Advances in Complex Systems
© World Scientific Publishing Company

PHONOLOGICAL DIVERSITY, WORD LENGTH, AND POPULATION SIZES ACROSS LANGUAGES: THE ASJP EVIDENCE

SØREN WICHMANN
Department of Linguistics, Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6, D-04103 Leipzig, Germany
wichmann@eva.mpg.de

TARAKA RAMA
Department of Swedish Language, University of Gothenburg,
Gothenburg, 405 30, Sweden
taraka.rama.kasicheyanula@gu.se

ERIC W. HOLMAN
Department of Psychology, University of California,
Los Angeles, California 90095-1563, USA
holman@psych.ucla.edu

Received (received date)
Revised (revised date)

Previous literature has reported a positive correlation between phoneme inventory sizes and population sizes for languages, indicating that larger languages tend to make more phonological distinctions, and claims have also been made that average word length and phoneme inventory sizes are negatively correlated. Yet another relevant variable is geography, since the spatial propinquity of languages influences the similarity of their overall typological profile; moreover, specific historical events affecting language distributions, such as migrations or the development of certain cultural advantages, are usually also anchored geographically. In this paper we replicate previous findings on a substantially larger set of data drawn from comparative wordlists in the database of the Automated Similarity Judgment Program (ASJP) and discuss the relationships among the three variables mentioned in the title of the paper as well the influence of geography, including the idea that phonemic diversity across the world's languages provides evidence for an out-of-Africa model of the expansion of languages.

Keywords: demography, historical linguistics, linguistic diversity, ASJP, n-gram.

1. Introduction

Hay & Bauer [10] report a positive correlation between population sizes and phoneme size inventories. Atkinson [1] replicates this result, and additionally reports a negative correlation between phoneme size inventories and the distance from Africa of a given language. Neither Hay & Bauer nor Atkinson cite Nettle

[19–21], who has suggested, based on a small sample of languages, that phoneme inventory sizes and mean word length are inversely related.

In this paper we investigate the replicability of the results of [10] and [19–21] on a larger dataset comprising not just a few dozen languages as in the Nettle studies or a few hundred as in Hay & Bauer's study, but more than 3000 languages carrying unique ISO 639-3 codes, a sample which represents close to one half of the world's spoken languages as defined in [13]. This dataset, known as the ASJP Database and available as [27], consists of lists of 40 standard concepts and the corresponding words for these concepts in different languages. Different sections draw upon specific subsets of the database, as will be specified. The total sample includes languages from 109 out of the world's 121 linguistic families, 46 out of 123 isolates and unclassified languages, and 40 out of 122 creoles, mixed languages, and pidgins. Each family or non-genealogical group (such as 'Unclassified' and 'Creole') is represented by around one half of its members in the database.

The ASJP Database thus has the advantage of a large coverage of languages, but it is possibly disadvantaged for the purposes of the present study by containing words and not actual phoneme inventories and furthermore by a transcription procedure by which certain phonological distinctions are merged. In Sec. 2 we therefore describe the nature of the data further and investigate the degree to which the segments represented in the word lists (henceforth SR for 'Segments Represented') are numerically proportional to phoneme size inventories. In Sec. 3 we correlate SR with mean word length and in Sec. 4 with population sizes. Patterns of residence and migration will influence the distribution of languages and therefore also the linguistic typological profiles of different world areas. It has been claimed in [1] that the distribution of phoneme size inventories reflects migrations pertaining to the very first movements of humans out of Africa. While it is inherently doubtful that phoneme inventory sizes can change quickly enough to 'stay in tune' (i.e., correlate) with population sizes *and* at the same time preserve a signal many thousans of years old, there is no doubt that migrations underlie some typological distributions. In Sec. 5 we therefore discuss the claim of [1]. Sec. 6 summarizes the findings.

2. SR as predictor of phonological inventory sizes

As a first step toward using the ASJP data [27] for the study of the worldwide distribution of phoneme inventory sizes we correlate SR (segments represented in the word lists) with the segment inventory sizes of UPSID [18], which contains information on phonological segments for 451 languages. The UPSID sample was designed to serve as a source for [17] and was chosen so as to be representative of each of the world's language groups, but presumably also to maximize the coverage of the variation in sound structures across the world's languages. It includes as many as 919 different phonological segments.

We would like to know how representative SRs are of given language's inventory of phonological segments. There are two obvious sources for potential discrepancies between SRs and UPSID inventories. The first is the inherent limitation of word lists: we cannot expect a short list of words to contain all of the phonological segments in a language, especially if the language has a very large segment inventory. The second source for potential discrepancies is the way in which segments are transcribed in the ASJP lists, henceforth ASJPcode. Thus, in the following paragraphs we provide some more detail on the nature of the word lists and on ASJPcode.

The word lists comprise a 40-item subset of the so-called Swadesh list, where the concepts were selected for their higher stability [11] , i.e. for the tendency for words for these concepts to only be slowly replaced by new words. As a rule of thumb the database normally only includes lists that are at least 70% complete, i.e., which contain at least 28 items on the 40-item list. Less complete lists are only included in a handful of exceptional cases where the importance of the list was judged to override the usual criterion. The concepts on the 40-item subset of the Swadesh are: BLOOD, BONE, BREAST, COME, DIE, DOG, DRINK, EAR, EYE, FIRE, FISH, FULL, HAND, HEAR, HORN, I, KNEE, LEAF, LIVER, LOUSE, MOUNTAIN, NAME, NEW, NIGHT, NOSE, ONE, PATH, PERSON, SEE, SKIN, STAR, STONE, SUN, TONGUE, TOOTH, TREE, TWO, WATER, WE, YOU (SG). It is often the case that more than one word is available for a given concept, i.e., synonyms, near-synonyms or phonological variants. In the present study we arbitrarily use only the first item in a list of alternative forms to avoid the introduction of biases from the nature of the sources (large dictionaries vs. shorter vocabularies or the individual practices of different transcribers) and to enhance tractability of the results: using only one synonym per word allows us to equate the number of attested concepts with the number of words in a list.

One might wonder whether some other selection of concepts would be more adequate for sampling phonological segments. A larger list would obviously increase the probability that all segments of language are represented in the list, but the strength of this relationship is an empirical question.

We do find a small positive correlation ($r = .17$) between the number of words attested (which ranges from 23 to 40 with an average of 35.7) and SR for our total sample of 3169 languages (see Table 1 in Online Supplementary Materials). So the number of words matters, but since the frequency distribution of phonological segments presumably has a Zipfian nature in texts we would expect sort of relation between word list size and SR which would produce diminishing returns with more words.

As for the selection of concepts one may pause to consider which sorts of words ought to be included in a list for the segments to be maximally representative of the total inventory. Since the relation between sound and meaning in language is mostly arbitrary a word for any concept can potentially contain any phoneme. But there are two special classes of words that can exhibit phonemes otherwise not attested in the standard vocabulary, namely onomatopoeic words and loanwords. The ASJP lists do not include concepts that are likely to be subject to onomatopoeia in the narrow sense of sound imitation, even if they do include concepts which in some languages are prone to sound symbolism in a broader sense, cf. [25]. So rare phonemes confined to onomatopoetic expression are not expected to be represented. But such phonemes are in any case often treated as marginal or as not belonging to regular segment inventories by descriptive phonologists. The lists do, however, often include loanwords, maybe on the order of 5% or so on average (currently we only have estimates available for longer versions of the Swadesh list, indicating an average of 8.5% for lists of 99 Swadesh items across a sample of 36 languages, cf. [11]).

Thus, as regards the size and nature of the selection of concepts it does not seem that there is any particular reason to expect that ASJP lists could not be representative of at least a regular proportion of segment inventories. We now turn to the issue of transcription.

The transcription system, ASJPcode, was first presented in [5]. The system operates with 34 basic consonantal symbols and 7 vowel symbols, cf. Tables 1–2 below. The representation of vowels is limited to at most 7 different qualities, as reflected in the 7 symbols, but in addition nasalization can be indicated by an asterisk following a vowel symbol. The 34 consonantal symbols can be combined freely to represent phonetically complex segments that are subsequently treated as single phonological units. The symbols ∼ and $ follow sequences of respectively two and three consonant symbols to indicate that such sequences are to be treated as units. For instance, kw∼ indicates a labialized k, and kwy$ a labialized k with a palatal offglide. Finally, the modifier ” indicates glottalization or implosion.

Because of the modifiers ∼ and $ ASJPcode is quite versatile, but it also has some limitations. A relatively major limitation is the failure of the system to capture the distinction between retroflex and non-retroflex consonants, a distinction which is a common in South Asia and Australia, for instance. Another limitation, especially worth noting here, is the convention according to which all click sounds are reduced to just one symbol. While this is a severe deficiency, it fortunately applies to a narrowly circumscribed set of languages only, namely those claimed to belong to the so-called Khoisan family in [13].

This is not the proper place to either defend or critizice ASJPcode in any major way. We would like to stress that we see no good reason for the reduction of information caused by such major limitations as the merging of some voicing distinctions or the neglect of retroflection. On the other hand, a possible different strategy of using a full equivalent of the International Phonetic Alphabet (IPA) may not be a viable alternative. It should be kept in mind that the data on which ASJP word lists are based vary in quality and suffer from overall inconsistency. The latter point is particularly important. For some languages IPA-style transcriptions are available, but the vast majority of the data are available either in a practical orthography or in some regionally preferred transcription system (for instance, Africanists and Americanists tend to cater towards different sets of symbols, not all of which belong to the IPA). In this situation the use of a transcription system making fine

Table 1. ASJP consonant symbols.

ASJP symbol	Description
p	voiceless bilabial stop and fricative
b	voiced bilabial stop and fricative
m	bilabial nasal
f	voiceless labiodental fricative
v	voiced labiodental fricative
8	voiceless and voiced dental fricative
4	dental nasal
t	voiceless alveolar stop
d	voiced alveolar stop
s	voiceless alveolar fricative
z	voiced alveolar fricative
c	voiceless and voiced alveolar affricate
n	voiceless and voiced alveolar nasal
S	voiceless postalveolar fricative
Z	voiced postalveolar fricative
C	voiceless palato-alveolar affricate
j	voiced palato-alveolar affricate
T	voiceless and voiced palatal stop
5	palatal nasal
k	voiceless velar stop
g	voiced velar stop
x	voiceless and voiced velar fricative
N	velar nasal
q	voiceless uvular stop
G	voiced uvular stop
X	voiceless and voiced uvular fricative, voiceless and voiced pharyngeal fricative
7	voiceless glottal stop
h	voiceless and voiced glottal fricative
l	voiced alveolar lateral approximate
L	all other laterals
w	voiced bilabial-velar approximant
y	palatal approximant
r	voiced apico-alveolar trill and all varieties of ‘r-sounds’
!	all varieties of ‘click-sounds’

distinctions, such as one between high and mid back rounded vowels (different o-sounds), would induce overdifferentiation with respect to the information usually available. For instance, given a single o-sound a source will usually just use the

Table 2. ASJP vowel symbols.

ASJP symbol	Description
i	high front vowel, rounded and unrounded
e	mid front vowel, rounded and unrounded
E	low front vowel, rounded and unrounded
3	high and mid central vowel, rounded and unrounded
a	low central vowel, unrounded
u	high back vowel, rounded and unrounded
o	mid and low back vowel, rounded and unrounded

symbol o for the mid back rounded vowel rather than the IPA symbol for a low back rounded vowel even if the latter is phonetically more adequate. If ASJPcode were enriched with more symbols for vowel qualities, including a symbol for the low o-sound, transcribers would in the majority of cases not know which symbol to use when encountering an o in a source for lexical data. Thus, ASJPcode often induces a loss of information, but it usually protects against arbitrary transcription decisions. A somewhat more adequate transcription system is imaginable, but for the purpose of transcribing words of all the world's languages given the nature of the sources at hand ASJPcode cannot in a trivial manner be replaced with IPA (or ascii equivalents such as SAMPA or the system used in UPSID).

We now estimate how strongly SR is related to UPSID segment inventory size. For this purpose we first need to match languages in UPSID with languages in ASJP. This matching cannot be perfect by any criterion except by the criterion that the source for the UPSID data should be the same as that for the ASJP such that the two datasets could be said to derive from the same 'doculect' (to use a term which has recently become current and which refers to a language variant as defined by a particular source for its description). Normally this criterion is too strict to be applicable in practice, but a looser version according to which the data should be produced by the one and the same linguist working on the same dialect can sometimes be applied. We have applied this criterion whenever possible. Other less stringent criteria, which were used when the one just mentioned could not be applied, are (in descending order as criterial for a given decision): the data should pertain to the same geographically defined dialect; the variants should have similar names; the ASJP word list chosen from a set of otherwise equally good alternatives should be the more complete one. In a couple of cases UPSID seems to generalize over several languages or variants of languages as listed in [13], such that an average of SRs from several ASJP lists seemed to be the most adequate point of comparison, i.e., Southern and Northern Itelmen and varieties of Dani. The result of this identification process was a matching between 392 of the 451 languages in UPSID with ASJP data. The names (in UPSID and ASJP) and ISO 639-3 codes as well as the full results of the comparisons are given in the Online Supplementary Materials, Table 2.

Here we restrict the results of the comparisons to a statistical summery. For this purpose we exclude a single outlier, the language called !XU in UPSID, which has a great number of click sounds that are disregarded in ASJPcode. It is the only click language in the sample.

The linear correlation between SR and UPSID inventory size, graphically displayed in Fig. 1, is a solid $r = .60$. (Here and elsewhere in the paper we use adjusted r-values). The cone-shaped distribution of the datapoints displays a regularity in the proportion between the two variables. The average ratio of SR to UPSID segment inventory size is .840 with a standard deviation of .195. The ratios between the SR and UPSID inventory sizes across languages are uncorrelated ($r = -.04$) with the number of concepts attested in the ASJP lists. Thus, for all practical purposes we can ignore the number of attested concepts when SR is used as a proxy for total segment inventory size.

As a point of minor interest we note that SRs sometimes exceed UPSID segments in number (cf. cases where dots fall below the dotted line in Fig. 1). The main reason why the number of segments in word lists can apparently exceed the actual number of segments in a language's inventory is that transcribers and/or sources of the ASJP data may apply analyses that differ from those of UPSID, treating complex consonants as single segments through use of the transcriptional modifiers $\sim$ and \$. On the ASJP website, navigable through [27], the sources, the transcribed data, and even transcriber identities are available, making it possible to study such cases in more detail. For instance, for the extreme case of the language called GBARI in ASJP and GWARI in UPSID, where ASJP has 40 segments and UPSID only 26, the transcriber assumed that all combinations of a consonant symbol and the palatal glide (y in ASJPcode) are palatalized single phonemes, that the sequence ts is one phonological unit, and that all consonants followed or preceeded by a nasal are likewise single segments. There are often possibilities for alternative analyses in phonology, and the different analyses of GBARI/GWARI could, at least in principle—we are not going to make an actual judgment in the case—, both have arguments in their favor. As a matter of fact, [17] (p. 6) explicitly says that when there was a choice "between a unit or sequence interpretation of, for example, affricates, prenasalized stops, long (geminate) consonants and vowels, diphtongs, labialized consonants, etc.", then there was "some prejudice in favor of treating complex phonetic events as sequences (i.e. as combinations of more elementary units)."

We conclude from the correlation in Fig. 1 that SRs in ASJP word lists are approximately proportional to segment inventory sizes to a degree where it is meaningful to use SRs as proxies for segment inventory sizes when it comes to investigating correlations with other features, such as word length, population size, and geographical distances—the topics of the next sections. If, for instance, we find a correlation between SR and average word length then this should reflect a similar or even

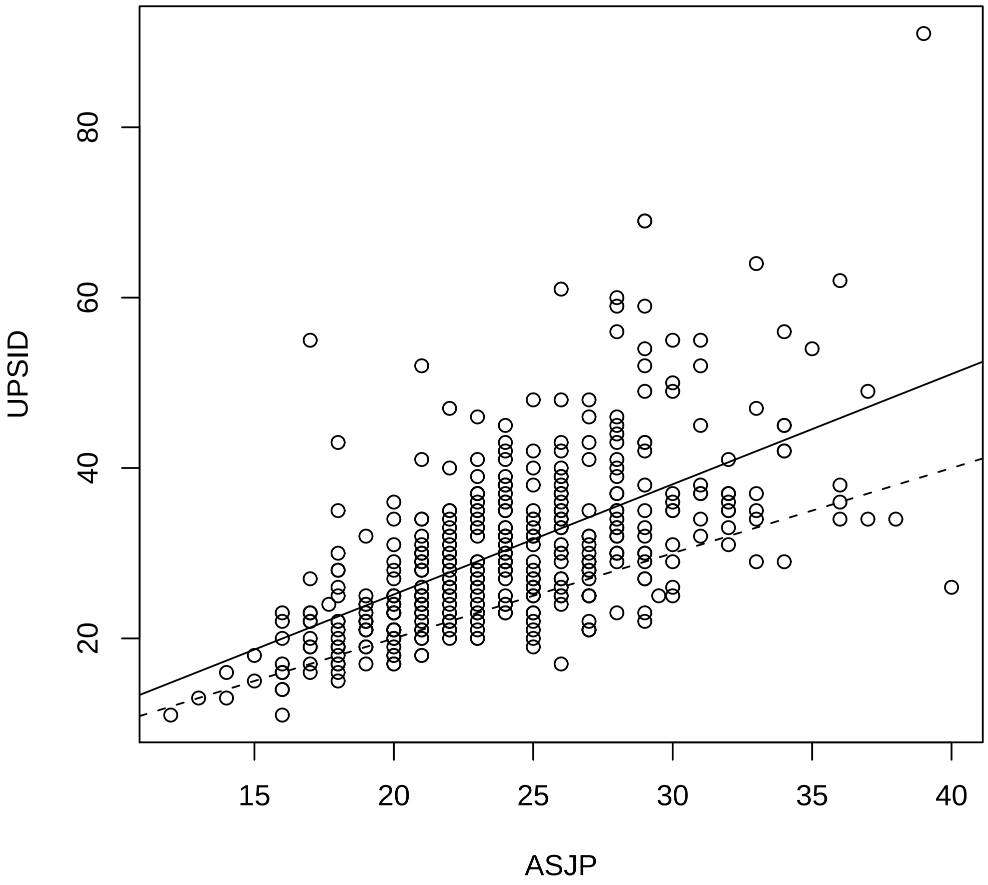

Fig. 1. Numbers of segments in languages that are in both ASJP and UPSID. Solid line is based on a linear model predicting UPSID segment inventory size from the number of segments (SR) in ASJP word lists. Dotted line, with an intercept at (0, 0) and slope of 1, separates cases with respectively more (above the line) or fewer (below the line) UPSID segments than SRs.

stronger correlation between segment size inventories and average word length.

3. SR and word length

In two papers based on different samples of data Nettle [19, 20] (cf. also [21]) argues that word length is inversely correlated with the size of phonological inventories across languages. The latter is defined as the number of phonological segments available, including vowel length and tones (where the number of tones is multiplied by the number of vowels). A mean word length used for the correlation is arrived at by using 50 random dictionary entries, making sure that the dictionaries used were roughly equally sized since larger dictionaries will tend to contain longer words on average. In [19] 10 languages from a world-wide sample are used and in [20] a sample of 12 languages of western Africa. We are interested in testing whether a correlation still holds up when the much larger sample of languages in the ASJP database is used. Before presenting our results we need to discuss aspects of the findings in [19, 20] that make them different from ours in some respects even if the overall success in replicating the findings will turn out to be positive.

The are several reasons why we cannot expect the findings to be completely similar. One obvious reason, already discussed in the previous section, is that we depend on ASJPcode and a count of segment types which is often incomplete. Therefore we expect to find a weaker correlation. Another reason, which has somewhat less obvious ramifications, is that our sample is quantitatively and qualitatively different from those of [19, 20]. The sample of [19] includes languages with some of the world's smallest as well as largest segment inventories, and a smattering of languages covering the range in between these extremes. The West African sample of [20] also seems to be biased towards a coverage of the range of variation in segment inventory sizes (this time of a particular geographical area). In contrast, our sample is not biased in any particular way, but simply contains random representatives of nearly all the world's language families where around one half of the members of each family is in the sample, as described in the beginning of Section 1 above.

Yet another difference is Nettle's use of segment inventory sizes (henceforth S) rather than SR. The distribution of S is positively skewed, with a long upper tail, which leads to a nonlinear relationship with mean word length (henceforth MWL) appromixating a power law. To replicate this finding with a larger set of data we can again use the UPSID data [18] on segment inventory sizes, and the ASJP data will furnish us with information on MWL. Nettle got MWL data for each language from 50 randomly selected dictionary entries, while we will use MWL counts of the ASJP word lists. The lists used for this exercise contain from 24 to 40 words and exactly 37 words on average. Thus, this way of getting MWLs is not vastly different from Nettle's approach. For the count of S we use the number given explicitly in UPSID and do not include tonal distinctions, which is somewhat different from Nettle's count, which does include tonal distinctions. The result of plotting MWL as a function of S is shown in Fig. 2. As in Nettle's studies, the distribution of S is positively skewed and the relationship is nonlinear. The two outliers to the right, which account for much of the nonlinearity, are !XU with 141 segments (in the UPSID count, which, interestingly, is different from Nettle's) and Archi with 91 segments.

Our next step is to use all of the 3168 ASJP languages to count MWL and SR (rather than S) (See data in Table 1 of Online Supplementary Materials). The

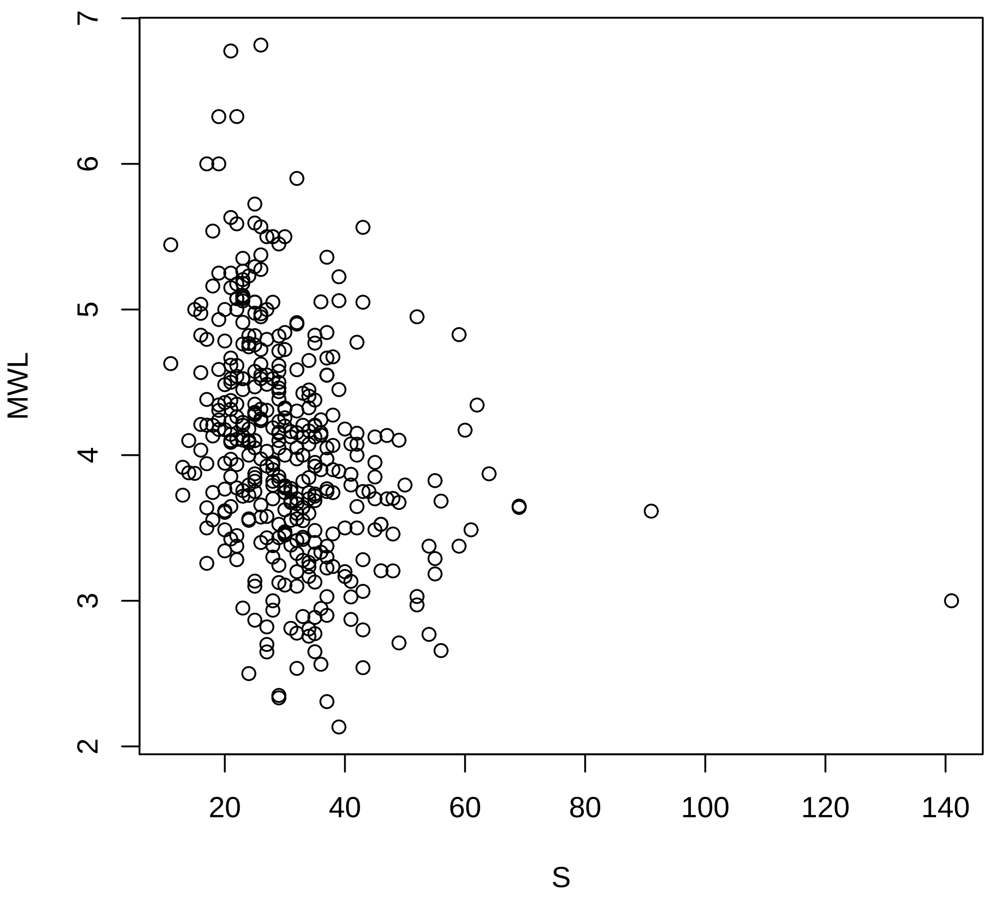

Fig. 2. Mean word length from ASJP data plotted against segment inventory sizes from UPSID.

result, displayed in Fig. 3, shows that the distribution of SR is approximately normal and the relationship is approximately linear, with $r = -.31$. Thus, for all practical purposes we can use linear regression.

The preceding observations making reference to [18–20] had to be made in the absence of significance testing because of a sampling bias in these sources towards including the range of variation in S as well as the lack of a control for areal and genealogical effects. In contrast, a correlation found in the ASJP data can be tested for statistical significance since, as mentioned in Section 1 above, our sample is

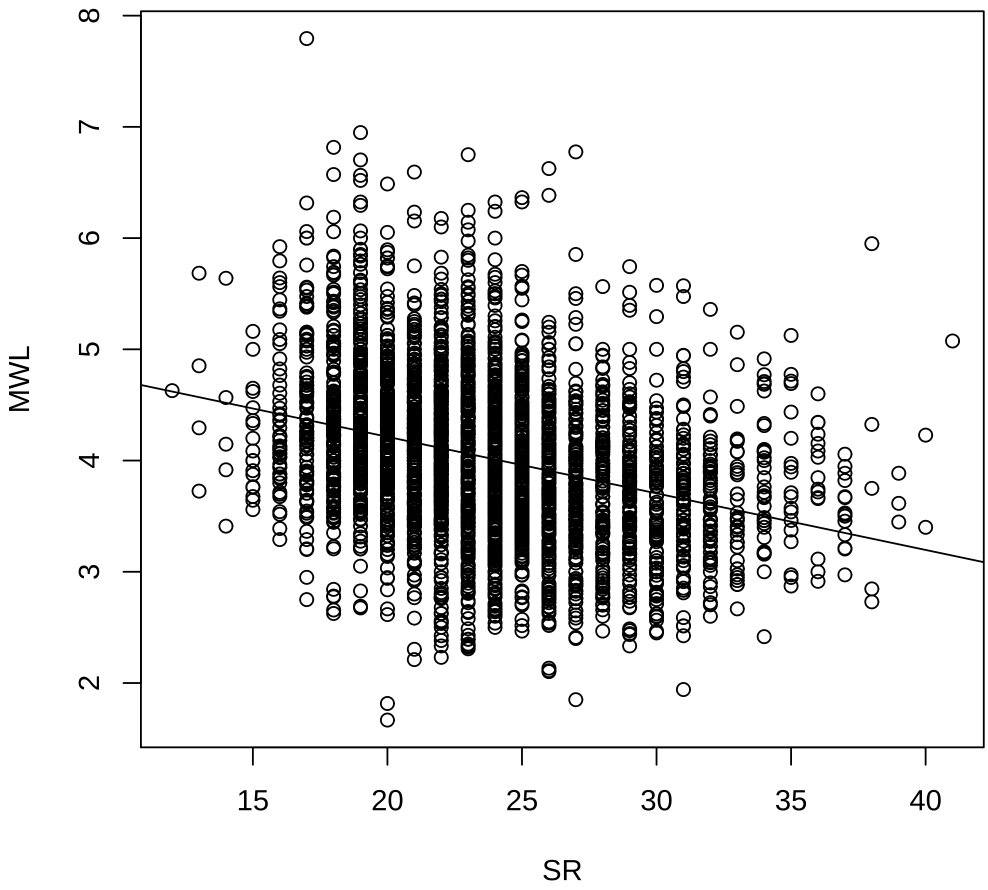

Fig. 3. Mean word length plotted against the number of segments represented, data from ASJP.

unbiased (in the sense that we are using whatever source was available) and also covers around half of the languages in nearly all the world's linguistic families that still have members which are being spoken. To derive a conservative p-value for the correlation between SR and MWL we control for two complicating factors. The first is genealogical: languages in the same family are related through inheritance and cannot be treated as statistically independent. To deal with this problem we take averages of respectively SR and MWL within language families, and then we use families as the units of analysis in the correlations. Visual inspection of frequency histograms of SR and MWL within larger families confirmed that the distributions are approximately normal, justifying taking averages. The data for average MWL (henceforth MMWL) and SR (henceforth MSR) within families are given in Online Supplementary Materials, Table 3, and are plotted in Fig. 4. The correlation is still negative: $r = -.23$ across 157 families, including isolates and unclassified languages.

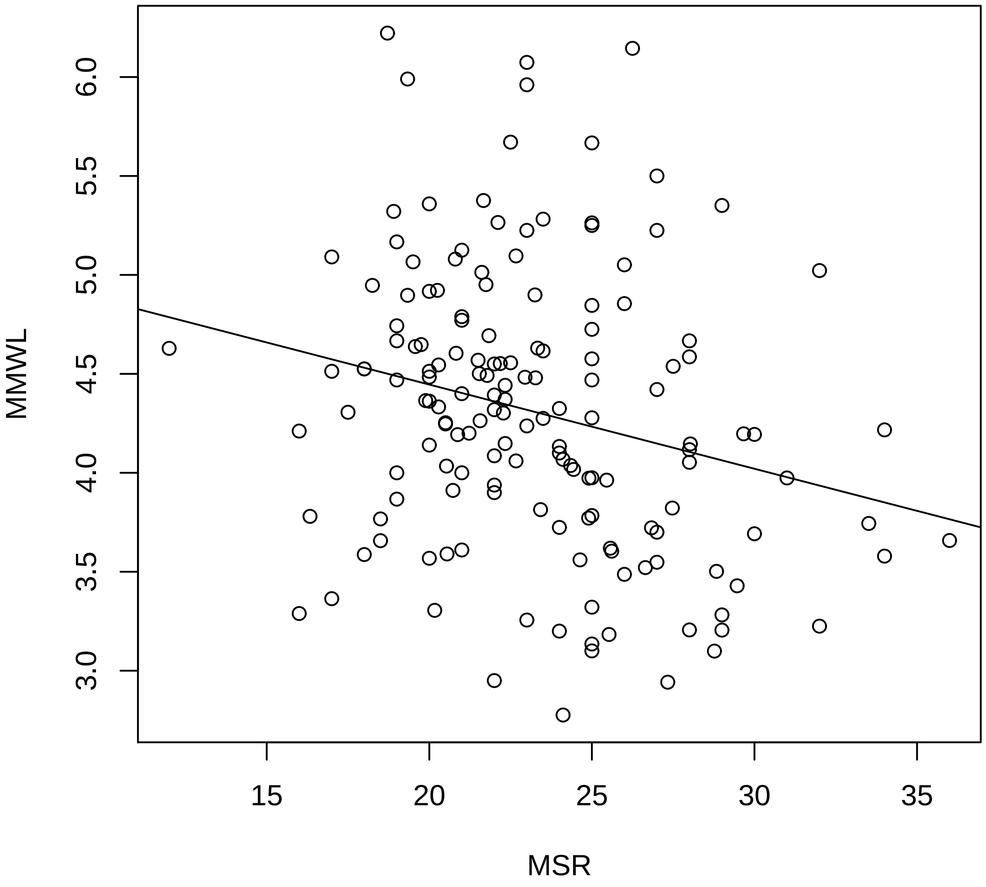

Fig. 4. Mean word length plotted against mean SR for individual families (using ASJP data only and including all isolates and unclassified languages in the sample).

The second complicating factor is geographic: processes such as language contact and migration tend to increase the similarity between geographically contiguous languages even if they are genealogically unrelated [**?**]. In particular, there is smaller variation within geographic macro-areas as defined in [8] than in the world at large and enough variation between macro-areas to require us to take areas into account. This point is illustrated by the boxplots shown in Figs. 5-6. (For convenience all languages of a given family are assigned to a single macro-area, even in the few cases where a family extends over two areas, such as Misumalpan and Austro-Asiatic, where the macro-area containing the majority of the languages is chosen to be representative of the family at large.)

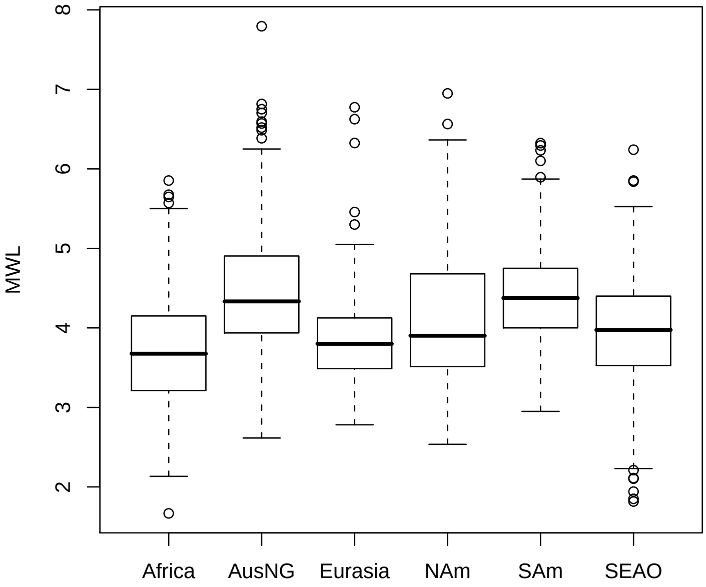

Fig. 5. Box plots of MWL for languages pertaining to the six macro-areas of [8].

We are now, in principle, equipped to establish the p-value for the correlation between MSR and MMWL. To control for areal effects we treat the macro-area to which each family belongs as a random effect in a linear mixed model using [2, 3]. The p-value is estimated using the MCMC method implemented in R [24] as the pvals.fnc function of [2]. The correlation in Fig. 4 proves to be significant (p = .0011). This correlation may be overly conservative because it is based on all families, including those containing a single language in our sample, as well as

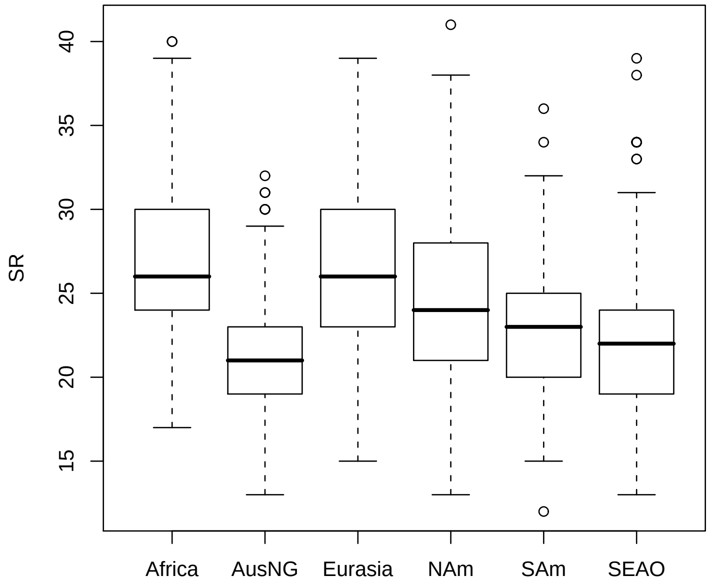

Fig. 6. Box plots of SR for languages pertaining to the six macro-areas of [8].

isolates and unclassified languages. MSR and MMWL in small families are based on small samples and are therefore subject to more random sampling error than in large families, which will weaken the correlation. In fact, the correlation tends to grow when smaller families are excluded. Following the practice of [1] we can use the criterion that a family must contain at least two members to be included. Across the 90 families that satisfy this criterion $r = -.28$ for the linear correlation between MMWL and MSR and $p = .0026$ for MSR as a predictor variable in the linear mixed model.

In conclusion to this section, mean word length and segment inventory sizes are significantly correlated when using a genealogically and geographically balanced sample containing close to one half of the world's languages—even when the data are somewhat impoverished because of simplified transcription and samples of segments from word lists rather than full inventories. Thus, in general, we can confirm the findings of Nettle [19, 20], although the size of the correlation in the larger sample is not as great as reported in these papers.

4. SR and population sizes

Early studies of speaker population sizes and phoneme inventory sizes relied mainly on anecdotal evidence. Haudricourt [9] discussed extra-linguistic factors that may influence inventory sizes, such as geographical isolation leading to the maintenance of large inventories and bilingualism, which may lead to the mergers of some contrasts (cf. [12] for support), but which could also imply the introduction of new sounds from other languages (cf. also [22]). He noted that smaller populations are more likely to have a high proportion of bilinguals than larger populations, but did not explicitly state whether this should lead us overall to expect larger or smaller inventories in small populations—the effect of biligualism can pull in both directions. Trudgill [23] qualified this model of opposed effects further, suggesting that it is mainly bilingualism among children which may lead to enriched inventories, whereas it is mainly adult bilingualism which causes simplification.

Hay & Bauer [10] were the first to report a statistically validated positive correlation between speaker population sizes and phoneme inventory sizes. The sample of [10] consists of 216 languages, and the selection ultimately derives from a textbook [4] where languages were chosen such as be widely representative of different areas and linguistic families or simply to be of special interest to a linguistics student. The authors report that Spearman's $rho = .37$ for the correlation of total phoneme inventory sizes and logarithms of population size across the total set of languages. A low value of $p < .0001$ is also given, but is not to be trusted because of failure to control for interdependence of datapoints. They also present an analysis factoring in language family as an independent variable, where families for which 7 or more languages were available were treated as separate groups, and where other datapoints were lumped together in an 'other' category. The result is a correlation of $r = .49$. Finally, the correlation was also tested by using means of logarithms of population sizes and phoneme inventory sizes for each family, giving $rho = .46$, $p = .003$. The authors discuss possible factors such as the ones mentioned in the previous paragraph that may cause population sizes and phoneme inventory sizes to be intertwined, but refrain from pushing any particular explanation.

Atkinson [1] finds support for Hay & Bauer [10] and furthermore identifies an overall negative correlation between phoneme inventory sizes and the distance of languages from Africa. The two observations are brought to bear on one another by Atkinson, but presently we will focus on the first observation only, while the second will be the topic of our next section. Atkinson uses WALS [8] data for his study. It is to be noted that [8] operates with categorical values for segment inventory sizes rather than absolute numbers, e.g., consonant inventory sizes are put in categories from 'small' to 'large' with 3 intermediate categories. In order to arrive at values for the total inventories, [1] combines information from three different WALS chapters [14–16]. For an uncontrolled correlation between phoneme size inventories and log population sizes based on 503 languages [1] reports that $r = .39$, and an analysis which controls for genealogical relatedness using language family means yields $r = 0.47$, $df = 49$, $p < 0.001$ among families and also an effect within families. The results are similar to those of [10].

Following [10] and [1] we replicate these results, first by finding the uncontrolled correlation for our total sample of single languages, and then by controlling for genealogical relatedness using averages over language families. The result of the first step is plotted in Fig. 7. Our sample of 3153 languages having 1 or more speakers is more than 14 times as large as the sample of [10] and more than 6 times as large as that of [1]. We get a correlation which is weaker ($r = .22$), either because of the nature of the ASJP data or the more complete sample or a combination of these two factors.

To test the significance of the relation between and SR and the log of population sizes we now average over families. To stay on the conservative side we use all the 90 families with two or more members as in [1]. A slight gain in correlation would be gotten by using families with more than 6 members as in [10], but it is not clear which criterion to apply when excluding data points other than those representing one-member families. The result, plotted in Fig. 8, shows a correlation of $r = .19$, which is statistically significant, even if not highly so ($p < .043$).

The fact that we get a significant correlation between population size and the number of unique segments in ASJP word lists, the latter serving as a proxy for segment inventory sizes, supports the claim of a relation between these two variables in [10] and [1]. Due to the nature of the ASJP data our results are somewhat inconclusive as regards the magnitude of the correlation, which is likely somewhat higher than what we find, and most likely somewhere in between our $r = .19$ and Atkinson's $r = .47$. But given the size of our sample we can firmly support the *existence* of the correlation first identified by Hay & Bauer [10].

5. SR and geography

The purpose of this final section is to check the claim in [1] that segment inventory sizes tend to be smaller the further removed a language is from Africa. As in the preceding sections we need to average within families as one of the requirements for establishing statistical independence of datapoints. In the following we briefly describe the details of how we proceed.

In order to choose a single geographical coordinate for each family, [1] used the location of a centroid language. We use a more principled approach, taking the putative homeland as inferred by the method of [26], which identifies the homeland of a given family with the language which is most diverse in the specific sense defined in [26]. The difference in approaches has negligible effects since geographical ranges of language families are small in comparison to the distances between the various populated continents and Africa.

We more-or-less arbitrarily choose Addis Ababa as the point of origin of human kind within Africa. This choice does not introduce a bias in the hypothesis-testing, because the location of Addis Ababa is roughly equidistant to the coordinates that

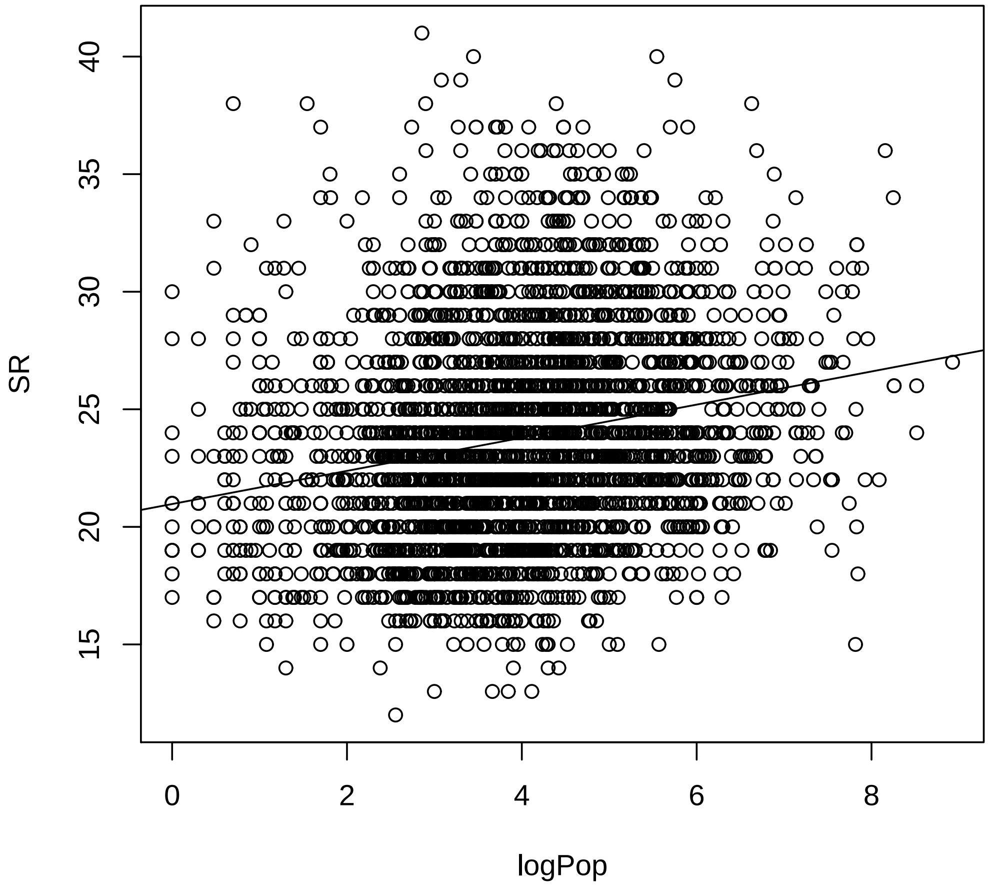

Fig. 7. Segments Represented in ASJP word lists for individual languages with one or more speakers plotted against log of population sizes from [13].

we use for three of the African families (3367 km from Afro-Asiatic, 3862 km from Khoisan, and 3676 km from Niger-Congo), while having a relatively short distance (1099 km) to the family with the smallest MSR (Nilo-Saharan, with a MSR of 25.45). A place of origin favoring the hypothesis of an inverse correlation between distance from the origin and phoneme size inventories would be closer to Khoisan and Afro-Asiatic, which have the largest MSR (respectively 28.77 and 28.03), or one could simply choose a best-fit origin as in [1].

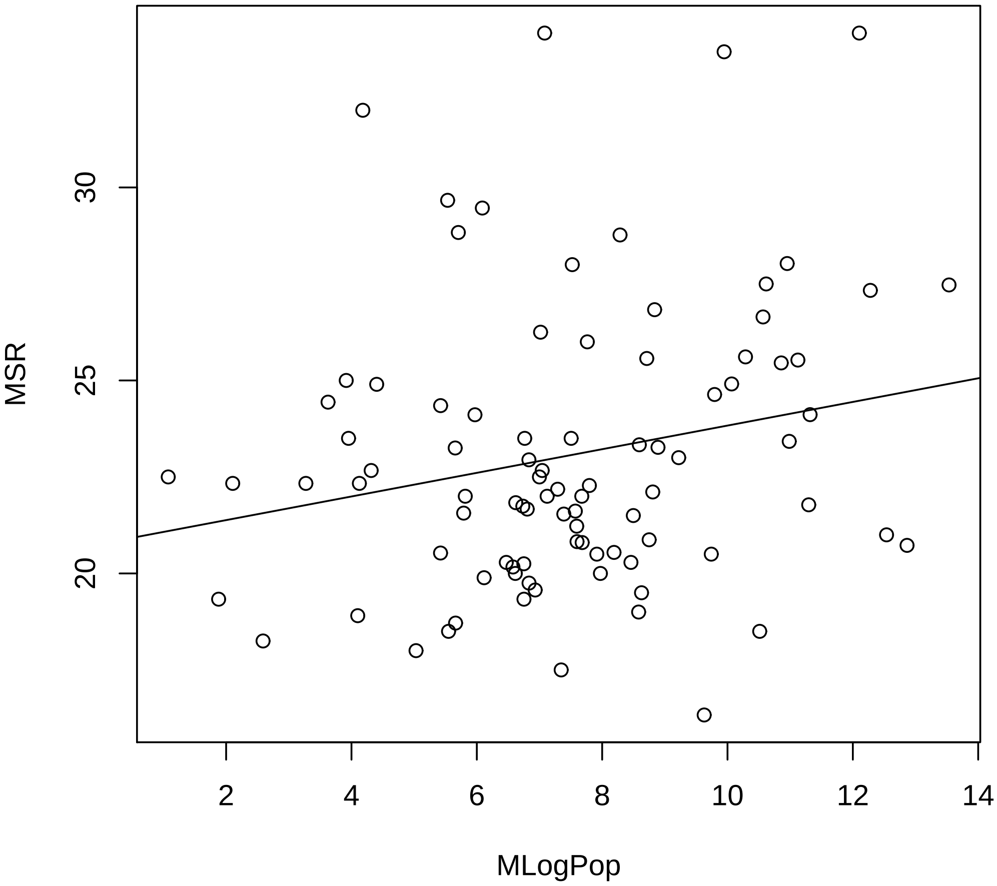

Fig. 8. Mean of Segments Represented in ASJP word lists within families with 2 or more speakers plotted against the mean of log of population sizes.

Other than in these details our approach is similar to that of [1]. We use families for which the sample includes at least two members, and we constrain migrations in a simplified manner using the waypoints of [1], i.e., Cairo, Istanbul, Phnom Penh, Bering Strait, and Panama. The major difference is in the datasets, where ours includes 90 families with a total of 3059 languages and that of [1] which includes 50 families with a total of 445 languages. As usual, a further difference concerns our use of SR as a proxy for phoneme inventory sizes. The results are plotted in Fig. 9.

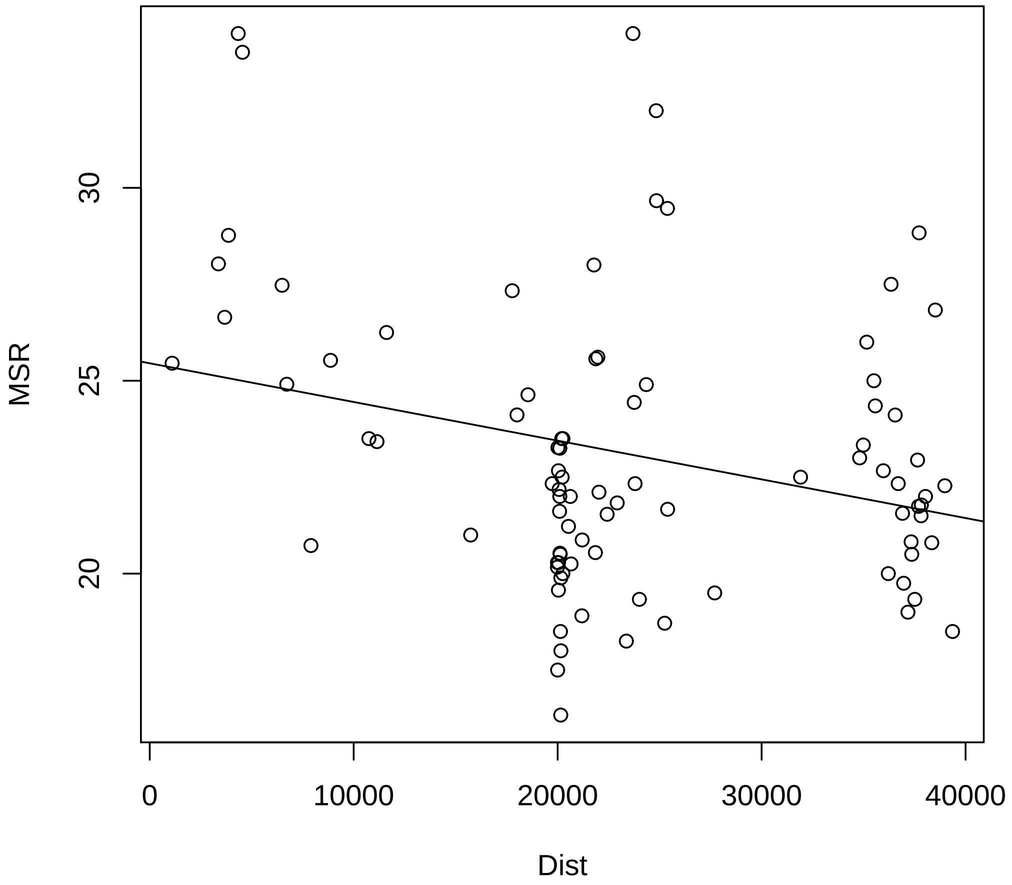

Fig. 9. Mean of Segments Represented in ASJP word lists within families with 2 or more speakers plotted against the distance (in km) from Addis Ababa.

A simple regression of the mean of the logarithm of population size and distance from Addis Ababa (henceforth 'Africa') for the same language families gives $r = -0.34$, $p = 0.0005$. This correlation is larger and most robust than that of MSR in relation to distance from Africa, which raises the question whether the inverse correlation between MSR and distance from Africa is a conspiracy of the facts that population sizes diminish with the distance from Africa (using language families as units of analysis) plus the fact that population size is correlated with MSR. The statistical significance of the relation plotted in Fig. 9 is therefore assessed through a multiple regression with MSR as the dependent variable and distance and log population size as predictor variables. This produces $R^2 = .072$ for the overall multiple regression, with $p = .0369$ for distance and $p = .2337$ for log population. Thus, the multiple regression confirms distance as a significant predictor of MSR with population size controlled.

What we have found, then, is that an analysis similar to Atkinson's [1] confirms his claim that phoneme inventory sizes really do seem to overall grow smaller as the distance from Africa increases. The question remains whether this is a secondary effect of a preference for longer words as the distance from Africa increases. This possibility is tested by another multiple regression, with MSR as the dependent variable and MMWL, distance, and log population size as predictor variables. This produces $R^2 = .106$ for the overall multiple regression, with $p = .0410$ for MMWL, $p = .0806$ for distance, and $p = .5585$ for log population. The one significant effect confirms the negative relation between MMWL and MSR, with population as well as distance controlled. The effect of distance is not significant, suggesting that its relation to MSR may be indirect, but it is not so far from significance as to refute Atkinson's interpretation. The role of MMWL is investigated by a final multiple regression, with MMWL as the dependent variable and MSR, distance, and log population size as predictor variables. This produces $R^2 = .163$ for the overall multiple regression, with $p = .0410$ again for MSR, $p = .2457$ for distance, and $p = .0141$ for log population. The new significant effect is a negative relation between population and MMWL, with MSR and distance controlled. This effect, along with the lack of a significant effect of population on MSR in the previous regression with MMWL and distance controlled, suggests that MMWL may mediate the correlation observed between population and MSR.

6. Discussion & Conclusion

In this paper we have tested different claims in the literature and were able to confirm that languages tend to have larger phoneme inventories when they have shorter words (or the other way around), that larger populations are associated with larger phoneme inventories (and therefore shorter words), and that, finally, phoneme inventories diminish with the distance from Africa. Multiple regression analyses suggest, however, that some of these relations may be indirect. How might one best account for these relations?

The relation between word length and phoneme inventory sizes is relatively straightforward. When the number of phonemes available decreases such that the probability for homonymy increases it makes sense that words (i.e., lexical roots or stems) should grow longer. We can empirically observe a lower limit to the number of phonemes that a language can do with, cf. a language such as Rotokas with 11 phonemes according to [18], and most languages prefer to not stay close to the limit, but rather to have a surplus of expressive means. Inversely, if for some reason—for instance through phonological erosion—a language undergoes a change towards shorter words it may need to increase its inventory of phonemic distinctions. Chinese is an example where this sort of development is historically documented. Finally, it is also reasonable to expect that a change towards longer words (derivations, compounds) can lower the pressure on the phoneme inventory or that the acquisition of new phonemes can lower the pressure on the word formation.

It is harder to explain why there is a positive correlation between population size and phoneme inventory size. Hay & Bauer [10] refrained from insisting on an explanation. They failed to cite Nettle [19, 20], where the key to a possible explanation might be found. If word roots and stems tend to get regularized towards shorter canonical forms in larger populations this would account for the correlation. This explanation is consistent with the mediating role played by mean word length in the multiple regressions.

As regards the inverse correlation between phoneme inventory sizes and distance from Africa Atkinson [1] attempts to explain this in terms of prehistorical bottlenecks. If, along migration routes, migrants pass barriers that reduce populations this could have an effect on phoneme inventory sizes, given that the two seem to be correlated. We are not quite convinced that this explanation holds. One problem with it is that prehistorical societies would have been small in any case, whether they had to pass a bottleneck or not. Moreover, a bottleneck in the normal genetic usage, which is also appropriate for the current discussion, is a historical event leading to the reduction in diversity, where only a small population is singled out for further reproduction. In a historical linguistic context this could mean the survival or passage through a migration point of a single language which could be unrepresentative of the total linguistic diversity. Such a language might have any number of phonemes. Thus, for instance, if the language(s) which made it to the Americas happened to have large phoneme inventories then this characteristic would be transferred to modern descendants of the language. Indeed, the Northwest Coast of the Americas is famous for languages having large phoneme size inventories. Thus, we doubt that bottleneck effects could pull in a specific direction. Rather, they would seem to upset any kind of regularity in the development towards smaller or greater phoneme inventory sizes.

In an alternative explanation for the out-of-Africa correlation it is possible to leave out population sizes altogether. Given a wave-model of migrations where each successive wave of migrating populations passes by previous populations the ancestors of the languages which are currently most removed from Africa would have passed by more populations than the ancestors of languages which are currently less removed. From the discussions in [9] and [23] we can pull opposed hypotheses about the effect to expect on phoneme size inventories from language contact: language contact can lead to the acquisition of new phonemes, particularly with children as second language learners, or to the loss of phonemes, particularly with adults as second language learners. If an explanation of the out-of-Africa correlation is to be based on language contact it would seem that L2-learning by adults had a more dominant effect than L2-learning by children. Thus, a model of successive events of language shift by adults as they migrated along the routes from Africa to the rest of the world, passing by populations already *in situ*, would account for the weak but nevertheless statistically significant decline in phoneme inventories as the distance from Africa increases.

References

[1] Atkinson, Q. D., Phonemic diversity supports a serial founder effect model of language expansion from Africa, *Science* **332** (2011) 346–349.

[2] Baayen, R. H., languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics", R package version 0.955 (2009).

[3] Bates, D. M. and Maechler, M., lme4: Linear mixed-effects models using S4 classes, R package version 0.999375-32 (2009).

[4] Bauer, L., *The Linguistics Student's Handbook* (Edinburgh University Press, 2007).

[5] Brown, C. H., Holman, E. W., Wichmann, S., and Velupillai, V. Automated classification of the world's languages: A description of the method and preliminary results. *STUF–Language Typology and Universals* **61** (2008) 285–308.

[6] Haspelmath, M., Dryer, M., Gil, D., and Comrie, B. (eds.), *The World Atlas of Language Structures Online* (Max Planck Digital Library, Munich). http://wals.info/, accessed on 2011-05-09.

[7] Dryer, M. S. Large Linguistic Areas and Language Sampling, *Studies in Language* **13** (1989) 257–292.

[8] Haspelmath, M., Dryer, M. S., Gil, D., and Comrie, B. (eds.), *The World Atlas of Language Structures* (Oxford University Press, 2005).

[9] Haudricourt, A. G., Richesse en phonèmes et richesse en locateurs, *l'Homme* **1** (1961) 5–10.

[10] Hay, J. and Bauer, L. Phoneme inventory size and population size, *Language* **83** (2007) 388–400.

[11] Holman, E. W., Schulze, C., Stauffer, D., and Wichmann, S. On the relation between structural diversity and geographical distance among languages: observations and computer simulations, *Linguistic Typology* **11** (2007) 395–423.

[12] Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., Brown, P., and Bakker, D. Explorations in automated language classification, *Folia Linguistica* **42** (2008) 331–354.

[13] Labov, W., *Principles of Linguistic Change* (Blackwell, 1994).

[14] Lewis, M. P. (ed.), *Ethnologue*, 16th ed. (SIL International, 2009).

[15] Maddieson, I., Consonant inventories, in *The World Atlas of Language Structures Online*, eds. Haspelmath, M., Dryer, M., Gil, D., and Comrie, B. (Max Planck Digital Library, Munich). http://wals.info/, accessed on 2011-05-07.

[16] Maddieson, I., Vowel quality inventories, in *The World Atlas of Language Structures Online*, eds. Haspelmath, M., Dryer, M., Gil, D., and Comrie, B. (Max Planck Digital Library, Munich). http://wals.info/, accessed on 2011-05-07.

[17] Maddieson, I., Tone, in *The World Atlas of Language Structures Online*, eds. Haspelmath, M., Dryer, M., Gil, D., and Comrie, B. (Max Planck Digital Library, Munich). http://wals.info/, accessed on 2011-05-07.

[18] Maddieson, I., *Patterns of Sounds* (Cambridge University Press, 1984).

[19] Maddieson, I. and Precoda, K., *UCLA Phonological Segment Inventory Database*. http://web.phonetik.uni-frankfurt.de/upsid.html, accessed on 2011-05-09.

[20] Nettle, D., Segmental inventory size, word length, and communicative efficiency, *Linguistics* **33** (1995) 359–367.

[21] Nettle, D., Coevolution of phonology and the lexicon in twelve languages of West Africa, *J. Quantitative Linguistics* **5** (1998) 240–245.

[22] Nettle, D., *Linguistic Diversity* (Oxford University Press, 1999).

[23] Nichols, J., *Linguistic Diversity in Time and Space* (University of Chicago Press, 1992).

[24] Trudgill, P., Linguistic and social typology, in *The Handbook of Language Variation and Change*, eds. Chambers, J. K., Trudgill, P., and Schilling-Estes, N. (Oxford, Blackwell, 2002), pp. 707–728.

[25] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

[26] Wichmann, S., Holman, E. W., and Brown, C. H. Sound symbolism in basic vocabulary. *Entropy* **12** (2010) 844–858.

[27] Wichmann, S., Müller, A., and Velupillai, V., Homelands of the world's language families: a quantitative approach, *Diachronica* **17** (2010) 247–276.

[28] Wichmann, S., Müller, A., Velupillai, V., Brown, C. H. Holman, E. W., Brown, P., Sauppe, S., Belyaev, O., Urban, M., Molochieva, Z., Wett, A., Bakker, D., List, J.-M., Egorov, D., Mailhammer, R., Beck, D., and Geyer, H., *The ASJP Database (Version 13).* http://email.eva.mpg.de/ wichmann/languages.htm, accessed on 2011-05-09.