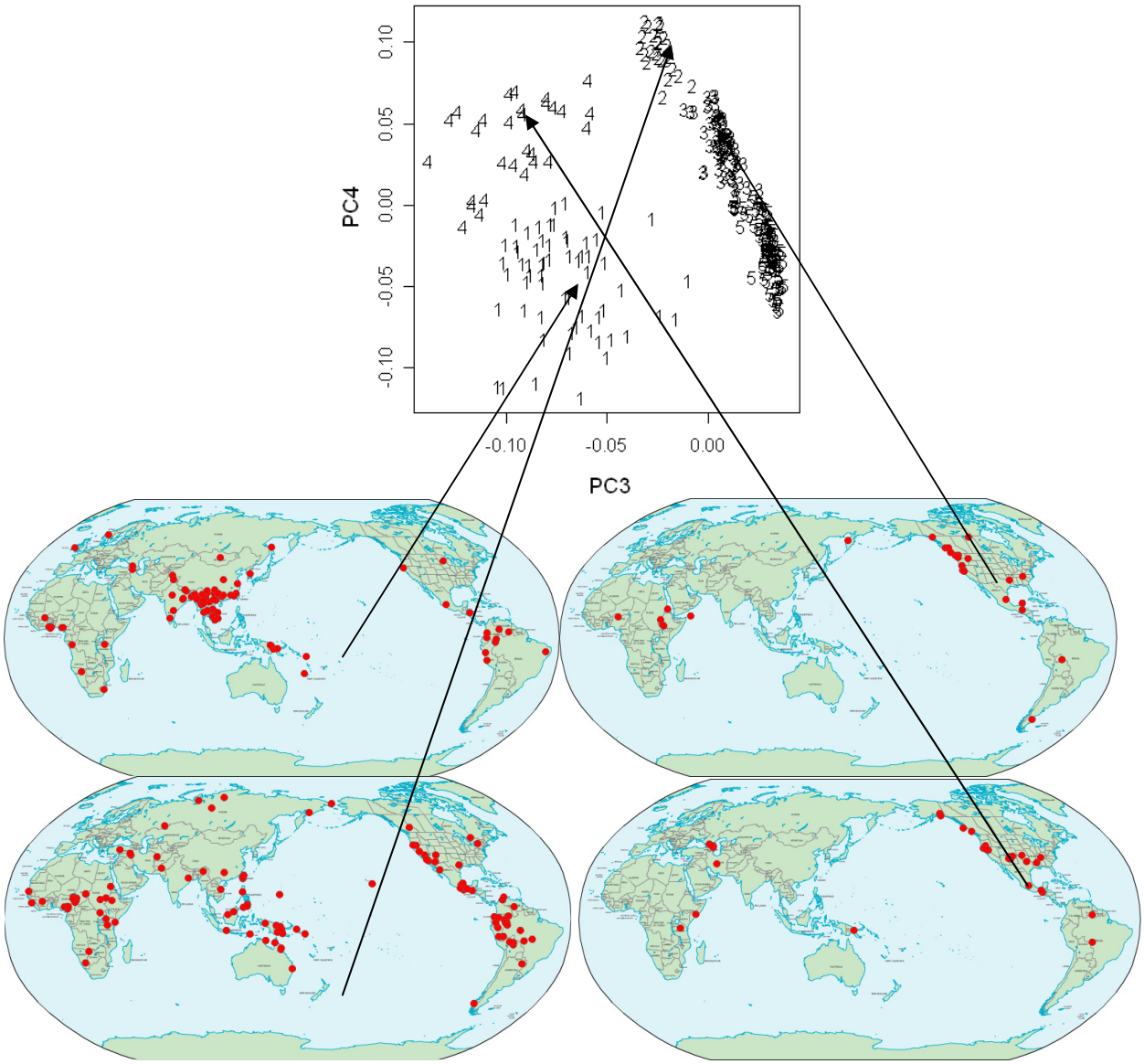
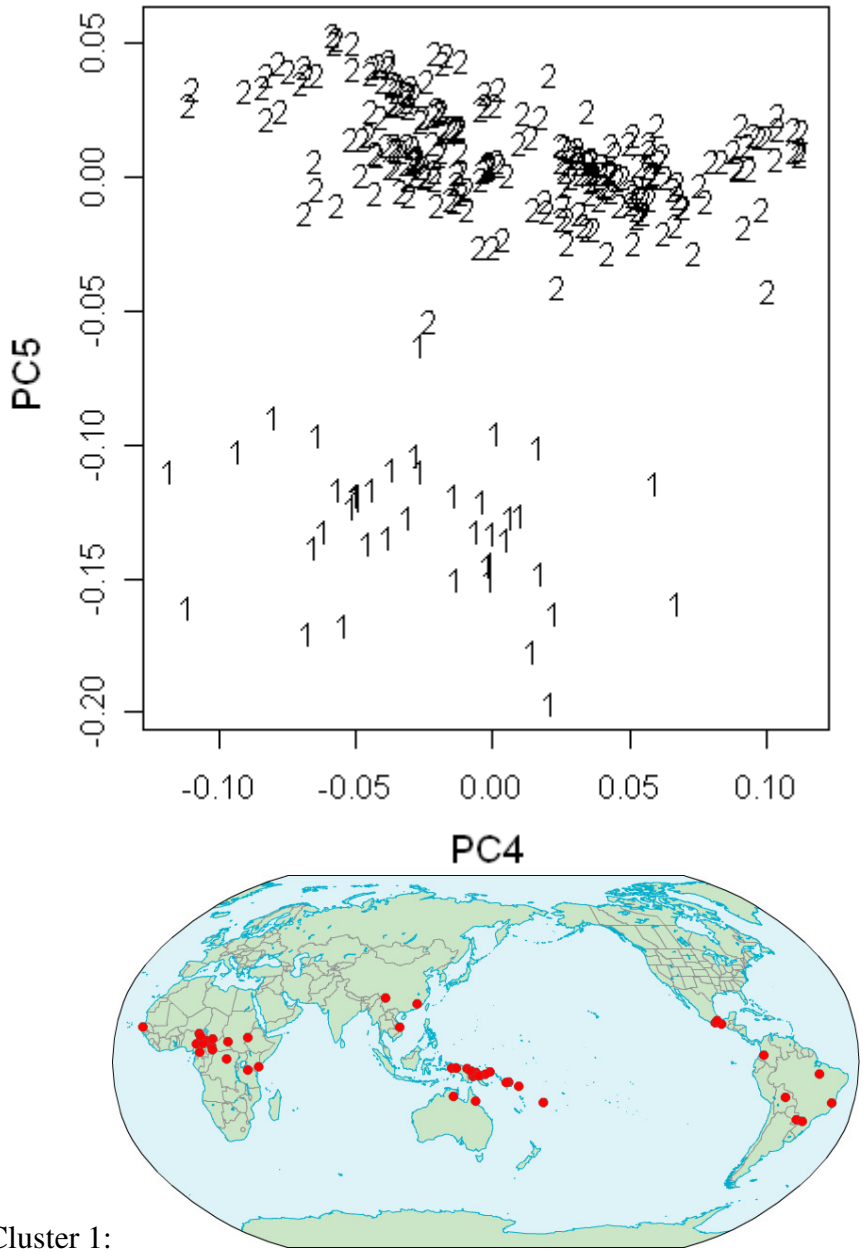


Figure 9. Languages categorized by type of glottal articulation (PC3-PC4)

The smaller cluster extracted by PC5 contains some Niger-Congo languages in Africa and many Papuan languages, as shown in Figure 10, and is distinguished by the presence of voiced prenasalized stops /mb/ and /nd/ (r=-.91).



Cluster 1:

Figure 10. The clusters of stop systems extracted by PC5.

To summarize the results of the present section, the main dimensions of variation among stop consonant systems are 1) voicing, 2) the type(s) of glottal articulations present in the system, and 3) whether the inventory contains prenasalized stops. Aspirated stops and ejectives are distinguished as distinct subtypes of stops accompanied by glottal articulations, as indicated by the fact that they both favor high scores on PC3 but are separated by PC4. Languages fall into distinct clusters along these dimensions, suggesting the presence of attractor states that are more

stable than other states and are thus likely to arise in the course of language change. For instance, a language is unlikely to have just a single ejective or an aspiration contrast at only one place of articulation. Rather, a language either does or does not make use of the ejective feature, resulting in discontinuities in the distribution of sound inventories along the 'ejective-use' dimension.

5. Clustering stops

The present section reports a hierarchical clustering analysis, in which stops were clustered based on similarities in their distribution across inventories. For all the cluster analyses below, a PCA was performed first on the relevant segments-by-languages matrix. For all classes of segments, whether stops, fricatives, sonorant consonants, or vowels, there was a distinct cluster of 'basic' segments, which was separated from the other segments at the very top. Figure 11 shows the cluster of basic stops. The lower the merging point between two nodes or clusters in the dendrogram, the more similar the clusters or nodes are to each other in terms of their distribution across languages. If and only if the merging point is at zero, then the distributions of the two segments are exactly the same. Figure 11 shows that the basic stops separate along the dimension of voicing, rather than place, indicating that languages tend not to differ in which of the basic places of articulation they utilize. This is consistent with the results presented in the previous section where voiced stops of different places of articulation behaved similarly and differently from the voiceless stops. It is interesting that there is an asymmetry between voiced and voiceless segments: while the voiced labial is more 'basic' than the voiced velar and thus patterns more like the coronal, the voiceless velar is more basic than the voiceless labial (cf. Pericliev 2004, who suggests that /g/ implies /b/ but /p/ implies /k/, with some exceptions). Differences in heights of joining indicate that the difference between /b/ and /g/ is smaller than that between /p/ and /k/ while the difference between /k/ and /t/ is smaller than that between /b/ and /d/. Finally, labials and velars are not grouped together as 'peripherals', rather the more basic peripheral is grouped with the coronal.

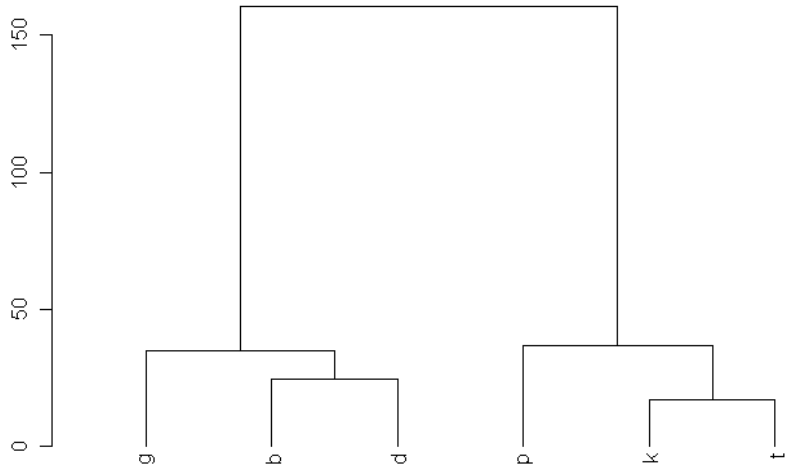


Figure 11. Basic stops

All other segments form a second major cluster, which is in turn separated into two subclusters. The first of these in the case of the stops appears to contain mostly singly elaborated segments (Maddieson 1984) and is shown in Figure 12. The most distinctive group of segments within this cluster are the aspirated voiceless stops at basic places of articulation, which are grouped together with the glottal stop but not the ejectives at a higher level. Another very distinct subcluster are the prenasalized voiced stops.

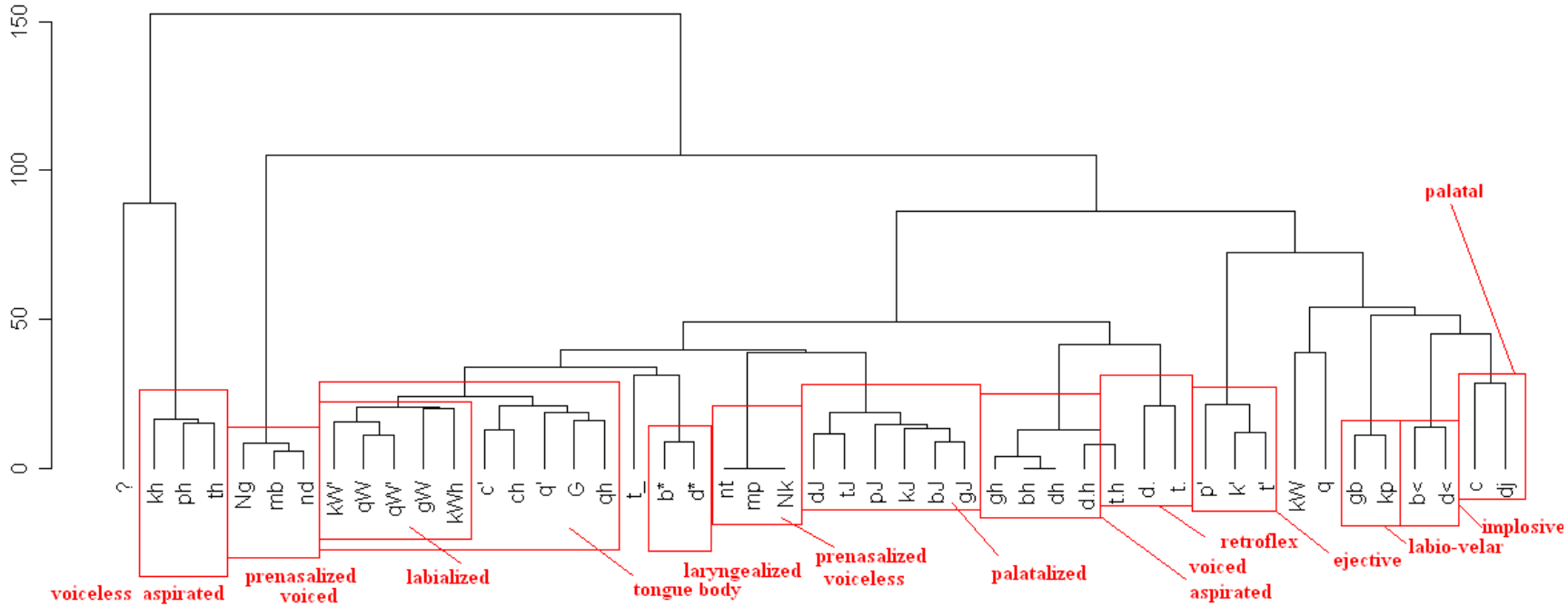


Figure 12. Singly elaborated stops.

Finally, the subclusters of the largest subcluster of the second cluster, which is always the largest and contains the rarest segments are difficult to interpret. Within this subcluster, clustering is not as reliable because of insufficient data for these rare segments: a rare segment by definition occurs in only a few languages, hence the sets of languages containing different rare segments are unlikely to overlap. On the far right of the subcluster, clustering can be seen to produce "chaining", where there is extensive uncertainty about proper grouping, as multiple languages merge at almost the same height. The subcluster is shown in Figure 13.

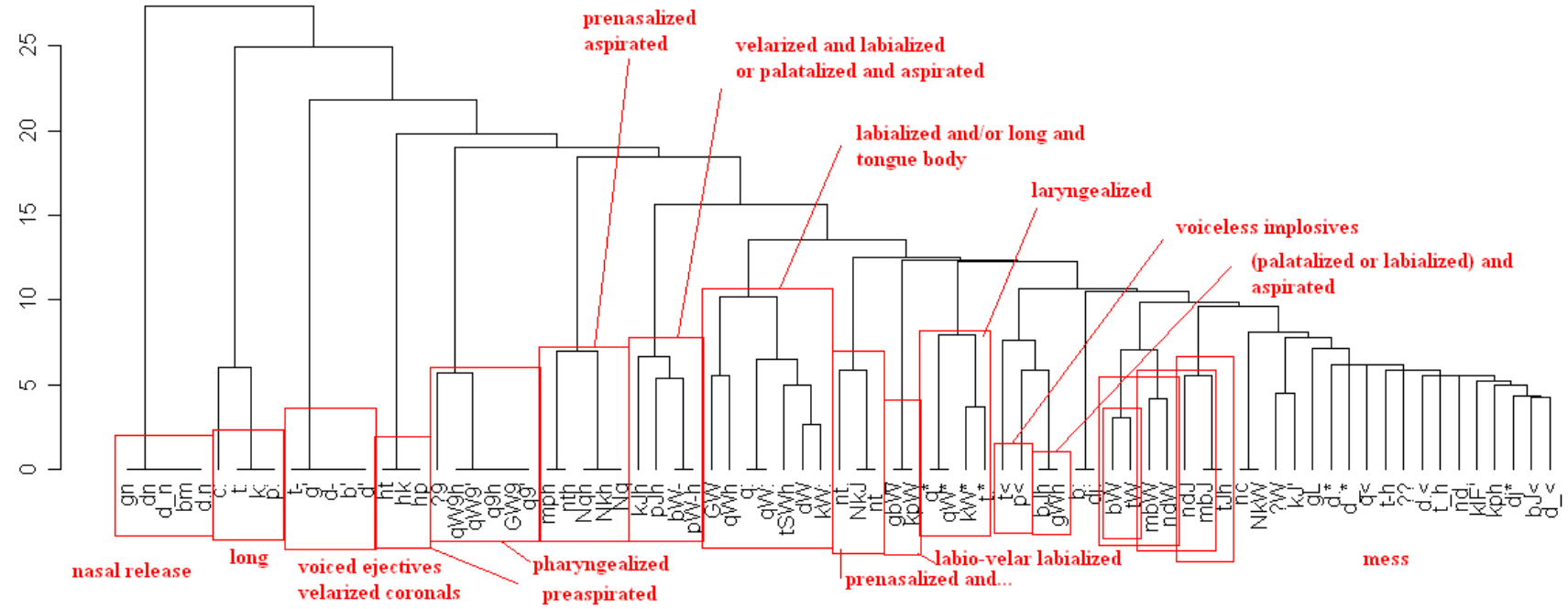


Figure 13. The rare stops.

To summarize the present section, hierarchical clustering is shown to produce featurally interpretable groupings of stops as long as the clustered phonemes are reasonably common. This finding indicates that if a language has one segment with a certain feature, it is also likely to have other segments with the same feature. This result is consistent with Maddieson's (1984) proposal that once a language starts to make use of a feature, that feature is available to be used in multiple segments. However, it is important to note that linking at the zero level is quite rare for common phonemes (Figures 12-13). Almost no two phonemes have the same exact distribution across languages (/b^h/ and /d^h/ and /mp/, /nt/, and /ŋk/ are the only exception in Figures 12-13). Thus, not all combinations of available features must be present in a language.

.

6. Principal components of vowel systems

Like stop systems, existing vowel systems do not appear to be distributed randomly in the space of possible vowel systems (Liljencrantz and Lindblom 1972). In particular, there is clear clustering along principal components 2, 3, 6, and 7. The three clusters of vowel systems along PC2 are shown in Figure 14. High scores on PC2 correlate with the presence of lower mid vowels /ɛ/ and /ɔ/ and the absence of mid vowels /e/ and /o/.

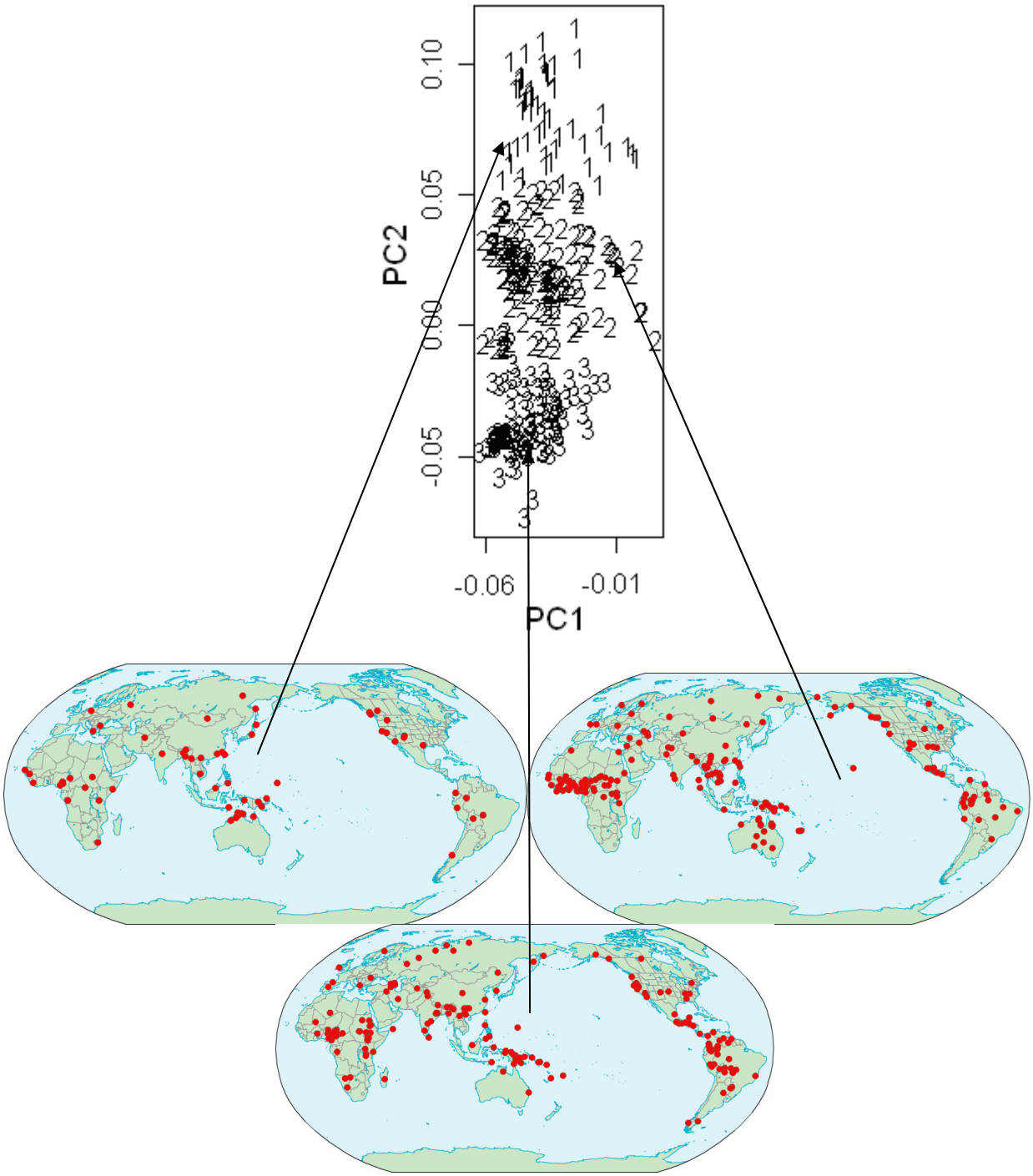


Figure 14. Vowel systems clustering along PC2.

It is possible that the clustering along PC2 is driven by differences between fieldworkers, as opposed to languages since mid and lower-mid vowels are difficult to distinguish and symbols for mid vowels are more readily available in font sets. This interpretation appears to be less likely for clustering seen along PC3, which distinguishes languages based on the presence/absence of distinctively nasalized vowels. The map in Figure 15 shows languages that have contrastive vowel nasalization.

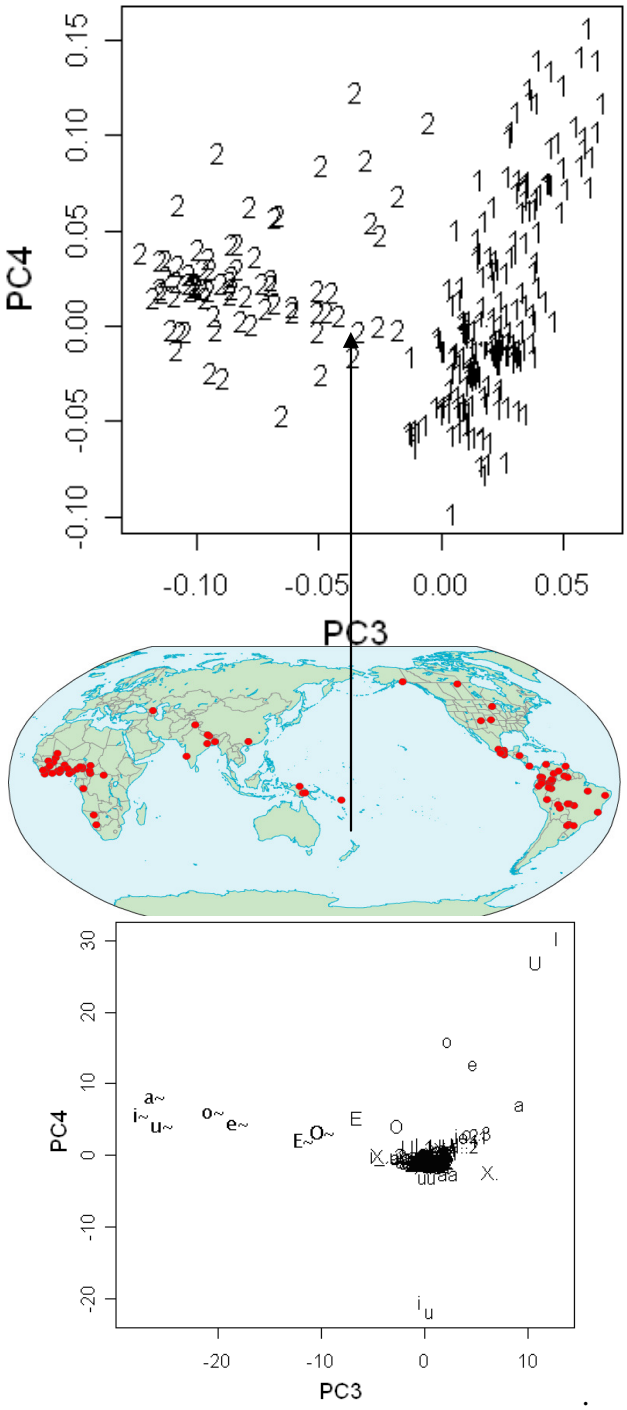


Figure 15. Languages with and without nasalized vowels.

Finally, Figure 16 shows that the 6th and 7th principal components are very sensitive to the presence of /ə/ in the inventory. The clustering of vowel inventories along the 6th and 7th principal components is shown in Figure 17. The upper left in Figures 17-18 (Cluster 1) is associated with the presence of /ə/ and, to a lesser extent, with the presence of /ɨ/ and /ɔ/ and the absence of /a/.

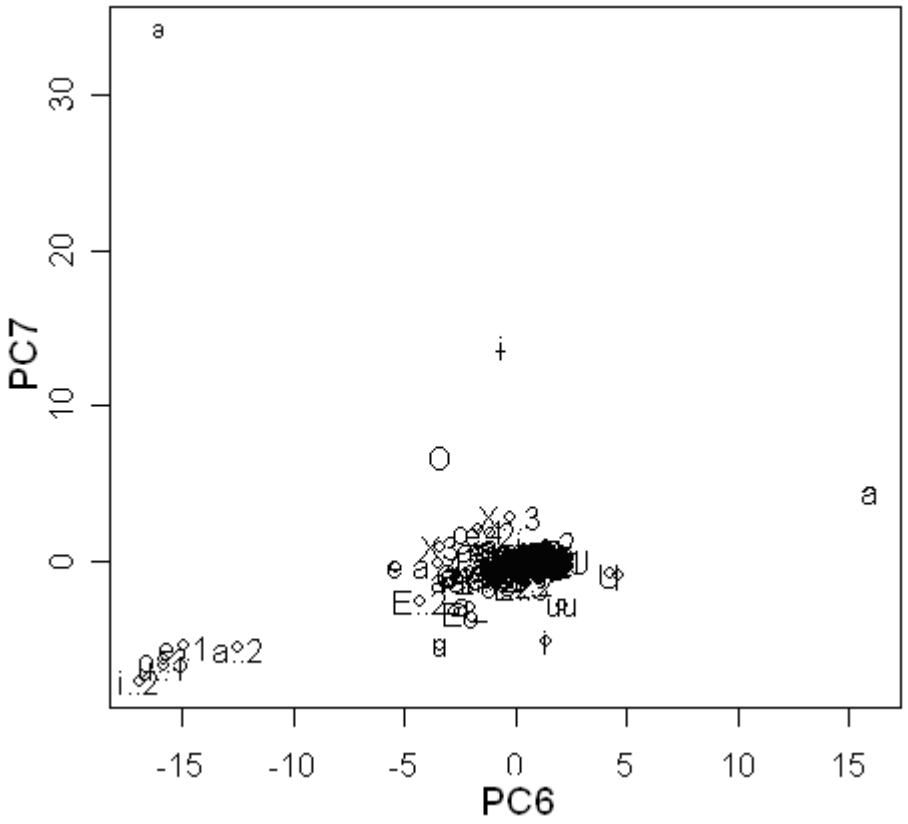


Figure 16. The 6th and 7th principal components defining vowel inventories.

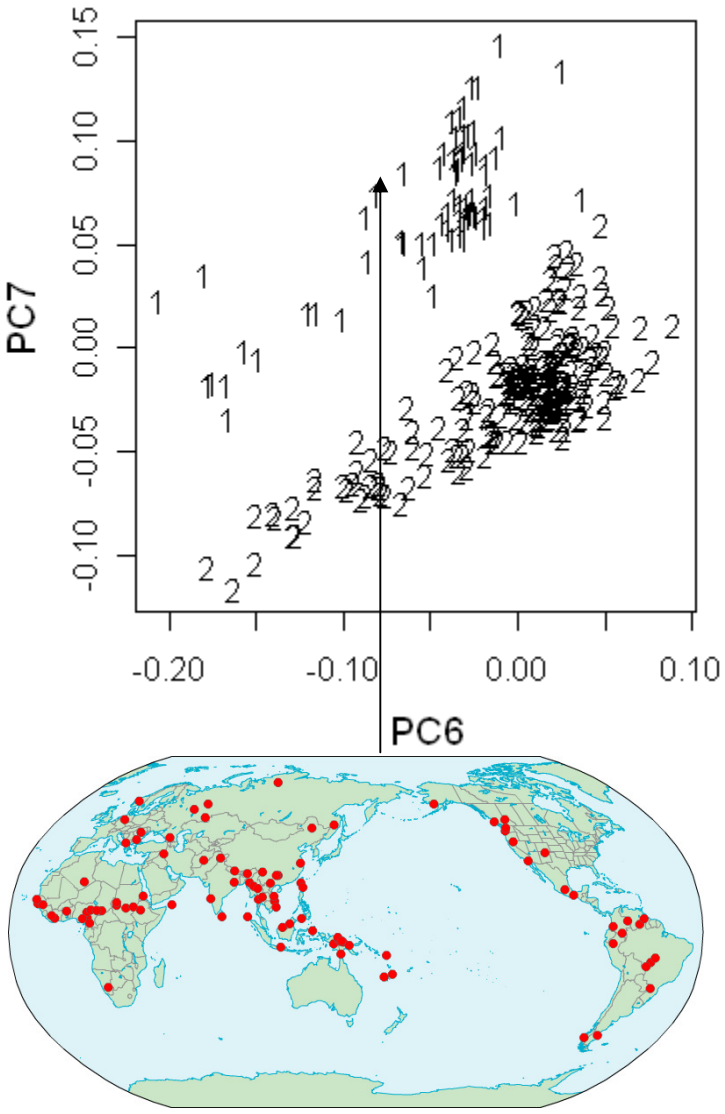


Figure 17. Languages that have /ə/ (and/or /ɨ/ and/or /o/) and/or lack /a/.

To summarize the present section, vowel inventories in UPSID cluster along the following dimensions: 1) whether the inventory contains (mostly) mid or lower mid vowels, 2) whether the inventory contains nasalized vowels, and 3) whether the inventory contains schwa or /a/. Especially in the case of the mid vs. lower-mid distinction, it is possible that the clustering is due to differences in transcription methods between fieldworkers, rather than between languages. Unfortunately, this extraneous source of variance cannot be eliminated in working with a database of grammars composed by a large number of researchers. An additional source of discontinuity in the data is its discrete nature. For instance, a segment inventory either does or does not contain a schwa. If the presence of schwa is assigned a strong weight in determining the location of a language in space, this categorical difference can split languages into two distinct clusters, as seen in the PC6-PC7 space.

7. Clustering vowels

Like in the case of the stops, the first cluster is formed by phonemes that are most frequent across languages. Figure 18 shows that these include the cardinal vowels /i/, /a/, and /u/, which are more frequent than the second subgroup comprising the peripheral mid vowels /e/ and /o/. It is interesting that none of these most common vowels are central, preferring to be located at the extremes of the vowel space in the front/back dimension. It is also interesting that /o/ and /e/ are grouped together, as do /ɪ/ and /ʊ/, and /ɛ/ and /ɔ/ in Figure 19. Thus, vowels seem to come in front-unrounded/back-rounded pairs, suggesting systems that are symmetrical in the front/back dimension and maximize differences in F2 (cf. Liljencrantz and Lindblom 1972).

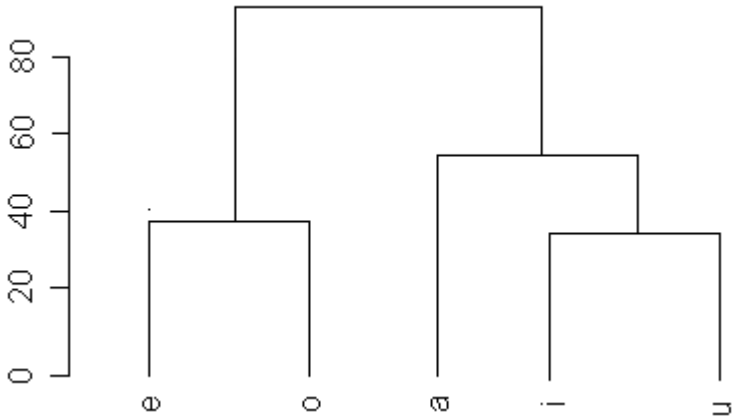


Figure 18. The basic vowels.

Figure 19 shows the first subcluster of the second cluster of vowels. It is interesting to note that the cardinal vowels cluster together and separately from the mid vowels regardless of length and nasality of the vowel, as well as of the presence/absence of frication and laryngealization. Thus, the same pattern of clustering is observed for the basic vowels in Figure 18 as well as for 'nasalized basic' and 'long basic' vowels in Figure 19, fricated vowels in Figure 21 and laryngealized vowels in Figure 22.

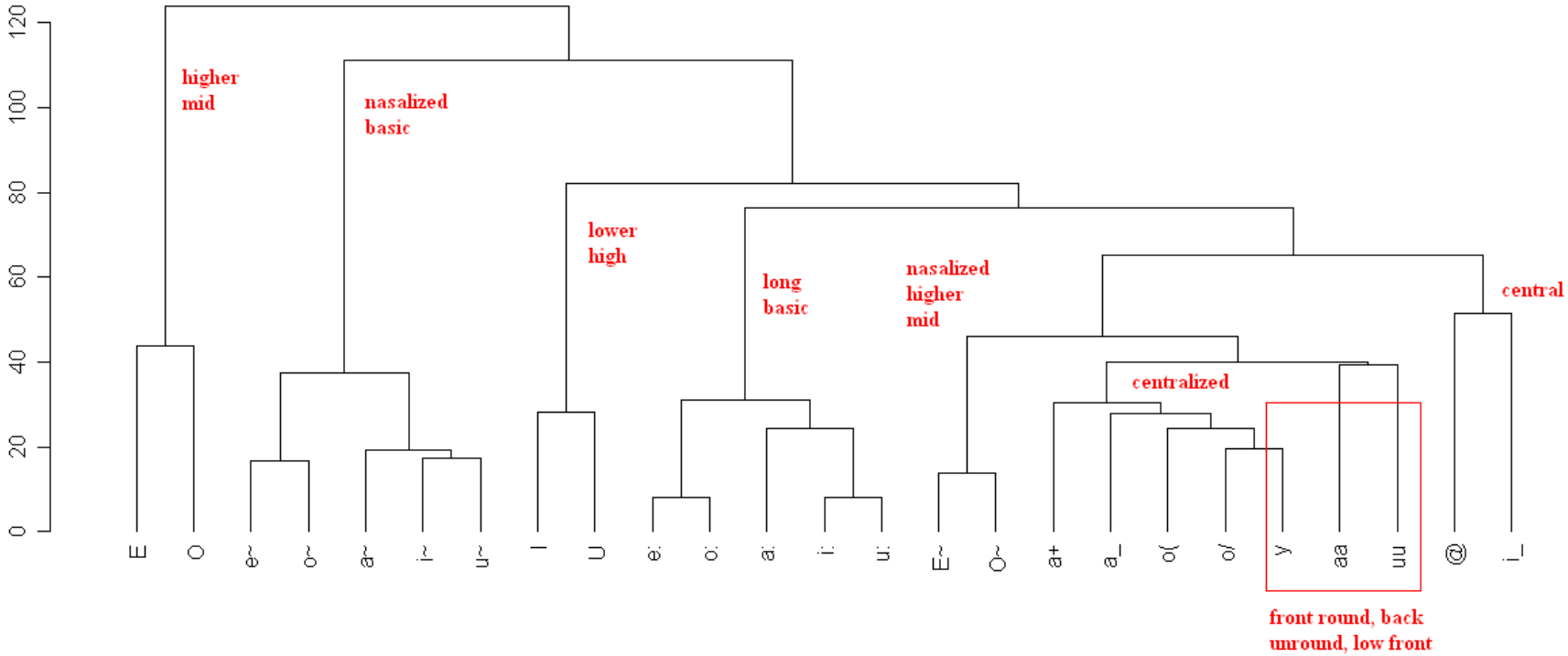


Figure 19. Second cluster of vowels.

The last subcluster of vowels, contains several distinct groupings of rare vowels. The most distinctive subcluster are short voiceless vowels, shown in Figure 20, which are not joined with the rest of the cluster until a height of 60. Other distinctive clusters are fricated vowels (Figure 21), laryngealized vowels (Figure 22), pharyngealized vowels (Figure 23), long vowels that are nasalized or are at a non-basic height (Figure 24), nasalized vowels in non-basic locations in the height/frontness space (Figure 24), fricated non-cardinal vowels (Figure 25) and nasalized laryngalized vowels (Figure 25).

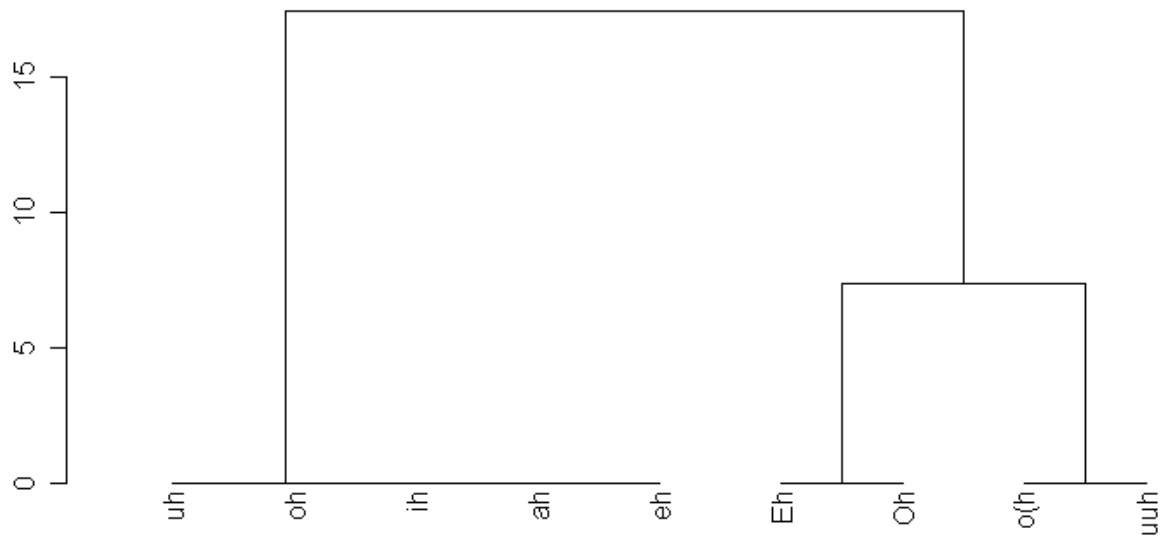


Figure 20. Voiceless vowels

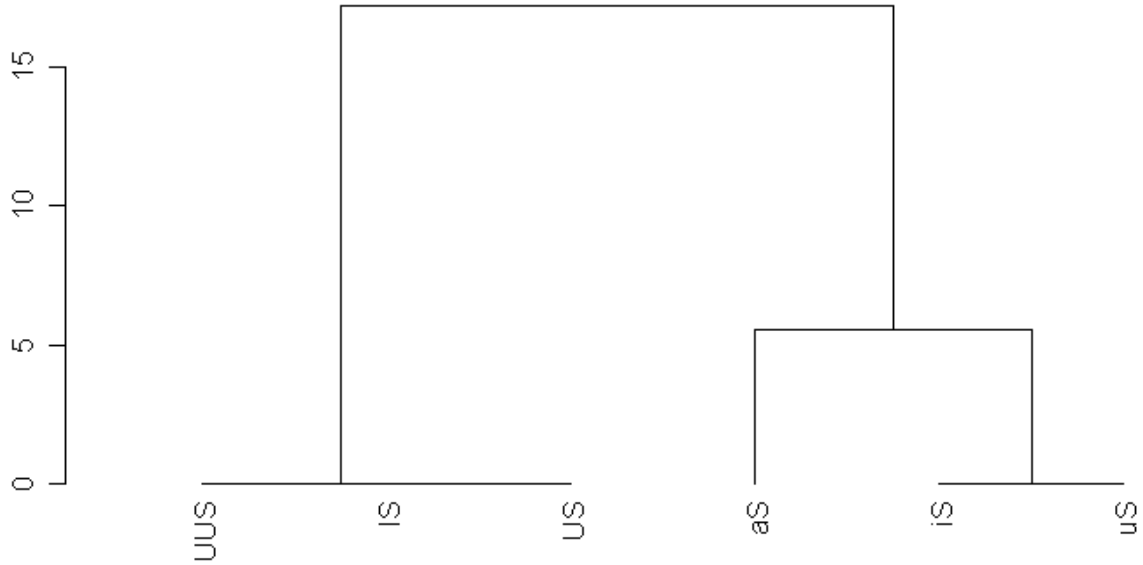


Figure 21. Fricated vowels.

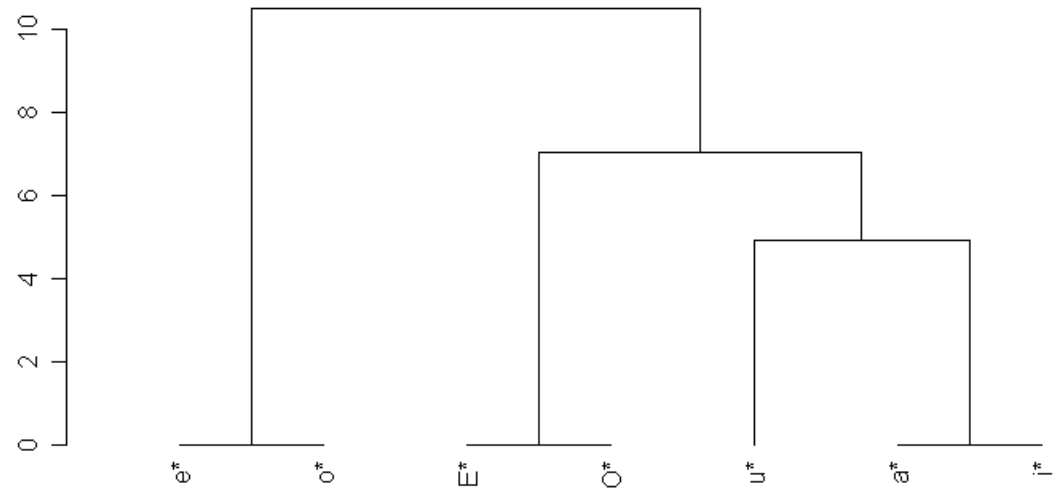


Figure 22. Laryngealized vowels

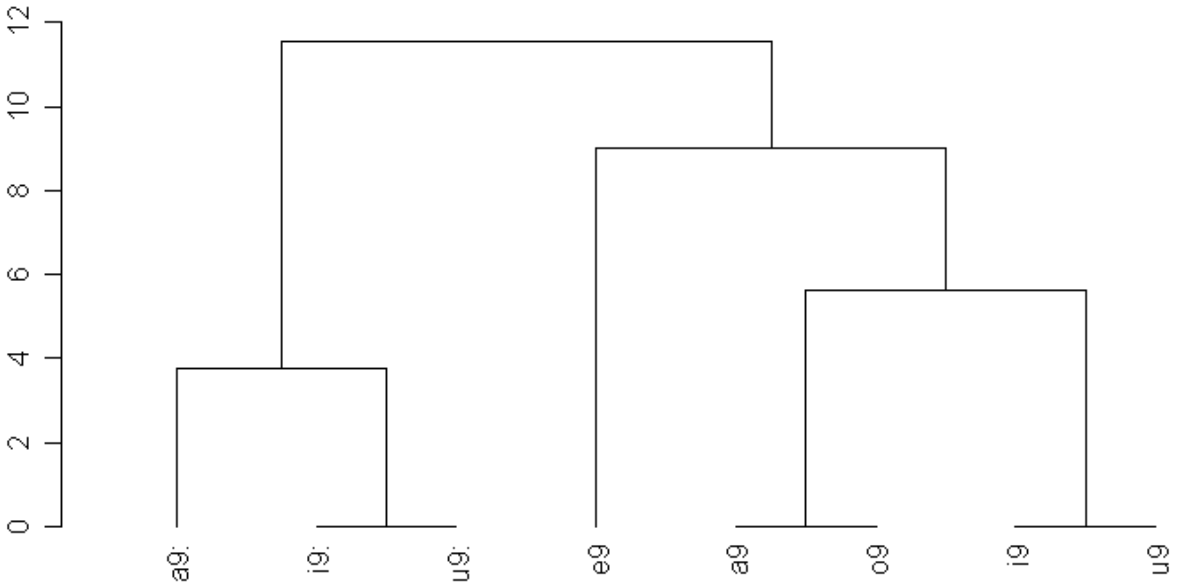


Figure 23. Pharyngealized vowels.

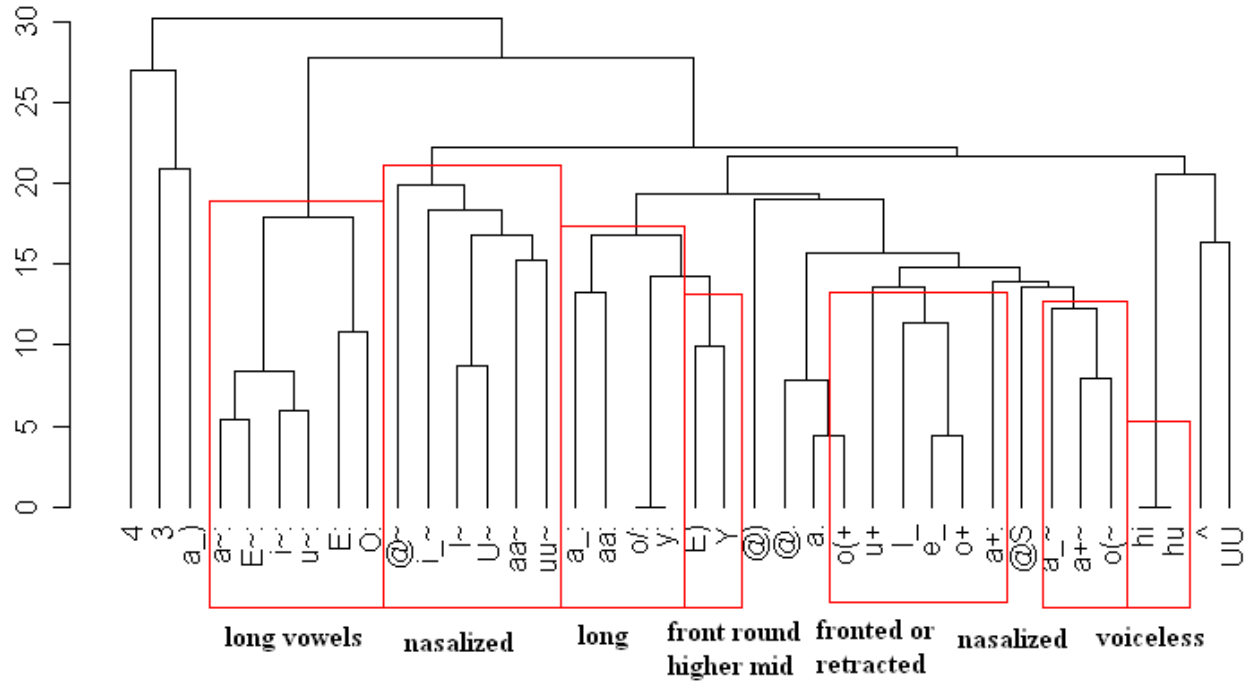


Figure 24. Almost the rarest vowels.

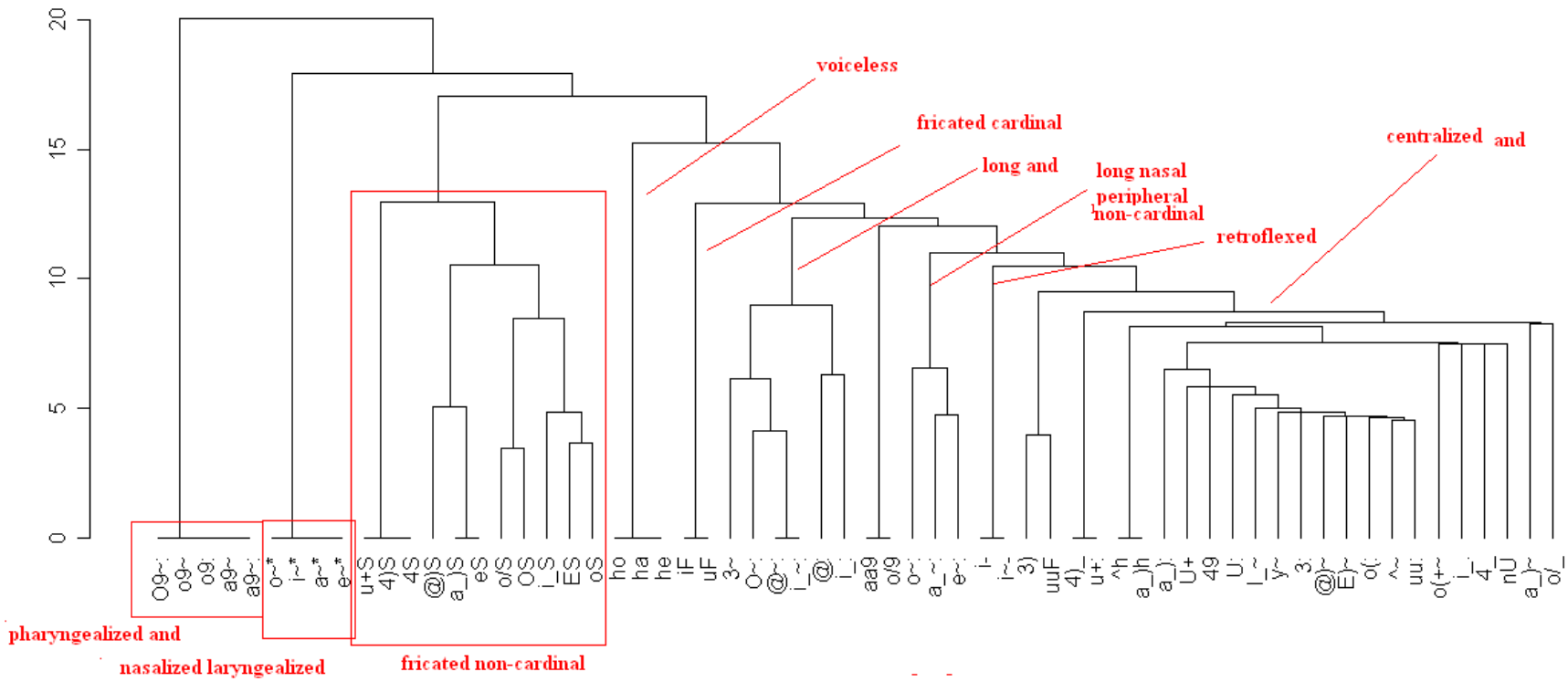


Figure 25. The rarest vowels.

Thus the clustering of vowels can be interpreted quite easily in terms of features coupled with the dimension of typological frequency (as long as the to-be-analyzed vowels are reasonably frequent). The most robust pattern in the clustering of vowels is that cardinal vowels cluster together and are distinct from non-cardinal vowels.

7. Non-plosive consonants.

Interestingly, fricative/affricate sets and sets of sonorant consonants do not appear to separate into distinct groupings the way that stop and vowel inventories do. The fricative/affricate and sonorant consonant inventories are distributed approximately normally in the space of possible inventories defined by principal components. Thus, it is not clear that there are attractor states that systems of non-obstruent consonants tend towards. Given this observation, the status of such consonant sets as systems of interacting elements is questionable. The clustering of fricatives/affricates and sonorant consonants is also more difficult to interpret in featural terms. Thus, I do not present separate analyses for non-plosive consonants.

8. A hierarchy of languages

In this last results section, I present a hierarchical clustering analysis of phoneme inventories, which provides a typological classification for the languages of the world based on the phonemes they utilize. Figure 26 provides the locations of the clusters that are shown in the following figures. As can be seen, clusters 2, 3, and 4 are joined together at a higher level into a single cluster that is very distinct from cluster 1. Languages within cluster 1 are grouped together at a rather low level, suggesting a high level of similarity between the phoneme inventories, while languages in the second major cluster seem more heterogeneous. This is explained by the fact that languages in the second major cluster tend to have larger phoneme inventories and thus are less likely to share a high percentage of phonemes. However, clusters within the second major cluster are also more interpretable because, given the large number of rare phonemes featured by Cluster 2 languages, phonemes occurring within a cluster are unlikely to occur outside of that cluster.

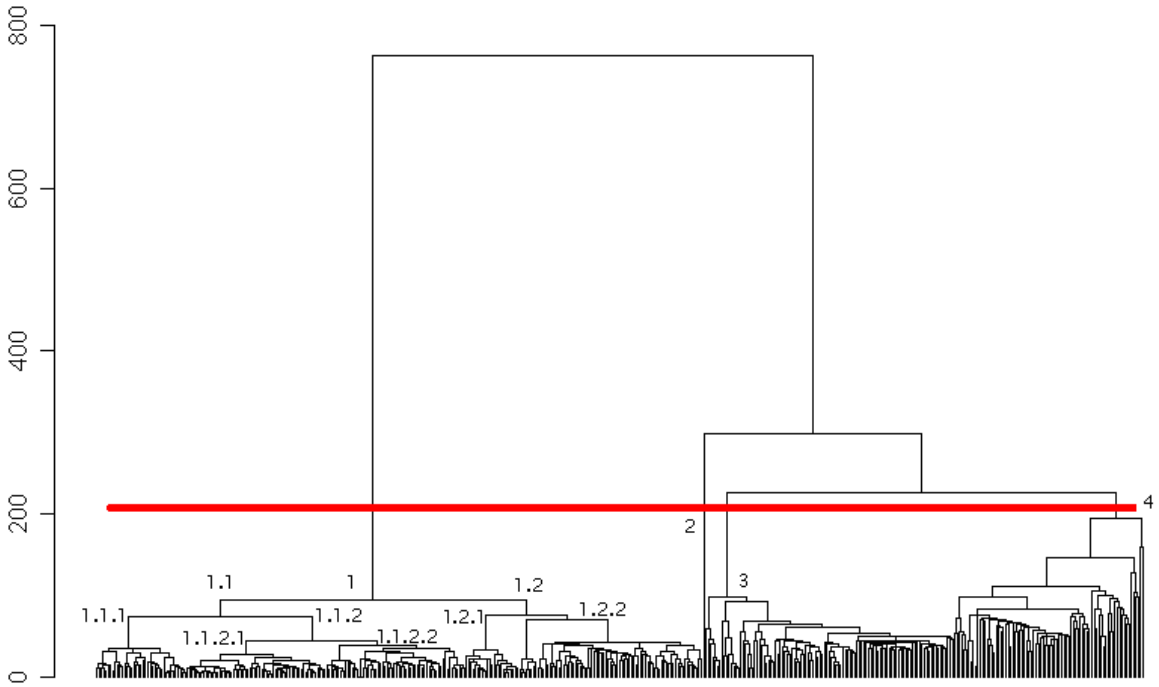


Figure 26. The clustering of the world's languages in terms of similarities between phoneme inventories. Numbers correspond to numbers of clusters presented in the following figures and are shown above the clusters they refer to (at the merging point of the cluster).

The resulting clusters have what Wittgenstein (1953) called a family resemblance structure. Languages in a given cluster may not be in the same cluster because they share some (rare) segments but rather because each of them shares some (rare) segments with some third language in the same cluster. The presence/absence of a particular phoneme rarely correlates perfectly with membership in a particular cluster. Two of the rare exceptions in the present dataset are the CHANGZHOU-JAVANESE cluster in Figure 35, where the two languages in the cluster are the only languages that have obstruents with a distinctively breathy release, and the SHILHA-TAMASHEQ cluster (Figure 35) where the three languages in the cluster share certain pharyngealized obstruents and are the only languages to have the sounds. The more typical case can be illustrated by the example of the ACOMA-NAXI cluster shown in Figure 27 below. Membership in the cluster correlates most with the presence of the voiceless aspirated palatal sibilant affricate (r=.57), which occurs only in GELAO, LAI and NAXI, and the retroflexed mid central unrounded vowel (r=.57), which occurs only in MANDARIN, NAXI and GELAO. No segments whose presence is significantly associated with membership in the cluster are present in all of the languages within the cluster. In fact, the only two segments present in all of the languages within the cluster are /m/ and /u/, neither of which is significantly associated with membership in the cluster. Figure 27 shows, however, that there is a network of family resemblances between the sound inventories of the languages in question.

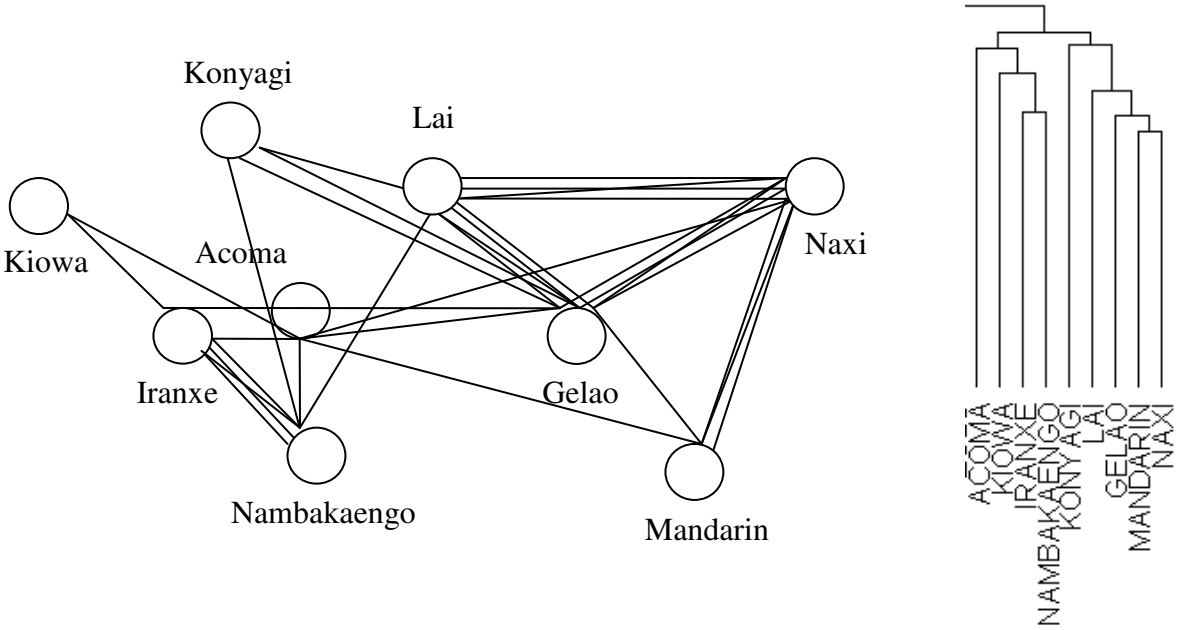


Figure 27. Connections between languages in the ACOMA-NAXI cluster. A connection indicates that the connected languages share a segment whose presence is significantly correlated with membership in the cluster.

In Figure 27, there is no distinct intermediary that connects together two disparate language classes. A somewhat different case is shown by the NENETS-RUSSIAN cluster in Figure 28. Here, almost all the languages within the cluster, except PO-AI, share palatalized consonants, including /r^{j}/, which is only present within the cluster. PO-AI does not have palatalized consonants and is connected to the cluster by virtue of sharing overshort vowels with NENETS. Thus, NENETS acts as intermediary between PO-AI and the other languages in the cluster,

which would otherwise not be connected to PO-AI. It is clear that the resulting connection between PO-AI and, say, RUSSIAN is an accidental one and is much weaker than the connection between NENETS and RUSSIAN, which is not reflected in the clustering diagram. Thus, the clustering diagram has to be interpreted with caution by examining the specific reasons for the clustering decisions made by the algorithm.

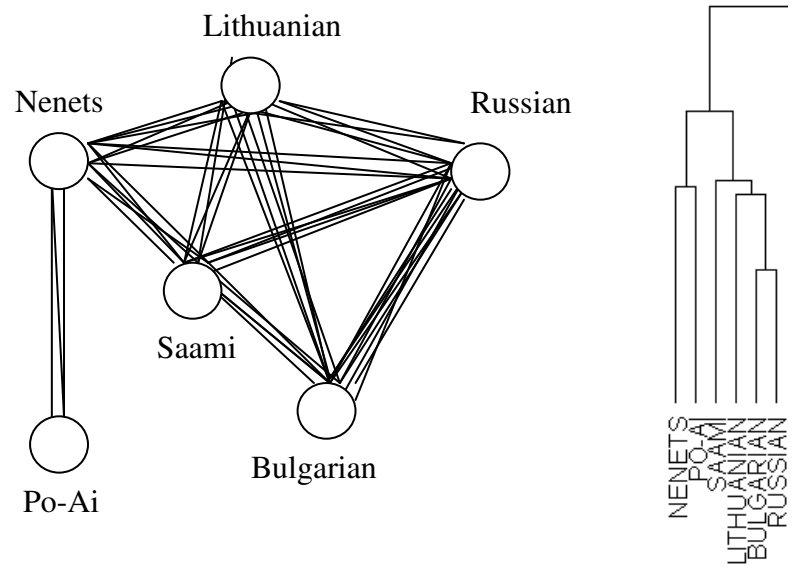


Figure 28. Connections between languages in the RUSSIAN-NENETS cluster.

An imperfect correlation between cluster membership and the presence of a particular segment can have three possible interpretations. First, the segment may occur only within the cluster but may not occur in all of the languages in the cluster. This is exemplified by /r^{j}/ and the overshort vowels in Figure 28. Second, the segment may occur in all languages within the cluster but also occur in some languages outside of the cluster. This is the case for /r^{j}/ in the subcluster of languages including SAAMI, LITHUANIAN, RUSSIAN and BULGARIAN because /r^{j}/ also occurs in NENETS. Third, the languages in the cluster may simply be more likely to have the segments than languages outside of the cluster. This is the case for the correlation between /t^{j}/ and the RUSSIAN-NENETS cluster. The phoneme /t^{j}/ occurs in only 16 languages worldwide, five of them being within the RUSSIAN-NENETS cluster.

In the second case, there may be a subcluster within the cluster that contains all languages that feature the segment but it could also be the case that languages featuring the segment are distributed throughout the cluster without forming a distinct subcluster. For instance, the the retroflexed mid central unrounded vowel can be associated with the MANDARIN-GELAO subcluster in Figure 27, which would include all and only the languages that have the sound. On the other hand, the voiceless aspirated palatal sibilant affricate cannot be associated with a cluster that would contain all the languages that have the sound and only those languages because MANDARIN is categorized as being more similar to NAXI than GELAO and LAI are, despite the fact that only NAXI, GELAO and LAI contain the sound. Palatalized segments are associated with five of the six languages in the NENETS-RUSSIAN cluster but the five languages do not form a distinct subcluster, since NENETS is grouped with PO-AI, rather than with the rest of the inventories containing palatalized segments.

In what follows, if there is a distinct subcluster that contains a segment, then the segment will be reported as being associated with the subcluster, rather than with the larger cluster, whereas if there is no distinct subcluster then the cluster will be associated with the segments. By a "distinct subcluster" I will mean a tree of at least three languages whose root is separated from the closest branch within the subcluster by less than it is separated from the closest branch outside the subcluster. According to the second criterion, the SAAMI-RUSSIAN subtree in Figure 28 is a subcluster of the NENETS-RUSSIAN cluster while the LITHUANIAN-RUSSIAN subtree is not.

It is important to note that the difficulties with interpreting correlations between cluster membership and the presence/absence of specific segments do not indicate that such correlations are meaningless. In particular, if all languages within a cluster share some segment, the presence of the segment is not necessarily the reason the languages are grouped together. If the segment also occurs frequently outside the cluster, its presence in the cluster's languages may well be accidental. If this is the case, there should be no correlation between cluster membership and the segment's presence/absence. Thus, the presence of a significant correlation between cluster membership and segment presence/absence is a necessary but not a sufficient condition for considering the segment's presence or absence an important influence on cluster assignment.

The analyses of clustering patterns reported below are incomplete and should be augmented by analyzing segment-language co-occurrence patterns in greater depth, as shown in Figures 27-28. However, even while the reasons for some of the cluster boundaries remain obscure, it is hoped that the clustering will be helpful as a classification of the world's sound inventories. The clustering is based on an objective definition of similarity that weights phoneme-language co-occurrences in a mathematically appropriate manner giving more weight to phoneme occurrences that are more surprising and taking into account co-occurrence patterns between phonemes. Thus, I would argue that the present clustering extracts much of the information about language relatedness, either areal or genetic, that is present in the phonemically transcribed phoneme inventories of UPSID.

Figures 29-35 suggest that some information about language relatedness is present in sound inventories. Cluster 1.2.1 contains only languages of central and western Africa. All but two are Niger-Congo languages. The exceptions are the Afroasiatic KERA and the Nilo-Saharan LUGBARA that are geographically adjacent to Niger-Congo languages. Cluster 1.2.2 contains the NGARINJIN-YOLNGU subcluster containing exclusively Australian languages and the the Mari-Turkish subcluster, containing only Uralic and Altaic languages (MARI, KIRGHIZ, BASHKIR, TUVA, TURKISH, MANCHU, AZERBAIJANI). Some examples of apparent areal groupings are also present in Cluster 3. Languages of India cluster together, whether they are Indo-European (HINDI, BENGALI, NEPALI), Austro-Asiatic (MUNDARI, KHARIA), Tibeto-Burman (NEWARI), or Dravidian (TELUGU). Another interesting subcluster within Cluster 3 is the cluster of Northwestern American languages (TLINGIT, BELLA COOLA, CHEHALIS, HAIDA, KWA'KWALA, LUSHOOTSEED, NUUCHAHNULTH, and QUILEUTE), which are grouped together whether they are Salishan, Wakashan, Na-Dene, or Haida. Cluster 4 contains the NENETS-RUSSIAN cluster discussed above contains Balto-Slavic Indoeuropean languages and languages in heavy contact with Balto-Slavic languages (NENETS, SAAMI), with the exception of PO-AI.

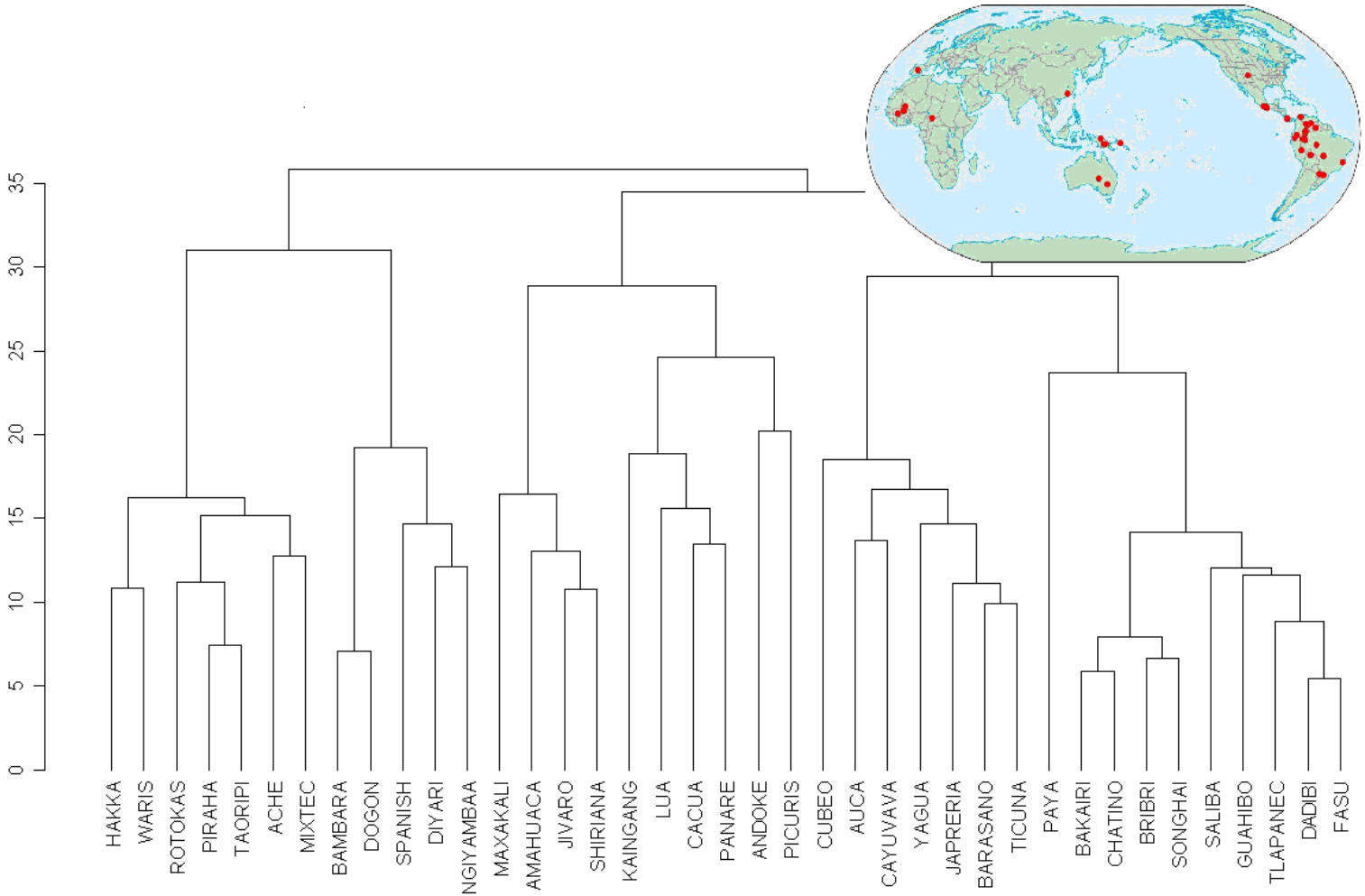


Figure 29. Cluster 1.1.1. Membership in this cluster correlates with the presence of nasal vowels (.22<r<.55 depending on vowel) and the absence of basic nasal consonants /n/ and /m/ (-.40<r<-.25). Languages in the BAMBARA-NGIYAMBAA cluster are associated with the voiced alveolar tap (r=.78). The MAXAKALI-SHIRIANA cluster is associated with the nasalized high back unrounded vowel (r=.75). The KAINGANG-PICURIS cluster is associated with the nasalized higher mid central unrounded vowel (r=.65). All subclusters except for the HAKKA-NGIYAMBAA cluster are associated with nasal vowels. The HAKKA-MIXTEC cluster and, to a much lesser extent, the KAINGANG-PICURIS cluster and the CUBEO-TICUNA cluster are associated with the absence of /n/.

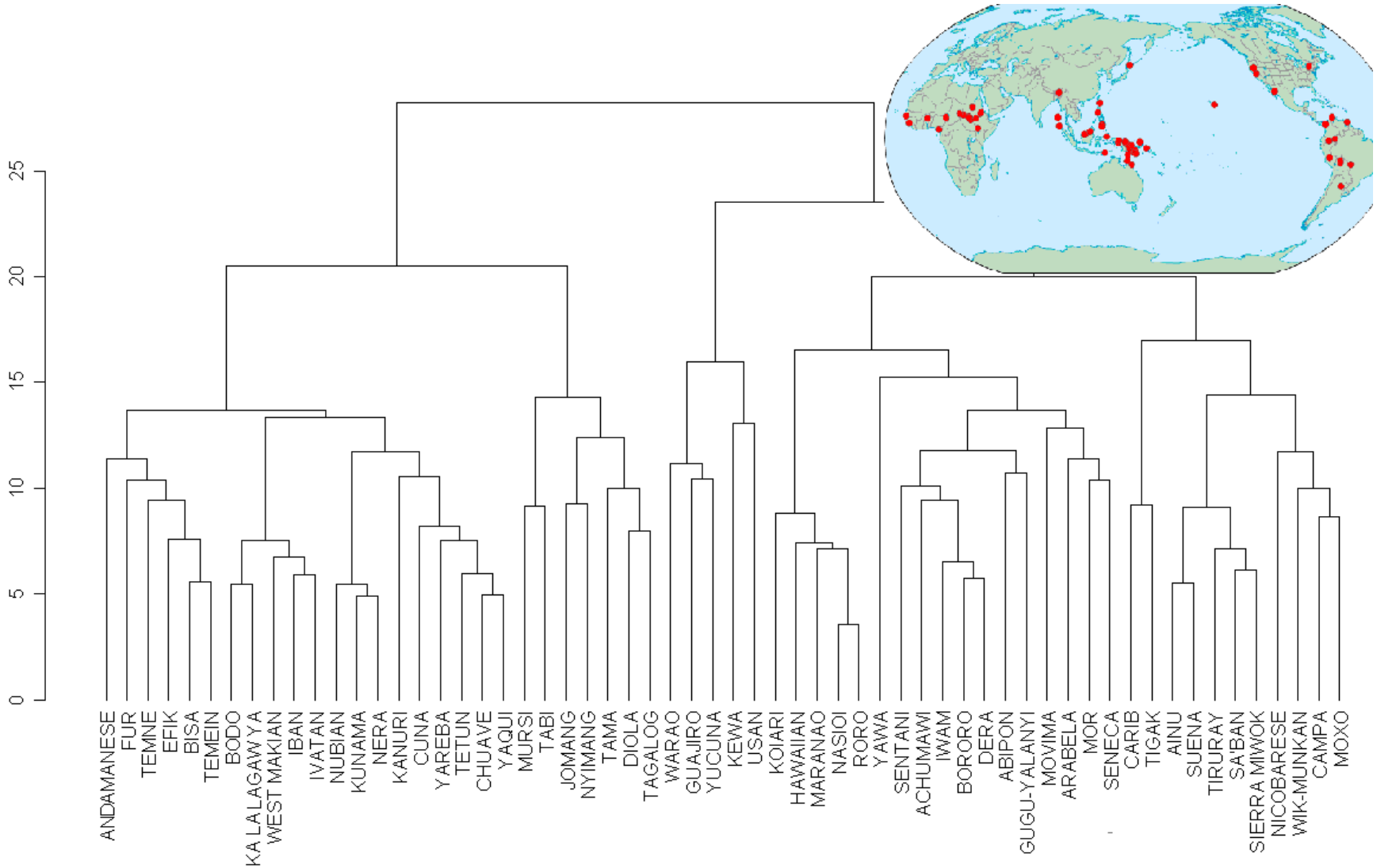


Figure 30. Cluster 1.1.2.1. Membership in this cluster correlates with the absence of basic voiceless aspirated stops and /ts/ (r=-.21). Membership in the WARAO-USAN subcluster is correlated with the presence of the voiced alveolar lateral flap (r=.51).

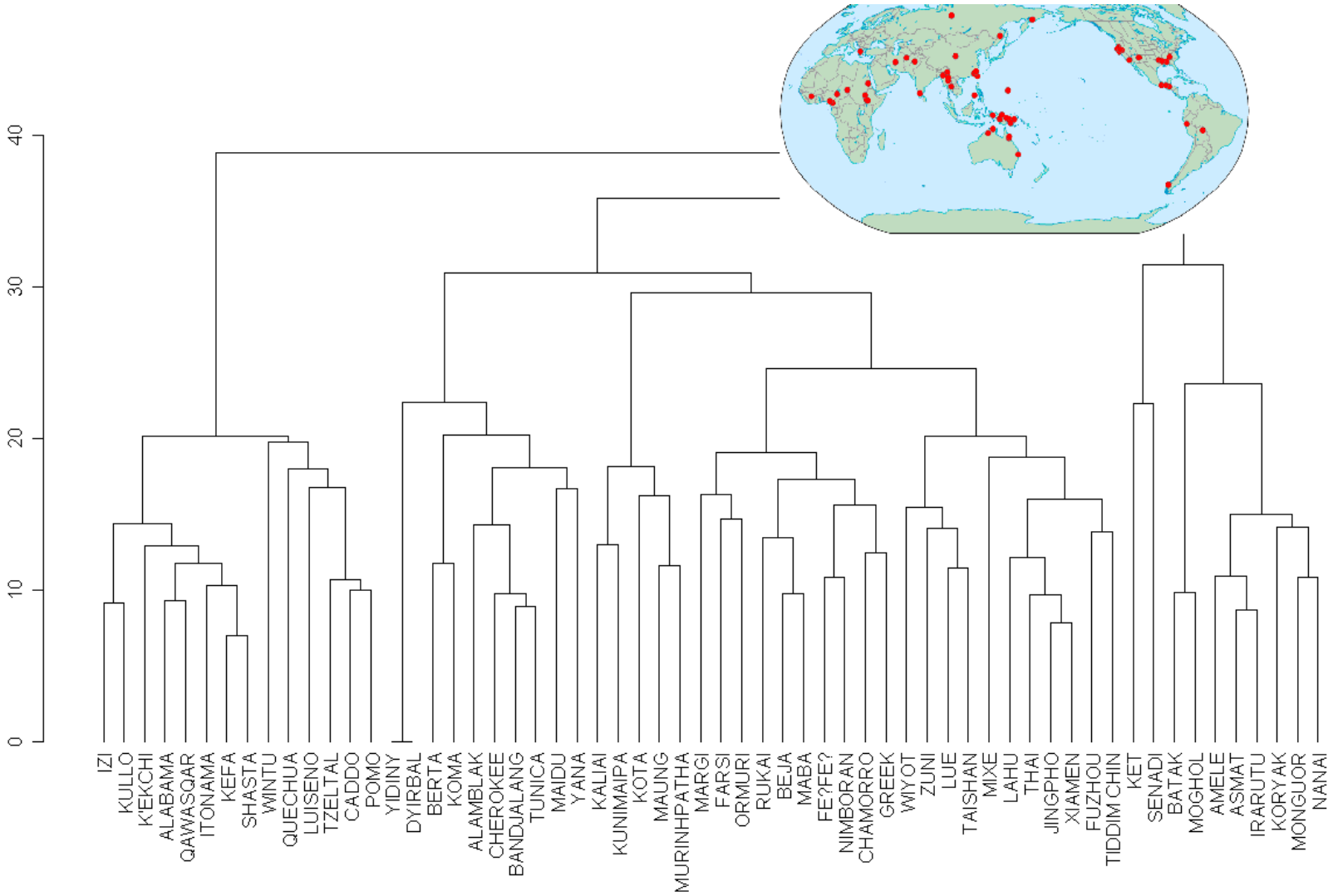


Figure 31. Cluster 1.1.2.2. Membership in the IZI-POMO cluster correlates with the presence of ejectives (.21<r<.36). Membership in the KALIAI-MURINHPATHA cluster correlates with the presence of a voiced retroflex flap (r=.59). Membership in the KET-NANAI cluster correlates with the presence of the voiced palatal fricative (r=.72), the absence of /j/ (r=-.26), and, outside of the AMELE-NANAI cluster, the presence of the voiced uvular trill (r=.59).

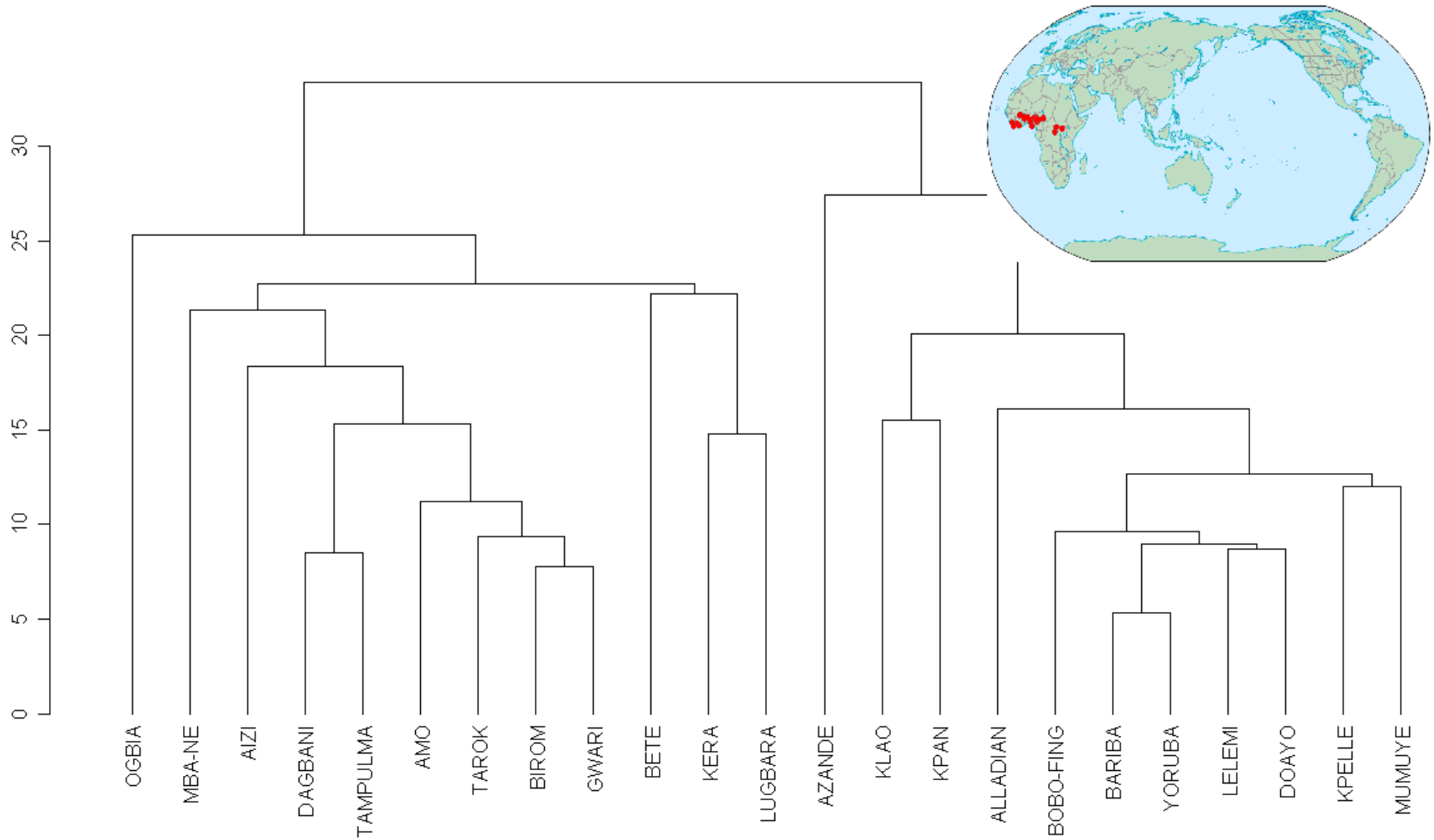


Figure 32. Cluster 1.2.1. Membership in this cluster correlates with the presence of the doubly articulated stops /kp/ and /gb/ (r=.76), the fricative /v/ (r=.37), /f/ (r=.28), and /z/ (r=.29). Membership in the AZANDE-MUMUYE cluster correlates with the presence of lower mid vowels /ɔ/ and /ɛ/ (r=.23) and their nasalized counterparts (r=.35). Membership in the OGBIA-LUGBARA cluster correlates with the presence of the voiced bilabial implosive (r=.21).

Membership in the NGARINJIN-YOLNGU subcluster within cluster 1.2.2, which contains all and only Australian languages in the whole cluster, is correlated with the presence of retroflex and, to a lesser extent, palato-alveolar consonants: the voiced retroflex approximant (r=.68), the voiced retroflex nasal (r=.57), the voiced retroflex lateral (r=.53), the voiceless retroflex stop (r=.47), the voiceless palato-alveolar stop (r=.38), the voiced palatoalveolar lateral (r=.29) and the voiced palatoalveolar nasal (r=.24), as well as with the absence of /s/ (r=-.39). The retroflex segments occur in all languages within the cluster but differ in how frequently they occur outside of the cluster. Thus, the approximant occurs in nine languages outside the cluster while the nasal occurs in 16, accounting for the fact that the approximant is a better predictor of cluster membership. The palatoalveolar stop and nasal occur only within the GARAWA-YOLNGU subcluster within the cluster but also occur in some languages outside the cluster. The palatoalveolar lateral occurs only in the GARAWA-WESTERN DESERT subcluster and in some languages outside the cluster.

Membership in the DANGALEAT-NYANGI cluster, which contains only African languages, correlates with the presence of implosives (labial, alveolar, velar, or palatal, .3<r<.71), particularly the palatal implosive (r=.7), and the voiceless palatal stop (r=.21). The associated implosives occur in all languages within the cluster, except for the velar implosive, which occurs only in the MAASAI-NYANGI subcluster. The voiceless palatal stop occurs in all languages within the cluster except MAASAI.

Membership in the AGHEM-TEKE cluster, consisting of Niger-Cordofanian Bantoid languages, correlates with the presence of the voiced labiodental affricate, which occurs only in this cluster and is present in all languages within it (r = 1). The KANAKURU-SAVOSAVO cluster is associated with the prenasalized voiced palatal plosive (r=.7) as well as other prenasalized stops (r=.3). The associated prenasalized stops occur in all three languages within the cluster. The GADSUP-TOTONAC cluster is associated with long vowels (.3<r<.35). No long vowel occurs in all languages within the cluster. Thus, /o:/ occurs in GADSUP, NONI, DIEGUENO, BRAHUI, and TONKAWA, /i:/ occurs in TIGRE, ADZERA, NONI, BRAHUI, ATAYAL, TONKAWA, and TOTONAC, /u:/ occurs IN GADSUP, TIGRE, NONI, BRAHUI, ATAYAL, TONKAWA, and TOTONAC, /æ:/ occurs in GADSUP and TIGRE, while /a:/ occurs in ADZERA, NONI, DIEGUENO, TONKAWA, and TOTONAC. However, every language in the cluster except COFAN has long vowels. COFAN is grouped together with ADZERA and NONI because of the presence of a voiced velar approximant in these languages.

The SINHALESE-KOMI cluster is associated with the presence of palatal affricates (r=.67). The voiceless palatal affricate occurs in all languages in the cluster except KOYA and TULU. The voiced counterpart occurs in ALBANIAN, HUNGARIAN, IATE, KLAO, KOMI, and SINHALESE. Both also occur outside of the cluster with the voiced affricate being less frequent overall. The BASQUE-TSOU cluster is associated with retroflex fricatives and the voiceless retroflex affricate (.4<r<.62). The voiceless affricate and fricative occur in all languages within the cluster while the voiced fricative occurs only in the PAIWAN-TSOU subcluster. The EJAGHAM-HIXKARYANA cluster is associated with the presence of a voiced palatoalveolar stop (r=.7), which is present in all languages in the cluster. The PASHTO-PAPAGO cluster is associated with the voiced retroflex lateral flap, which occurs in all

languages within the cluster (r=.7). Membership in the MARI-TURKISH cluster, which contains Uralic and Altaic languages, correlates with the presence of the front rounded vowels, particularly the mid front vowel (r=.68), which occurs in all languages within the cluster except TURKISH. The AZERBAIJANI_TURKISH subtree is distinguished by the lowering of the high front rounded vowel. A high front rounded or lowered high front rounded vowel is present in all languages within the cluster except MANCHU.

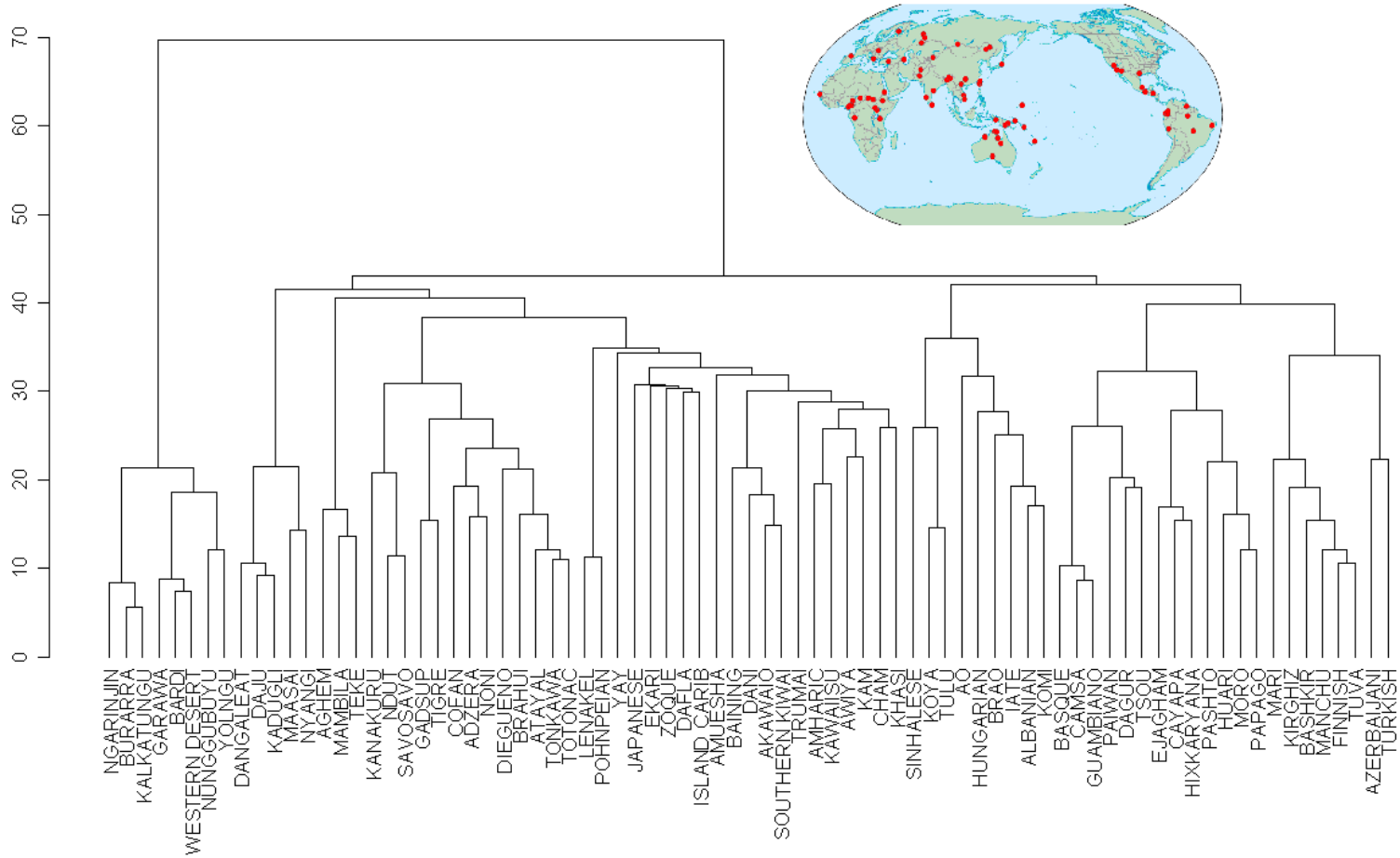


Figure 33. Cluster 1.2.2.

Cluster 2 is formed by a single language, !XU, which has a very large number of phonemes. Within Cluster 3, the HINDI-MUNDARI cluster, which contains languages of India, is associated with breathy voiced stops (.82<r<.93), the retroflex stops (r=.73 for the voiceless aspirated, r=.43 for the voiced, r=.37 for the plain voiceless), voiced aspirated affricates (r=.56 for the palatoalveolar, r=.43 for the alveolar) and the breathy lateral and nasals (.36<r<.44). The breathy voiced retroflex stop occurs only in languages in this cluster and occurs in all languages of this cluster except for NEWARI, which is the only language in the cluster that lacks retroflexes. Interestingly, this phoneme, which as both featural characteristics associated with the cluster (retroflex and breathy voiced stop) is the phoneme associated with the cluster most strongly. The breathy voiced bilabial and velar stops occur in all languages of the cluster but also in IGBO, PARAUK and, in the case of the velar, !XU outside the cluster. The associated retroflex segments occur in all languages within the cluster except NEWARI.

The TLINGIT-CHEHALIS cluster, containing Northwest American languages, is associated with the presence of labialized uvulars (.35<r<.85) and ejectives, especially uvular and velar ejectives and the voiceless alveolar lateral ejective affricate. Like in the case of the HINDI-MUNDARI cluster, this cluster is most strongly associated with a segment that features all of the associated characteristics, a labialized voiceless uvular ejective stop, which occurs in all languages within the cluster and only in LAK and RUTUL outside of it. The top two ejectives and the top two labialized uvulars are present in all languages within the cluster.

The PAEZ-SANGO cluster, containing languages from a rather small area of western Central Africa, except for PAEZ, is associated with prenasalized and doubly articulated obstruents (.28<r<.84). Like the two previous clusters, this cluster is associated most strongly with the consonant that has all of the associated features, the prenasalized voiced labial-velar plosive (r=.83), which occurs in all languages in the cluster except PAEZ and in MBA-NE outside the cluster. PAEZ, which, unlike the rest of the languages in the cluster, is South American, lacks the two strongest correlates of the cluster, /mv/ and /ŋmgb/, although other prenasalized obstruents are present. The marginal status of PAEZ within the cluster is reflected in the clustering solution since it is the last language to attach to the cluster.

The SUI-MIEN cluster is associated with the presence of voiceless nasals (.46<r<.73). The voiceless bilabial nasal is present in all languages within the cluster. The ARAUCANIAN-BURUSHASKI cluster is associated with the presence of the voiced retroflex fricative (r=.81), which occurs only in ARAUCANIAN, ARMENIAN, BURUSHASKI, and CHUKCHI, and the overshort mid central unrounded vowel (r=.6), which occurs only in ANGAATIHA, CHUKCHI, GEORGIAN, and SEBEI, the first three languages being inside the cluster. The NAVAJO-BATS cluster is associated with the dental/alveolar lateral affricates (r=.51 for the voiced, r=.4 for the aspirated, r=.32 for the ejective). The overarching YUCHI-JAQARU cluster is associated with the presence of ejectives, especially /t'/ (r=.41). The KWOMA-MAZATEC cluster is associated with prenasalized obstruents, especially the labialized prenasalized voiced velar stop (r=.59), which is present in KWOMA, all languages of the MBABARAM-KWAIO subcluster, which forms the core of the larger cluster, and NGIZIM.

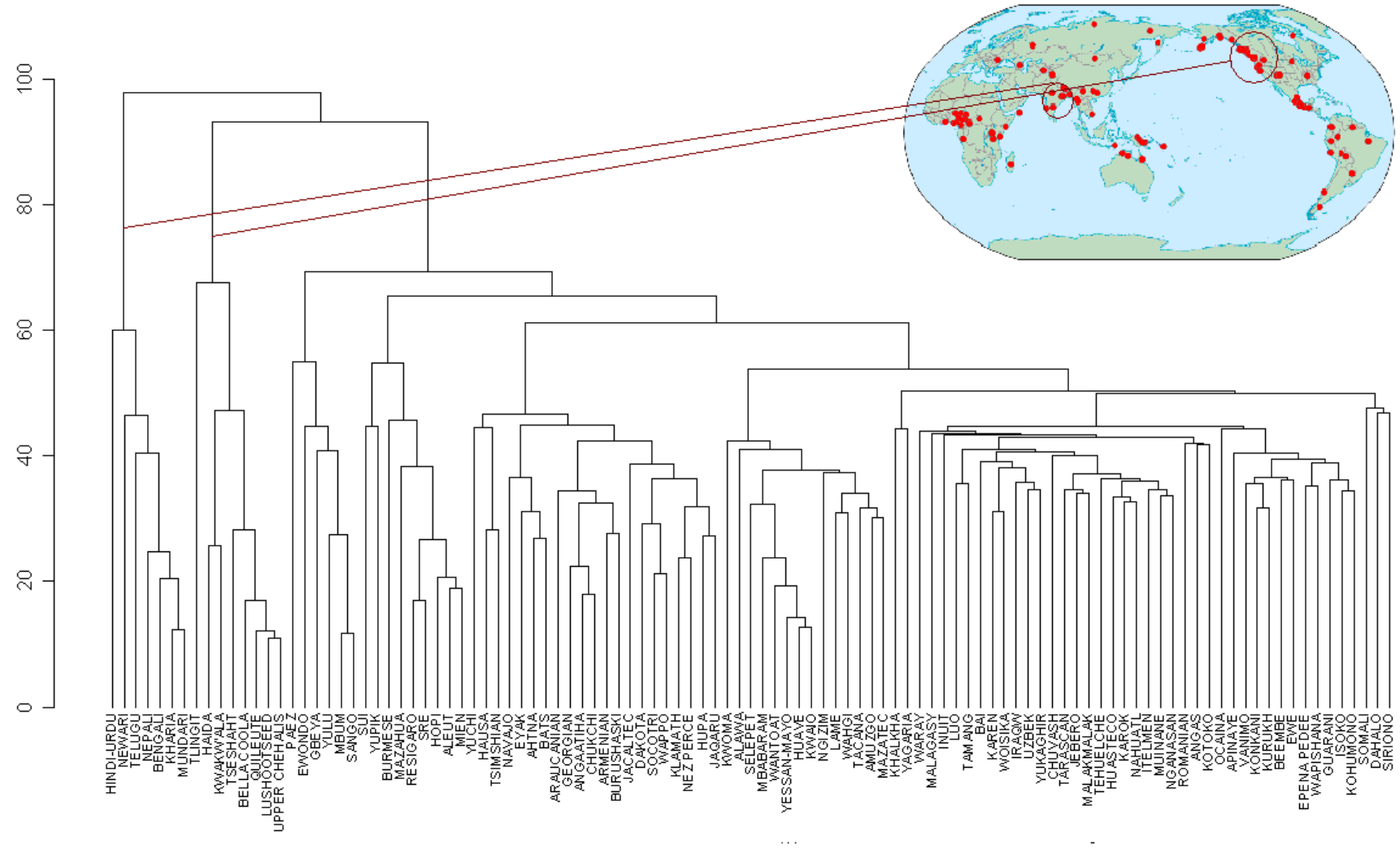


Figure 34. Cluster 3.

Within Cluster 4, the NENETS-RUSSIAN subcluster, whose internal structure is presented in Figure 28, is associated with the presence of palatalized consonants (.40<r<.92), especially /r^{j}/ (r=.91) as well as velarized palatoalveolar fricatives (r=.41). The ACOMA-NAXI subcluster is associated with the presence of the voiceless aspirated palatal sibilant affricate (r=.57), which occurs only in GELAO, LAI and NAXI, all within this subcluster, the retroflexed mid central unrounded vowel (r=.57), which occurs only in MANDARIN, NAXI and GELAO, the labialized alveolar voiceless stop, which occurs only in IRANXE and NAMBAKAENGO, the nasalized lower mid back rounded to high front unrounded diphthong, which occurs only in IRANXE and KIOWA, and the high central unrounded to high front unrounded diphthong, which occurs only in IRANXE and ACOMA. The BRUU-YUCATEC subcluster is associated with the presence of laryngealized vowels which occur only within the cluster in DINKA, SEDANG, SOUTHERN NAMBIQUARA and YUCATEC, and breathy vowels, which occur in BRUU, DINKA and NYAH KUR within the cluster and in PARAUK outside of it.

This section presented the results of a hierarchical clustering analysis of phoneme inventories of the languages of the world. In some areas of the world, including northwest North America, India, and Eastern Europe / Russia, clear areal patterns have been identified where all languages in an area, regardless of genetic affiliation, share certain rare phonemes or classes of phonemes. In other cases, the observed similarities may be due to genetic relatedness, as in the case of Turkic and Finno-Ugric languages and languages of Australia. Yet in other cases, the similarities are accidental, as is probably the case for NENETS and PO-AI. The data themselves do not allow us to determine the source of the observed similarities. An important topic for future research is to devise ways of reducing the number of accidental groupings. One possible source of such groupings is differences in transcription methods between scholars from different areas (and similarities between scholars with similar backgrounds). The influence of this factor may be observed by correlating principal components of the space in which languages vary with characteristics of the language descriptions, including background information about the transcribers.

Perhaps, a more interesting possible source of accidental groupings is the distance metric. In the present paper, as well as in all other extant clustering analyses of typological data (Albu 2007, Altmann 1971, Altmann and Lehfeldt 1973, 1980, Cysouw 2007, Szmrecsanyi and Kortmann to appear) clustering was accomplished on the basis of a Euclidean distance metric. To illustrate the influence of the distance metric on clustering decisions, let us consider two points (0,1) and (1,2). These points differ on two dimensions and differ by one unit on each of the dimensions. The distance between them under the Euclidean distance metric is the same as the distance between the two points on a sheet of paper: $\sqrt{(1+1)}$. The distance between them under the city-block, or Manhattan, distance metric is simply the sum of the amounts by which they differ on the two dimensions: 1+1. However, the distance between two points that differ on only one dimension is the same under both metrics. Thus, the use of the city-block metric in place of the Euclidean metric for clustering languages would move languages that differ on multiple dimensions further apart than they are when the Euclidean distance metric is used. Other distance metrics are also possible. Finally, the present clustering solution is based on weighing principal components based on how much variance in phoneme inventories they account for. Other weighting schemes

are likely to produce different results. For instance, it is likely that certain segments change more slowly than other segments (e.g., consonants vs. vowels). If this is the case, then the presence/absence of a slow-changing segment may be weighted more heavily than the presence/absence of an unstable segment for estimating genetic relatedness.

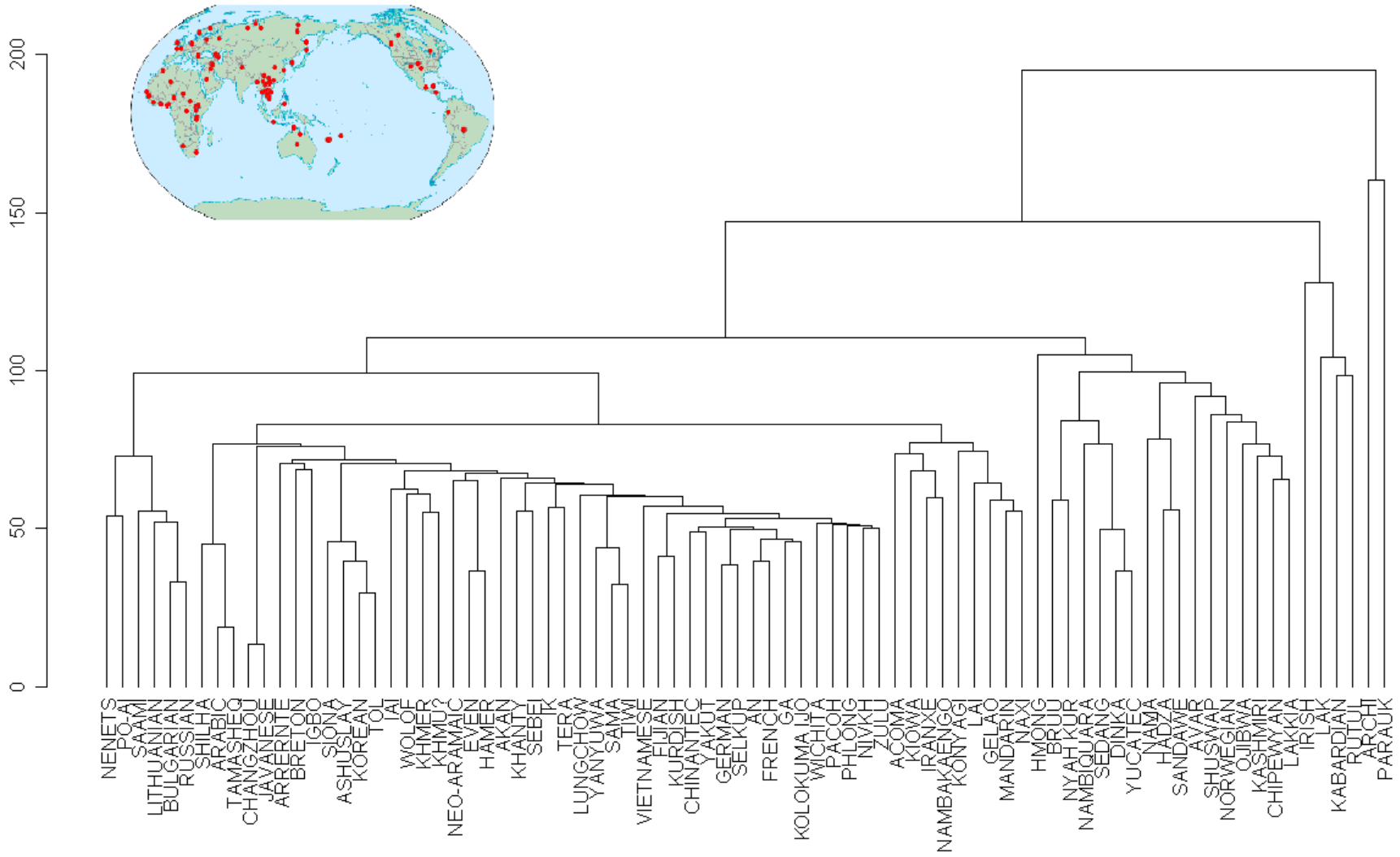


Figure 35. Cluster 4.

The clustering algorithm used in the present paper creates clusters of languages that are maximally similar in the principal components space. Then it looks at languages that are more isolated and assigns each of them to a cluster whose center is closest to the language's location in the principal component space. The use of cluster center locations, as opposed to, for instance, the cluster membership of k nearest neighbors is arguable. For the purpose of the present study, this clustering method seemed to produce the best results relative to other methods available in R. However, this does not exhaust the space of possible clustering methods.

The use of principal components, on the other hand, is well supported by the results of the present study. First, it allows for generalization across segments with a similar distribution across languages. For instance, languages in the GADSUP-TOTONAC cluster do not share any specific segments but almost all have a nasal vowel. The emerging categories often have a family resemblance or prototype structure. As shown above, multiple language clusters are associated with a prototypical segment that has all features typical of segments associated with the cluster of languages and is associated with the cluster of languages more than segments that have only some of the typical features. In the present study, this prototype structure is associated with the 'star' clusters, which contain languages that are clearly areally and/or genetically related and have many segments associated with them. The generality of this phenomenon remains to be tested.

9. Summary and conclusion

The present paper applied multivariate statistical techniques to the analysis of phoneme inventories in the UPSID database (Maddieson 1984, 1991, Maddieson and Precoda 1990). The results of this analysis indicated that phoneme inventories form three distinct clusters based mostly on the glottal articulations present in the stop inventory (aspiration, ejection, both or neither). The three clusters show significant geographical specificity with languages of the Americas, languages of North Africa and Europe and languages of Southeast Asia associated with different clusters. More detailed clustering analyses were shown to reveal more specific areas associated with uvular consonants, front rounded vowels, vowel nasalization, palatalized consonants, prenasalized obstruents, and breathy voiced stops. Larger clusters are also associated with the absence of common consonants, such as plain voiceless stops and /s/. Interestingly, these are the same phonological features that have been shown to produce classifications that are most consistent with classifications based on the full set of features present in the World Atlas of Linguistic Structures by Albu (2007). Albu used three different statistical methods to calculate the consistency between cluster assignments based on individual features vs. assignments based on the full set of features available in the database. Of the phonological features within the top dozen of most consistent features, all (except tone, which is not marked in UPSID) are associated with clusters in the present study. The only features that are associated with a cluster in the present study but not shown to be high in consistency by Albu (2007) are breathy voiced stops, associated with languages of India, and prenasalized obstruents, associated with some languages of Central Africa and, to a lesser extent, the South American language Paez. Thus, it is likely that the structure observed in the present language clustering analysis would be similar if non-phonological features present in WALS were included.

An interesting finding of the present study is that stop inventories and (monophthong) vowel inventories form distinct clusters in the space defined by principal components, whereas sets of continuant obstruents and sonorant consonants do not. This suggests that sets of stops and vowels function as interacting systems, for which certain configurations of space are more optimal than others, as suggested by Liljencrantz and Lindblom (1972) for vowels, whereas sets of continuant obstruents and sonorant consonants do not.

In both cluster analyses of vowel and stop systems as well as in the interpretation of language clustering, featurally similar segments were found to behave in a similar manner. Several language clusters were associated with feature sets, as opposed to individual segments. Thus, in at least one cluster, member languages had to have at least one nasal vowel but no nasal vowel was shared by all languages in the cluster. Other, more cohesive, clusters, were associated with a set of features in which the presence of all the associated features in a segment was most favorable for the segment being associated with the cluster with presence of some of the features being less favorable. Thus, northwest North American languages could be seen as being associated with uvular labialized ejective consonants and to a lesser extent with consonants that are uvular but neither labialized nor ejective or ejective and uvular but not labialized. Clusters of frequent phonemes were most often interpretable in terms of features as well as typological frequency / markedness. However, despite the fact that the segment inventories of languages do appear to have featurally-describable structure, languages do not simply have feature inventories because no two segments of reasonable frequency occur in the exact same set of languages.

On the methodological side, the present study suggests that multivariate statistical techniques can be useful for extraction of structure from typological data and classification of both typological features and their inventories. Since the original study of Altmann (1971), these techniques have been underused in typological research (see Cysouw 2007 for review). While they are unlikely to replace careful qualitative analyses, they provide an objective measure of similarity between sets of typological features and an objective way of classifying features or sets of features on the basis of similarity. Future typological research may fruitfully combine quantitative methods with the results of qualitative analyses. For instance, Albu (2007) correlates results of cluster analysis with information about genetic relationships between the clustered languages, It may also be possible to include information about genetic relationships produced by qualitative analysis as a variable in cluster analysis, which could potentially improve upon the results of both a purely statistical and a purely qualitative approach.

References:

Albu, M. 2007. *Quantitative analyses of typological data*. PhD Thesis: University of Leipzig.

Altmann, G. 1971. Die phonologische Pro"lähnlichkeit. Ein Beitrag zur Typologie phonologische Systeme der slawischen Sprachen. *Phonetica*, 24: 9-22.

Altmann, G., and W. Lehfeldt. 1973. Allemeine Sprachtypologie: Prinzipien und Meßverfahren. Munich: Fink.

Altmann, G., and W. Lehfeldt. 1980. Einführung in die Quantitative Phonologie. Bochum: Brockmeyer.

Biber, D. 1988. *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Bibiko, H.-J. 2005. The World Atlas of Language Structures: The Interactive Reference Tool. Oxford: Oxford University Press.

Croft, W., and K. T. Poole. To appear. Inferring universals from grammatical variation: multidimensional scaling for typological analysis. *Theoretical Linguistics*

Cysouw, M. 2007. New approaches to cluster analysis of typological indices. In Reinhard Köhler & Peter Grzbek (eds.) *Festschrift für Gabriel Altmann.* Berlin: Mouton.

Haspelmath, M., M. S. Dryer, D. Gil, and B. Comrie. 2005. *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Liljencrantz, J, and B. Lindblom. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 48, 839-62

Maddieson, I. 1984. *Patterns of Sounds*. Cambridge: Cambridge University Press.

Maddieson, I. 1991. Testing the universality of phonological generalizations with a phonetically specified segment database: results and limitations. *Phonetica* 48: 193-206.

Maddieson, I., and K. Precoda 1990. Updating UPSID. *UCLA Working Papers in Phonetics* 74: 104-114.

Pericliev, V. 2004. Universals, their violation, and the notion of phonologically peculiar languages. *Journal of Universal Language*, 5, 1-28.

Silnitsky, G. 2003. Correlation of phonetic and morphological systems of Indo-European languages. *Journal of Quantitative Linguistics*, 10, 129-41.

Szmrecsanyi, B., and B. Kortmann. To appear. The morphosyntax of varieties of English worldwide: A quantitative perspective. *Lingua*.

Wittgenstein, L. 1953. *Philosophical Investigations*. Malden: Blackwell.

Appendix 1: Coordinates and cluster assignments of UPSID languages in the PC2-PC3 space.

Cluster 1

Language	PC2	PC3
AGHEM	0.063	-0.001
AIZI	0.079	0.022
AJU	0.048	0.031
AKAN	0.046	-0.046
AKAWAIO	0.025	-0.004
ALBANIAN	0.046	-0.045
ALLADIAN	0.064	0.000
AMELE	0.052	-0.003
AMO	0.072	-0.003
AN	0.041	-0.005
ANDAMANESE	0.044	0.041
ANGAS	0.071	-0.002
ARABIC	0.056	-0.042
AUCA	0.016	0.019
AWIYA	0.007	-0.018
AZANDE	0.060	0.003
AZERBAIJANI	0.070	-0.033
BAMBARA	0.035	-0.020
BANDJALANG	0.074	0.018
BARIBA	0.038	-0.007
BASHKIR	0.046	-0.021
BASQUE	0.042	-0.007
BATAK	0.035	0.019
BEJA	0.022	-0.043
BENGALI	0.012	-0.031
BETE	0.080	0.039
BIROM	0.076	0.025
BISA	0.074	0.034
BOBO-FING	0.058	-0.035
BODO	0.030	0.007
BORORO	0.020	-0.008
BRAHUI	0.046	-0.033
BRETON	0.052	-0.005
BULGARIAN	0.058	-0.018
BURMESE	0.009	-0.030
CAMSA	0.015	-0.011
CARIB	0.021	0.028
CAYAPA	0.012	0.001
CHAMORRO	0.029	-0.002
CHUAVE	0.040	-0.006
CHUVASH	0.041	-0.030
CUNA	0.004	0.002
DAFLA	0.057	0.047
DAGBANI	0.068	-0.005
DAGUR	0.041	0.010
DANGALEAT	0.059	0.048
DERA	0.028	0.035
DINKA	0.062	0.056
DIOLA	0.055	0.040
DOAYO	0.053	0.003
DOGON	0.017	-0.006
DYIRBAL	0.084	0.047
EFIK	0.049	0.043
EJAGHAM	0.059	0.058
EVEN	0.066	0.024
EWE	0.053	-0.045
EWONDO	0.073	0.019
FE?FE?	0.051	-0.044
FINNISH	0.056	0.008
FRENCH	0.070	0.017
FUR	0.059	0.012
GA	0.066	-0.051
GBEYA	0.021	0.017
GERMAN	0.070	-0.007
GREEK	0.062	0.006
GWARI	0.045	-0.013
HAMER	0.037	0.002
HINDI-URDU	0.046	-0.043
HIXKARYANA	-0.004	-0.003
HUNGARIAN	0.064	-0.032
IAI	0.043	0.002
IBAN	0.023	0.000
IGBO	0.024	-0.024
IK	0.045	-0.013
IRAQW	0.023	-0.013
ISLAND CARIB	0.004	-0.020
ISOKO	0.063	0.015
IVATAN	0.040	0.016
JAPANESE	0.036	0.014
JOMANG	0.102	0.027
KALA LAGAW YA	-0.047	0.056
KANAKURU	0.022	-0.034
KANURI	-0.032	0.026
KAUGLI	-0.018	-0.017
KERA	0.060	0.012
KET	0.010	0.019
KHALKHA	0.042	-0.054
KHARIA	0.015	-0.007
KHASI	0.006	0.005
KIRGHIZ	0.061	-0.010
KLAO	0.015	-0.024
KOHUMONO	0.067	-0.030
KOIARI	0.038	-0.020
KOLOKUMA IJO	0.042	0.012

KOMA	0.010	-0.012	PAIWAN	0.033	0.030
KOMI	0.056	-0.009	PAPAGO	0.010	-0.012
KONKANI	0.015	-0.003	PARAUK	0.028	-0.008
KONYAGI	0.035	0.057	PASHTO	0.040	-0.034
KOTA	0.062	-0.037	ROMANIAN	0.042	-0.045
KOTOKO	0.041	-0.018	RUKAI	0.019	0.025
KOYA	0.032	-0.007	RUSSIAN	0.051	0.002
KPAN	0.046	-0.024	SAAMI	0.025	0.022
KPELLE	0.062	0.024	SA'BAN	0.037	-0.017
KUNAMA	0.059	-0.013	SAMA	0.017	0.006
KUNIMAIPA	0.049	0.024	SANGO	0.038	0.001
KURDISH	0.053	-0.018	SENADI	0.051	0.015
KURUKH	-0.002	-0.014	SENTANI	0.007	0.037
LAME	0.013	-0.008	SHILHA	0.057	-0.027
LELEMI	0.047	0.005	SINHALESE	0.052	0.014
LITHUANIAN	0.051	-0.025	SONGHAI	0.039	-0.009
LUGBARA	0.045	-0.017	SOUTHERN KIWAI	-0.090	-0.049
LUO	0.026	-0.012	SRE	0.012	0.008
MABA	0.071	0.014	SUENA	0.036	0.033
MALAGASY	0.047	-0.023	TABI	0.053	0.016
MAMBILA	0.036	0.004	TAGALOG	0.021	-0.007
MANCHU	0.047	-0.010	TAMA	0.069	0.030
MARANAO	0.009	0.013	TAMASHEQ	0.063	-0.031
MARGI	0.048	-0.028	TAMPULMA	0.084	0.033
MBABARAM	0.050	0.019	TAROK	0.058	-0.027
MBA-NE	0.046	0.041	TEKE	0.041	0.006
MBUM	0.056	0.004	TELUGU	0.028	-0.012
MIEN	0.007	-0.017	TEMEIN	0.065	0.051
MIXE	0.003	-0.009	TEMNE	0.010	0.023
MOGHOL	0.042	-0.010	TERA	0.036	0.021
MONGUOR	0.044	-0.032	TETUN	-0.002	-0.003
MORO	0.052	0.021	TICUNA	0.004	-0.020
MUINANE	0.021	-0.019	TIDDIM CHIN	0.020	-0.013
MUMUYE	0.052	-0.017	TIGAK	0.040	0.061
MUNDARI	0.012	-0.022	TIRURAY	0.035	0.025
MURINHPATHA	0.056	0.073	TULU	0.050	0.016
MURSI	0.042	0.030	TURKISH	0.061	-0.029
NANAI	0.046	0.020	TUVA	0.055	-0.014
NAXI	0.030	-0.019	USAN	0.019	0.027
NDUT	0.035	0.040	UZBEK	0.044	-0.033
NERA	0.055	-0.016	VANIMO	-0.004	0.007
NGANASAN	0.035	0.032	WEST MAKIAN	0.031	-0.003
NGIZIM	0.041	0.003	WOISIKA	0.067	0.030
NIMBORAN	0.026	0.002	WOLOF	0.043	0.037
NONI	0.055	0.000	YAGARIA	0.017	-0.008
NORWEGIAN	0.046	-0.032	YAKUT	0.080	0.008
NUBIAN	0.056	-0.042	YAQUI	0.003	-0.022
NYIMANG	0.064	0.038	YAREBA	0.026	-0.008
OGBIA	0.070	0.025	YAWA	0.001	0.012
ORMURI	0.060	-0.060	YAY	0.004	0.019

YIDINY	0.084	0.047
YORUBA	0.048	-0.025
YUKAGHIR	0.050	0.031
YULU	0.060	0.027
ZOQUE	0.007	0.013
Cluster 2		
!XU	0.018	-0.050
ACOMA	-0.004	-0.102
ADZERA	0.035	-0.067
AHTNA	-0.002	-0.099
ALAMBLAK	0.016	-0.079
AMHARIC	0.011	-0.065
ANDOKE	-0.011	-0.048
ARCHI	-0.012	-0.091
ARMENIAN	-0.053	-0.084
AVAR	0.004	-0.108
BAKAIRI	0.010	-0.034
BARASANO	-0.011	-0.039
BATS	-0.002	-0.135
BEEMBE	-0.037	-0.017
BELLA COOLA	-0.077	-0.029
BERTA	0.019	-0.044
BRIBRI	0.003	-0.038
BURUSHASKI	0.008	-0.059
CADDO	-0.033	-0.044
CAYUVAVA	-0.018	-0.016
CHANGZHOU	-0.049	-0.013
CHATINO	-0.016	-0.038
CHEROKEE	0.000	-0.043
CHIPEWYAN	-0.075	-0.064
COFAN	0.009	-0.051
CUBEO	-0.031	-0.028
DADIBI	-0.090	-0.021
DAHALO	-0.015	-0.067
DAKOTA	-0.058	-0.069
EPENA PEDEE	-0.028	-0.073
EYAK	-0.022	-0.107
FARSI	0.049	-0.107
GEORGIAN	-0.059	-0.097
GUAHIBO	-0.010	-0.034
HADZA	-0.029	-0.065
HAIDA	-0.091	-0.040
HAUSA	0.011	-0.054
HMONG	-0.044	-0.018
HUARI	0.014	-0.064
HUASTECO	-0.099	-0.044
HUPA	-0.071	-0.045
IRISH	0.018	-0.052
ITELMEN	-0.054	-0.018
ITONAMA	-0.048	-0.034
IZI	0.022	-0.082
JACALTEC	-0.068	-0.059
JAPRERIA	-0.070	-0.034
JAQARU	-0.078	-0.028
JINGPHO	-0.013	-0.063
KABARDIAN	-0.069	0.006
KAROK	-0.002	-0.032
KASHMIRI	0.065	0.039
KAWAIISU	-0.007	-0.068
KEFA	-0.095	-0.037
K'EKCHI	0.018	-0.087
KIOWA	-0.021	-0.084
KLAMATH	-0.017	-0.089
KULLO	0.010	-0.086
KWAKW'ALA	-0.025	-0.069
LAHU	0.005	-0.046
LAK	-0.009	-0.096
LUA	-0.023	-0.030
LUE	-0.012	-0.023
LUNGCHOW	-0.058	-0.017
LUSHOOTSEED	-0.026	-0.107
MAIDU	-0.062	-0.057
MAZAHUA	-0.063	-0.073
MIXTEC	-0.071	-0.023
NAVAJO	-0.063	-0.058
NEO-ARAMAIC	0.018	-0.052
NEPALI	-0.014	-0.039
NEWARI	0.009	-0.036
OCAINA	-0.030	-0.037
PICURIS	-0.044	-0.081
PIRAHA	-0.018	-0.032
POMO	-0.027	-0.036
QUECHUA	-0.056	-0.020
QUILEUTE	-0.048	-0.092
RESIGARO	-0.003	-0.055
RUTUL	-0.007	-0.114
SALIBA	-0.020	-0.043
SANDAWE	-0.012	-0.057
SENECA	-0.037	-0.012
SHASTA	-0.096	-0.031
SHIRIANA	-0.083	-0.022
SIRIONO	-0.062	-0.026
SOCOTRI	0.015	-0.065
SOMALI	0.028	-0.058
SOUTHERN NAMBIQUARA	0.004	-0.007
SUI	-0.039	-0.044
TARASCAN	-0.005	-0.033
TEHUELCHE	-0.048	-0.060
THAI	-0.024	-0.017
TIGRE	0.041	-0.084

TLAPANEC	-0.015	-0.055
TLINGIT	-0.037	-0.086
TSESHAHT	-0.094	-0.061
TSIMSHIAN	-0.033	-0.058
TUNICA	0.018	-0.064
TZELTAL	-0.021	-0.077
UPPER CHEHALIS	-0.102	-0.070
WAPISHANA	-0.031	-0.081
WAPPO	-0.030	-0.075
WICHITA	-0.058	-0.023
WINTU	-0.047	-0.089
WIYOT	-0.093	-0.028
XIAMEN	0.000	-0.051
YANA	-0.013	-0.108
YUCATEC	-0.091	-0.043
YUCHI	-0.048	-0.114
ZULU	-0.023	-0.028
ZUNI	-0.074	-0.027
Cluster 3		
ABIPON	-0.057	0.030
ACHE	-0.072	0.002
ACHUMAWI	-0.064	-0.012
AINU	-0.054	0.051
ALABAMA	-0.043	0.000
ALAWA	-0.028	0.111
ALEUT	-0.048	0.018
AMAHUACA	-0.080	0.001
AMUESHA	-0.061	0.010
AMUZGO	-0.085	0.007
ANGAATIHA	-0.056	0.055
ANI	-0.063	0.023
AO	-0.046	0.049
APINAYE	-0.053	0.005
ARABELA	-0.052	0.017
ARAUCANIAN	-0.027	0.095
ARRERNTE	-0.037	0.115
ASHUSLAY	-0.056	-0.001
ASMAT	-0.038	0.043
ATAYAL	-0.046	0.058
BAI	-0.021	0.013
BAINING	-0.046	0.085
BARDI	-0.020	0.122
BRAO	-0.021	0.025
BRUU	-0.010	0.013
BURARRA	-0.017	0.138
CACUA	-0.073	0.029
CAMPA	-0.076	0.001
CHAM	-0.052	0.030
CHUKCHI	-0.056	0.052
DIEGUENO	-0.044	0.043
DIYARI	-0.019	0.127
EKARI	0.015	0.063
FASU	-0.080	0.005
FIJIAN	-0.036	0.050
FUZHOU	-0.046	0.037
GADSUP	-0.024	0.035
GARAWA	-0.007	0.133
GELAO	-0.030	-0.002
GUAJIRO	-0.060	0.002
GUAMBIANO	-0.055	0.022
GUARANI	-0.052	0.005
GUGU-YALANYI	-0.029	0.115
HAKKA	-0.053	0.005
HAWAIIAN	-0.067	0.037
HIGHLAND CHINANTEC	-0.058	0.045
HOPI	-0.046	0.040
HUAVE	-0.027	0.017
IATE	-0.076	-0.001
INUIT	-0.044	0.046
IRANXE	-0.019	0.004
IRARUTU	-0.044	0.056
IWAM	-0.053	0.043
JAVANESE	-0.037	0.059
JEBERO	-0.044	0.067
JIVARO	-0.088	0.005
KAINGANG	0.052	0.029
KALIAI	-0.031	0.129
KALKATUNGU	-0.045	0.056
KAM	0.028	0.035
KAREN	-0.046	-0.017
KEWA	-0.017	0.028
KHANTY	-0.022	0.073
KHMER	-0.009	0.026
KHMU?	-0.038	0.019
KOREAN	-0.061	0.005
KORYAK	-0.037	0.046
KWAIO	-0.056	0.036
KWOMA	-0.067	0.007
LAI	-0.048	-0.002
LAKKIA	-0.056	-0.001
LENAKEL	-0.038	0.045
LUISENO	-0.050	0.014
MAASAI	-0.017	0.064
MALAKMALAK	-0.021	0.114
MANDARIN	-0.041	0.020
MARI	-0.006	0.029
MAUNG	-0.015	0.128
MAXAKALI	-0.096	-0.015
MAZATEC	-0.052	0.005

MOR	-0.075	0.014
MOVIMA	-0.083	0.020
MOXO	-0.013	0.010
NAHUATL	-0.070	0.009
NAMA	-0.055	-0.001
NAMBAKAENGO	-0.055	0.018
NASIOI	-0.042	0.045
NENETS	-0.025	0.029
NEZ PERCE	-0.074	-0.005
NGARINJIN	0.001	0.132
NGIYAMBAA	-0.047	0.116
NICOBARESE	-0.017	0.053
NIVKH	-0.030	0.034
NUNGGUBUYU	-0.023	0.125
NYAH KUR	-0.012	0.021
NYANGI	-0.009	0.087
OJIBWA	-0.038	0.015
PACOH	-0.036	0.048
PAEZ	-0.070	0.014
PANARE	-0.091	0.005
PAYA	-0.054	-0.006
PHLONG	-0.047	-0.005
PO-AI	-0.041	-0.002
POHNPEIAN	-0.013	0.091
QAWASQAR	-0.058	0.000
RORO	-0.058	0.011
ROTOKAS	-0.028	0.024
SAVOSAVO	-0.034	0.070
SEANG	-0.052	0.012
SEBEI	-0.019	0.085
SELEPET	-0.048	0.014
SELKUP	-0.021	0.043
SHUSWAP	-0.063	0.009
SIERRA MIWOK	-0.056	0.051
SIONA	-0.090	-0.016
SPANISH	-0.031	0.041
TACANA	-0.031	-0.003
TAISHAN	-0.053	0.008
TAMANG	-0.055	0.011
TAORIPI	-0.034	0.022
TIWI	-0.035	0.117
TOL	-0.082	-0.002
TONKAWA	-0.057	0.029
TOTONAC	-0.070	-0.006
TRUMAI	-0.045	0.020
TSOU	-0.040	0.027
VIETNAMESE	0.000	0.036
WAHGI	-0.047	0.073
WANTOAT	-0.053	0.057
WARAO	-0.038	-0.007
WARAY	0.003	0.113
WARIS	-0.051	0.043
WESTERN DESERT	-0.043	0.126
WIK-MUNKAN	-0.029	0.110
YAGUA	-0.066	0.015
YANYUWA	-0.018	0.111
YESSAN-MAYO	-0.043	-0.003
YOLNGU	-0.027	0.114
YUCUNA	-0.084	0.008
YUPIK	-0.024	0.021

Appendix 2: Significant correlates of belonging to Clusters 1, 2, or 3 in the PC2-PC3 space in UPSID notation.

Cluster 1

Phoneme	r
Present:	
d	0.66
b	0.65
g	0.61
f	0.41
gb	0.35
kp	0.33
dj	0.30
v	0.30
z	0.28
dZ	0.27
nj	0.26
O	0.24
N	0.21
d.	0.21
E	0.21
Absent:	
t'	-0.33
?	-0.32
k'	-0.32
tS'	-0.29
ts'	-0.28
p'	-0.28
th	-0.28
tsh	-0.27
kh	-0.26
ph	-0.24
kW	-0.22
kW'	-0.21

Cluster 2

Phoneme	r
Present:	
k'	0.63
t'	0.61
ts'	0.56
tS'	0.55
p'	0.54
kW'	0.40
th	0.40
tsh	0.39
q'	0.39
kh	0.38
tSh	0.36
ph	0.34
tlF'	0.32
?	0.31
h	0.31
qW'	0.28
S	0.28
X	0.26
kWh	0.26
qh	0.26
XW	0.25
hlF	0.24
tlF	0.24
ts	0.24
dlF	0.23
tS	0.22
c'	0.21
s'	0.21
Absent:	
N	-0.33
k	-0.27
nj	-0.25
O	-0.24
t	-0.23
E	-0.22
s	-0.28
k'	-0.24
dz	-0.23
ts'	-0.23
dj	-0.22
t'	-0.22
Z	-0.21
gb	-0.21
S	-0.20
p'	-0.20
tS'	-0.20

Cluster 3

Phoneme	r
Present:	
t_	0.27
Absent:	
b	-0.75
d	-0.74
g	-0.68
dZ	-0.37
f	-0.32
z	-0.31