PRE-PUBLICATION MANUSCRIPT. FINAL VERSION AT http://dx.doi.org/10.1353/lan.2012.0087

Revisiting population size vs phoneme inventory size

Steven Moran, University of Washington *
Daniel McCloy, University of Washington
Richard Wright, University of Washington

stiv@uw.edu
drmccloy@uw.edu
rawright@uw.edu

* corresponding author

PRE-PUBLICATION MANUSCRIPT. FINAL VERSION AT http://dx.doi.org/10.1353/lan.2012.0087

Revisiting population size vs phoneme inventory size

PRE-PUBLICATION MANUSCRIPT. FINAL VERSION AT http://dx.doi.org/10.1353/lan.2012.0087

ABSTRACT. In this paper we argue against the findings presented in Hay & Bauer 2007, which show a positive correlation between population size and phoneme inventory size. We argue that the positive correlation is an artifact of the authors' statistical technique and biased data set. Using a hierarchical mixed model to account for genealogical relatedness of languages, and a much larger and more diverse sample of the world's languages, we find little support for population size as an explanatory predictor of phoneme inventory size once the genealogical relatedness of languages is accounted for.[*]

Keywords: typology, phoneme inventories, population size, sampling, mixed models

*ACKNOWLEDGMENTS. The development of PHOIBLE was partially funded by the University of Washington's Royalty Research Fund. We would also like to thank contributors to the project: Morgana Davids, Scott Drellishak, David Ellison, Richard John Harvey, Kelley Kilanski, Michael McAuliffe, Kevin Pittman, Brandon Plasters, Cameron Rule, Daniel Smith, and Daniel Veja, as well as Marilyn Vihman for providing the Stanford Phonology Archive data. Tristan Purvis and Christopher Green were integral in curating many of the phoneme inventories from African languages. We would also like to thank Paul Sampson, Theresa Smith, and Donghun Kim for statistical consultation, and the editor of *Language* and two anonymous reviewers for helpful comments. Any remaining errors are of course our own.

PRE-PUBLICATION MANUSCRIPT. FINAL VERSION AT http://dx.doi.org/10.1353/lan.2012.0087

1. INTRODUCTION. This paper addresses the relationship between a non-linguistic factor (population size) and a property of languages (phoneme inventory size). Studies of the relationship between linguistic and non-linguistic structures date back at least a century, when Sapir (1912) delineated influences on language in the physical and social environments. Sapir found the influence of non-linguistic factors to be most clearly reflected in a language's vocabulary, but also found it to affect the phonetic system and grammatical forms of languages. Despite any criticisms one might level at Sapir's methods, reasoning, or the representativeness of his sample of languages, it seems clear that certain non-linguistic contexts clearly favor differential enrichment of the lexicon, evidenced by the uneven distribution of domain-specific vocabulary in relation to the importance of those domains for different linguistic communities (cf. Nettle 1999a).[1]

Later, attempts to quantify environmental influences on language emerged. Trudgill 1974 introduced the gravity model from geography to dialectology, proposing that faster change occurs in geographically close dialects with more speakers, and slower change in geographically distant dialects with fewer speakers. Several other studies have investigated the influences of non-linguistic structures on language change, and many have proposed that diachronic linguistic change is in fact affected by social and environmental factors beyond just language contact (Haudricourt 1961, Nettle 1996, Trudgill 1989). Recently, research using statistical methods and typological databases has furthered the view that changes in language structure are not purely linguistically driven. These lines of research suggest that some typological patterns (be they synchronic or diachronic) may be related to (or even a consequence of) societal factors (see e.g. Atkinson 2011, Fought et al. 2004, Hay & Bauer 2007, Lupyan & Dale 2010, Munroe et al. 2009).

Among these studies, the question of whether speaker population affects linguistic structure or rate of language change has been intensely debated (Bakker 2004, Nettle 1999b, Nettle 1999c, Pericliev 2004, Trudgill 1996, Trudgill 1997, Trudgill 2002, Trudgill 2004a, Wichmann & Holman 2009, Wichmann et al. 2008, Donohue & Nichols 2011, Wichmann et al. 2011). Phonemic inventory size has often been the metric used in such studies, because it is easier to quantify than other linguistic attributes such as morphological structure. In this paper, we address several questions about the relationship between speaker population size and properties of the

phonemic inventory. We take as our starting point the short report by Hay and Bauer (2007), and set out to reproduce the correlations they found, using a larger data set (969 languages compared to Hay and Bauer's 216) and more nuanced statistical analysis techniques that a larger sample size affords. Our aim is not merely to add another voice to the debate, but also to illustrate what we think are the right ways to ask such questions, and to illuminate some of the pitfalls of previous studies.

2. PREVIOUS STUDIES.[2] One way in which societal factors have been claimed to influence linguistic structures is through language contact. Speculation on the correlation between language contact and phoneme inventory size began at least as early as 1961, when Haudricourt argued that small inventories are the product of impoverishment that is characterized by monolingualism, isolation, and/or by non-egalitarian bilingualism. Trudgill's work further argues in support of the influence of language contact on linguistic structure (particularly phonology). Trudgill presents a typology in which isolated low-contact languages (e.g. Hawai‘ian) tend to have small inventories, as do short-term high-contact situations lacking widespread bilingualism (e.g. pidgins); in contrast, long-term high-contact situations with child bilingualism (e.g. Ubykh) are claimed to tend toward large inventories (Trudgill 1997:356, Trudgill 1996). Later, Trudgill (2002, 2004a) expands this picture, reasoning that other social factors such as social network structure, amount of shared information among speakers, and community size undoubtedly play a role.

Regardless of the particular pattern argued for, what is important to note about the aforementioned studies is their reliance on evidence from case studies. There are at least two problems with this approach. First, patterns appearing in very small samples may not extend to larger samples. For example, if one looks at the Austronesian languages, they tend to have small, isolated communities and small phoneme inventories, and it is tempting to generalize this observation to all languages.[3] This can lead to the second problem with case-based reasoning: confirmation bias. That is, once a pattern has been identified, it is all too easy to find evidence everywhere one looks, while discarding conflicting cases as outliers or insignificant exceptions. Nevertheless, forming hypotheses from case-based observations is a reasonable first step in investigating these kinds of questions, but the crucial next step in a scientific approach to

language study is careful empirical analysis of as much data as can be gathered that bears on the question at hand. Without this further step, the claims of Haudricourt, Trudgill, and others can be no more than hypotheses — albeit thoughtful, refined, and educated hypotheses.

In contrast to the case-based approach, studies have emerged that test the relationship between linguistic and non-linguistic factors using computer simulations (particularly for examining language change). Nettle (1999b, 1999c) examines the relationship between community size and language change using computer simulations modeled on Social Impact Theory (Nowak et al. 1990, see also Nettle 1999c). Based on his simulations, Nettle argues that rate of language change, borrowing, and the emergence of marked structures are less likely to occur as the population gets larger. Wichmann et al. 2008 revisits Nettle's results, but whereas Nettle models competition between only two languages or two linguistic features (the original and the novel forms), Wichmann and colleagues use a simulation model allowing several language forms (each with several linguistic features) to compete simultaneously. Wichmann and colleagues also analyze a sample of 2140 languages with data from Ethnologue 15 (Gordon 2005) and the World Atlas of Language Structures (WALS; Haspelmath et al. 2005), for which they estimate the stability of each of 134 WALS features and use the stability of features to estimate rate of linguistic change for each language. The results from their empirical study suggest that speaker population has no correlation with rate of linguistic change, whereas their simulations show both the presence and absence of some correlation, depending on whether linguistic diffusion is allowed to be global or if it is constrained to near neighbors in the social network. In more recent work, Wichmann and Holman test several different empirical data sets and statistical methods, and their findings 'strongly indicate that the sizes of speaker populations do not in and of themselves determine rates of language change' (Wichmann & Holman 2009:272).

A third approach to examining the relationship between linguistic and non-linguistic factors is to model the relationship statistically. Studies that directly test the relationship between speaker population and phoneme inventory size draw on a variety of sampling and statistical approaches with sometimes contradictory results (Bakker 2004, Pericliev 2004, Donohue & Nichols 2011, Hay & Bauer 2007, Wichmann et al. 2011). Bakker's study reexamines Trudgill's claims about the effects of language contact on phonological inventories, whereas Pericliev's

study targets Trudgill's claims about correlations between community size and phonological inventories. Both studies cast serious doubt on Trudgill's hypotheses, whereas Hay and Bauer's study seems to at least partially support Trudgill's claims; using a sample of 216 languages, they find correlations between speaker population and various measures of phonological inventory size (number of obstruents, number of monophthongs, etc). Although their correlations are modest (the Spearman coefficients they report range from 0.17 to 0.37), Hay and Bauer's findings seem to support the positive relationship between population and inventory size. On the other hand, Donohue and Nichols (2011) draw the opposite conclusion; they find that although correlations may be found within geographic areas, there is no worldwide correlation. Similarly, the findings in Wichmann et al. 2011 also seem to contradict Hay & Bauer 2007. Nevertheless, while their results draw on a much larger number of languages, Wichmann and colleagues estimate phoneme inventory size from Swadesh lists rather than descriptive grammars and use a coding scheme that collapses a number of phonemic contrasts, so the veracity of their data is perhaps subject to skepticism. In light of these conflicting results, we are still left with the question of whether speaker population and phoneme inventory size are correlated or not.

3. METHODOLOGICAL CONSIDERATIONS. There are a variety of methodological considerations that bear mentioning in light of the preceding discussion. First and foremost is the issue of sampling: each of the statistical studies mentioned above uses a different sample of languages. Wichmann and colleagues (Wichmann et al. 2008, Wichmann & Holman 2009) use a sample of 2140 languages drawn from WALS, Pericliev uses 428 languages drawn from the UPSID-451 database (Maddieson 1984, Maddieson & Precoda 1990), and Hay and Bauer use a sample of 216 languages drawn from Bauer 2007. Of these, Hay and Bauer's sample is decidedly non-random, as the aim of Bauer 2007 was to include languages of interest (widely spoken languages, well-known isolates, and languages exhibiting some typological rarity).[4]

Bakker's sample is also non-random, as his pilot study is effectively a series of case studies designed to illuminate the commonality of outliers with respect to Trudgill's hypotheses about language change. Though it would seem the WALS sample used by Wichmann and colleagues is the largest (and therefore the best) one, it should be noted that the data are incomplete, and Wichmann and colleagues divide the sample into four groups based on speaker population and

use the data available within each group to estimate the stability of linguistic forms for each group as a whole, rather than for each language individually. Furthermore, phonological inventories in WALS are categorized as 'small, average, and large' vowel quality inventories, and 'small, moderately small, average, moderately large, and large' consonant inventories (Maddieson 2011a, 2011b). Thus the smaller sample provided by the UPSID database and used by Pericliev is in some sense the best sample, in that it contains numerical (not categorical) data about phonological inventory size. UPSID also attempts to be genealogically balanced by including only one language from each 'small family grouping' (Maddieson 1984:5), even though such balance comes at the expense of failing to capture the typological diversity within each group.[5]

This sacrifice of typological diversity in UPSID is one example of another methodological challenge for studies like these: the avoidance of bias. The case of Swahili in UPSID (see Note 5) has been described as a BIBLIOGRAPHIC BIAS,[6] stemming from the fact that typological samples tend to include data from languages and language families that are well documented (Rijkhoff & Bakker 1998) and as many as two-thirds of all languages have no grammar or grammatical sketch (Bakker 2011).[7] A similar problem is purpose-built into the Bauer 2007 data set used by Hay and Bauer (2007, see Note 4), which has the additional problem of over-representing certain language families (notably Indo-European) and under-representing others (e.g. Niger-Congo; see below and Figure 4 for further discussion of the representativeness of Hay and Bauer's sample).

Further complicating the sampling problem is the inherent uncertainty surrounding the genealogical relatedness of languages. There is no radiocarbon dating for languages like there is for cultural artifacts or biological remains, and the malleability of language (both in isolation and in situations of contact) makes it difficult to distinguish similarities due to shared descent, areal diffusion, convergent linguistic evolution, or chance. As such, a variety of language grouping schemata have emerged at various levels. Perhaps most conservative is the notion of the LANGUAGE GENUS (Dryer 1989:267), which attempts to limit genealogical groupings to a maximum time-depth of 3500-4000 years (a criterion chosen to accord with the major established groupings within Indo-European, e.g. Romance, Germanic, Slavic, Celtic, etc). Higher-level groupings range from universally accepted (e.g. Indo-European) to highly speculative (e.g. Amerind), with a number of prominent controversial groupings under active debate (e.g.

Australian, Nilo-Saharan, et al). The highest-level grouping one can confidently establish has variously been called 'stock' (e.g. Cysouw 2005:555 and references therein) or 'phylum' (e.g. Perkins 1992), and historically the term 'family' has been used for groupings at a variety of levels (including this highest level). In what follows, we use the term FAMILY to refer to the highest-level grouping, and make no assertions as to the commensurability of these groupings with respect to equivalent time-depths or relative certainty of genealogical relatedness. We do so because in statistical modeling it is desirable to include the highest level groupings available to address the problem of (non-)independence of data points (the well-known 'nested data' problem).

A final methodological issue concerns interpretation of results. As data sets become increasingly large, the standard criterion of $p<0.05$ to determine statistical significance becomes easier and easier to attain, since in general a larger number of observations allows one to estimate variance with greater precision, hence potentially smaller standard errors and an increased likelihood of rejection of the null hypothesis. In such circumstances, an important part of statistical interpretation becomes the assessment of effect size: that is, if a statistically significant non-zero correlation exists, just how non-zero is it? For example, if each tenfold increase in speaker population yields an increase of 0.3 phonemes, is this finding interesting? Considering that the range of speaker populations spans about 9 orders of magnitude, the difference between the smallest and largest decades would be only 0.3×9, or a difference of less than 3 phonemes — well within the range of variability within any one order of magnitude. Such discussion of effect size is absent from the studies cited above; granted, many of these studies were intended as pilots or short reports, but that has not stopped them from influencing, or even becoming axioms, for other studies (e.g. Atkinson 2011).[8] Hay and Bauer's study in particular seems to be garnering quite a bit of attention outside of mainstream linguistics, so before presenting our own study we believe a close critique of their paper is in order.

As a short report, Hay and Bauer's approach to testing the hypothesized relationship between population size and phoneme inventory is admirable. However, in part because of the scope of the study, it suffers from a variety of methodological problems. The biggest problem is the non-independence of observations: data points within a given genealogical grouping are much more likely to have similar inventories than data points drawn from two different language families or

genera (this is the 'nested data' problem mentioned above). Hay and Bauer's main findings are based on Spearman rank correlations, a test that assumes that data are independent and thus does not take nesting into account. They attempt to account for the nesting problem by running two additional statistical tests. In the first test, each family with sufficient representation in their data was added to a multiple linear regression model as a categorical predictor.[9] In the second test, each language family was reduced to a single data point comprising the average speaker population and average phoneme inventory size of the languages from that family present in their data.

Hay and Bauer's results for the first test suggest that only Austronesian is a significant predictor of phoneme inventory size, while the variance seen in the other families is too great to conclude any significant difference among them (with population size remaining as a separate, significant predictor). This led them to provisionally conclude that the correlation they found was not simply an artifact of the particular languages and language families represented in their sample. However, multiple linear regression (like Spearman rank correlation mentioned in the previous paragraph) carries the assumption that each data point is independent, and failure to meet this assumption can lead to serious errors in estimations of statistical significance, even when the non-independence is partially accounted for by categorical predictors (Stevens 1986:75).

Hay and Bauer's multiple linear regression model has a further problem of radically unequal and unrepresentative group sizes: 44% of their languages fall into their 'other' category, while 23% are Indo-European (compare with Ethnologue, in which Indo-European contains only 6.4% of the world's languages). The overrepresentation of the Indo-European family, a well-described family with many languages boasting large populations and large phoneme inventories, may have biased their results toward finding a correlation where none exists. Hay and Bauer attempted to account for this possibility by employing a resampling procedure in which the same regression was run 200 times on different random samplings of their data. The results from the resampling procedure do seem to support the conclusion that population size is still a significant predictor of phoneme inventory size even after language family is taken into account. However, although resampling should effectively remove any bias due to INDIVIDUAL languages, the resampling alone is probably insufficient to overcome the strong bias introduced by the

overrepresentation of Indo-European in their data.

Moreover, radically unequal group sizes means that order of predictor entry matters in determining which predictors are significant (Stevens 1986:77–78). Although we are told that an automatic backwards step-down variable selection procedure was employed in their multiple linear regression model, we are not told what order of predictor entry yielded the reported results. For all we know, population may have been the first predictor added to the model, in which case language family is not being controlled for at all, but is merely added as a secondary predictor to help explain the residuals left after as much variability as possible has already been attributed to population size. All in all, the use of multiple linear regression with categorical predictors for the large families does not fully account for the influence of language family groupings, and thus does not conclusively demonstrate that their main findings about the influence of speaker population size are trustworthy.

In their second test regarding the influence of language family, Hay and Bauer reduce each family to a single data point comprising the average speaker population and average phoneme inventory size of the languages from that family present in their data. This reduced the 216 language-level data points to 46 family-level data points. Here the independence of observations is no longer a serious issue, since there is no established genealogical relationship between languages drawn from different (top-level) families, and thus no genealogical relationship between top-level aggregate data points.[10] However, sampling bias is still a serious concern. To take one example, their sample of Austronesian languages includes only languages from the Malayo-Polynesian branch, excluding all the Formosan languages (some of which have large phonemic inventories and very small populations — a combination that goes against the correlation that they seem to find). Moreover, although the choice of Spearman rank correlations is perhaps well-motivated based on the limited size of their data set, this approach is inherently more forgiving since it fits the data to any monotonic function instead of a straight line.[11]

To be clear: we believe Hay and Bauer did all they could to overcome the limitations of their data, and were meticulous in their analysis and their attempts to account for the problems they saw. They were also appropriately conservative in their interpretations. Nonetheless, their sample was simply too small and too biased to yield reliable results, limiting their statistical power and their choice of analytical tools. Given that their results are gaining acceptance among other

scholars (as mentioned above and in Note 8), we were inspired to conduct our own study on the influence of population size on phonological inventories.

4. CURRENT STUDY. Driven by Hay and Bauer's unexpected results, and in light of the methodological concerns expressed above, we set out to test the relationship between phonological inventories and speaker populations using our own data set. Our hypothesis was that once genealogical relatedness was properly accounted for, the correlations between speaker population and properties of the phoneme inventory would disappear. To test this, we used the PHOIBLE knowledge base[12] (Moran & Wright 2009): a data set of phonological inventories that subsumes and expands upon several extant sources, including UPSID-451 (Maddieson 1984, Maddieson & Precoda 1990), the Stanford Phonology Archive (SPA, Crothers et al. 1979), and *Systèmes alphabétiques des langues africaines* (Chanard 2006, Hartell 1993). Each language record includes population and genealogical information drawn from Ethnologue (Lewis 2009), geo-coordinates and genus-level classifications drawn from WALS online (Dryer & Haspelmath 2011), language family codes from Multitree 2009 and a variety of other types of data. Additionally, the segmental data are mapped to a set of phonological features so that the database may be queried based on features as well as segments. See Moran 2012 for full details; at the time of writing the database contained 1010 unique languages (not including some duplicate descriptions arising from the subsumption of other data sources). For the present study, pidgins, creoles, mixed languages, extinct or ancient languages, and languages lacking population data were excluded, leaving 969 languages to be analyzed (representing 321 genera and 100 families).

We first set out to reproduce the results found in Hay and Bauer using our data set and their statistical methods. Unsurprisingly, the results are quite similar: Spearman rank coefficients for our analyses range from 0.17 to 0.33, with statistically significant correlations between speaker population and inventories of sonorants, obstruents, all consonants, vowels, monophthongs, quality-only monophthong contrasts, and full phonemic inventories (see Figure 1). This suggests that our larger sample size has not radically changed the overall structure of the data; as mentioned above, however, we maintain that Spearman coefficients are not the most appropriate choice for assessing these data, since the data fails the criterion of independence of observations, and because the Spearman correlation is more forgiving than linear modeling.[13]

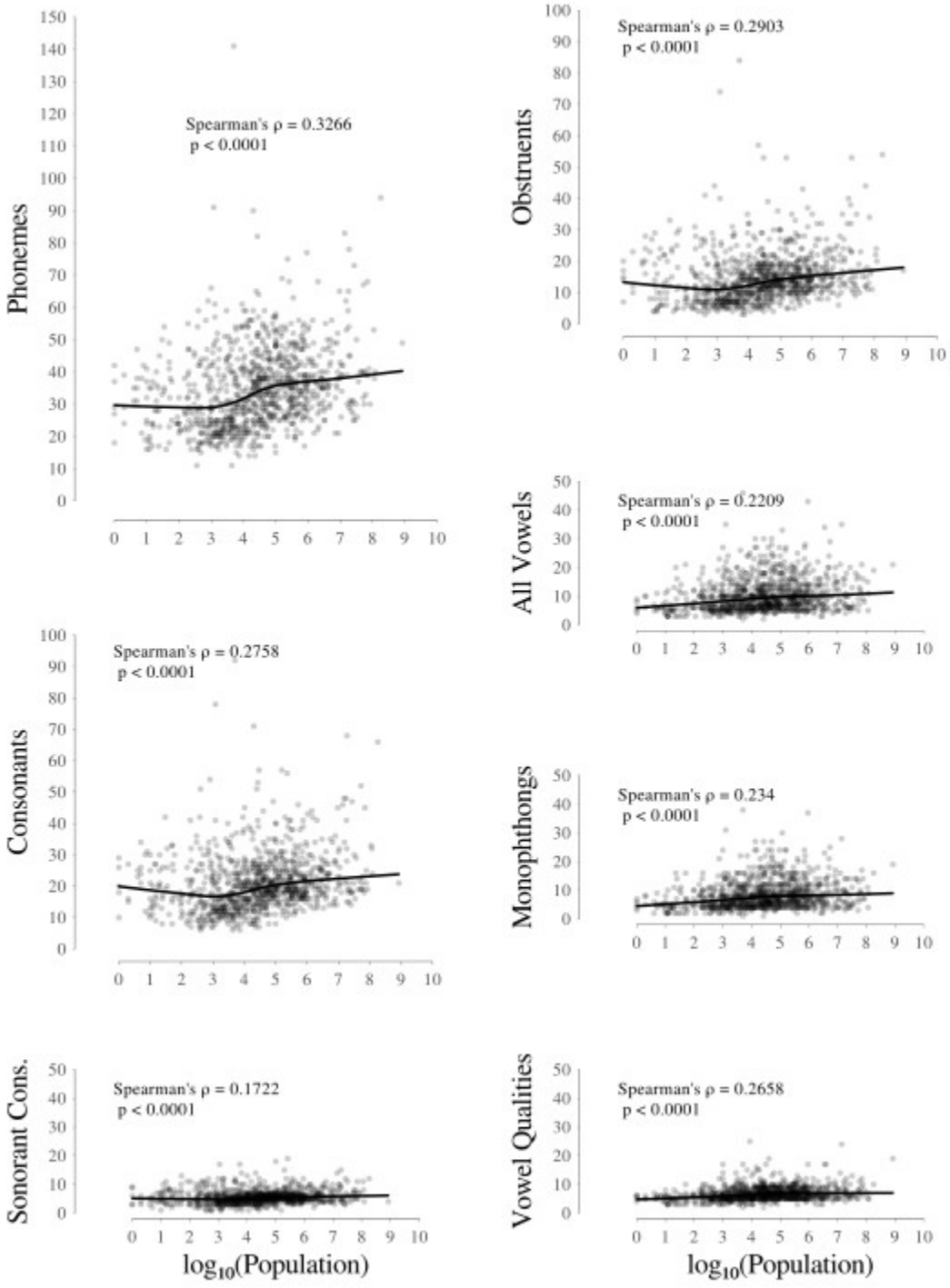

FIGURE 1: Scatterplots, LOWESS curves, and Spearman rank correlations for seven aspects of phoneme inventory size in the PHOIBLE data.

To address these issues, we re-examined the data using a hierarchical mixed effects model, which can handle nested data by allowing predictors at various levels. Mixed models also account for unequal group sizes by generating more precise estimates for groups with lots of data points and less precise estimates for sparsely populated groups. For maximum comparability to earlier studies, we created models in which log(population) is the independent variable (more commonly called a FIXED EFFECT PREDICTOR in the literature on mixed models). Language genus and language family were included as group-level predictors (a.k.a. RANDOM EFFECTS). Separate models were run with each of the following dependent variables: total phonemes, consonants, sonorant consonants, obstruent consonants, all vowels, monophthongs only, and quality-only monophthong contrasts (i.e. excluding contrasts of length, nasality, voice quality, or tongue root advancement/retraction). These were log-transformed to reduce non-normality of the residuals arising from the skewness of the outcome variables. In each case, we used a random-slopes-and-intercepts model (in which both the slope and intercept of the fitted line are allowed to vary across groups). The results are given in Table 1.

In interpreting each of these models, we are most interested in the coefficient for the log(population) fixed effect predictor (which represents the estimated slope of the line describing the relationship between phoneme inventory and population), and the amount by which this slope varies (represented by the standard error (SE) of the coefficient estimate, and by the variance (s^2) or standard deviation (s) of the random effects predictors GENUS and FAMILY). What is striking about many of these models is that the variability of the estimates for the influence of speaker population is in many cases smaller than the between-group variability for families. For example, the overall estimated slope for log(total phonemes)~log(speaker population) is 0.0093 with a standard error of 0.0041. But this slope is estimated to vary with a standard deviation of 0.0111 across families — a deviation nearly three times the coefficient estimate's standard error, and larger in magnitude than the coefficient estimate itself. This means that although the overall trend across languages is modestly positive, for many families we expect the relationship between phoneme counts and speaker population to be negative or zero. This can be seen in Figure 2, where the estimated slopes for the 12 largest families are plotted.

Outcome	Log-likelihood	Fixed effects			Random effects for genus (n=321)			Random effects for family (n=100)		
		Predictor	Coefficient (SE)	*t*	s^2	*s*	correlation	s^2	*s*	correlation
log(total phonemes)	697.6	intercept	1.4423 (0.0204)	70.8403	0.0000	0.0000	0.0000	0.0162	0.1272	-0.6540
		log(pop.)	0.0093 (0.0041)	2.2632	0.0001	0.0088		0.0001	0.0111	
log(consonants)	515.8	intercept	1.2173 (0.0289)	42.1727	0.0000	0.0000	0.0000	0.0385	0.1963	-0.7130
		log(pop.)	0.0091 (0.0058)	1.5848	0.0001	0.0093		0.0004	0.0212	
log(obstruent consonants)	398.4	intercept	1.0663 (0.0335)	31.7824	0.0008	0.0278	1.0000	0.0525	0.2291	-0.6740
		log(pop.)	0.0088 (0.0067)	1.3052	0.0001	0.0074		0.0006	0.0251	
log(sonorant consonants)	426.5	intercept	0.6194 (0.026)	23.7887	0.0226	0.1504	-1.0000	0.0169	0.1301	1.0000
		log(pop.)	0.0100 (0.0052)	1.9316	0.0006	0.0254		0.0000	0.0022	
log(vowels)	201.6	intercept	0.8483 (0.0259)	32.7772	0.0008	0.0287	1.0000	0.0135	0.1163	-1.0000
		log(pop.)	0.0157 (0.0053)	2.9515	0.0001	0.0099		0.0000	0.0059	
log(monophthongs)	150.8	intercept	0.7359 (0.0285)	25.8589	0.0012	0.0341	1.0000	0.0199	0.1412	-1.0000
		log(pop.)	0.0175 (0.0056)	3.1315	0.0001	0.0091		0.0001	0.0094	
log(quality-only vowel contrasts)	636.2	intercept	0.7238 (0.0174)	41.6384	0.0002	0.0158	1.0000	0.0068	0.0822	-0.6390
		log(pop.)	0.0095 (0.0037)	2.5583	0.0000	0.0063		0.0000	0.0060	

TABLE 1: Fixed effects estimates (left) and variance estimates (center, right) for models predicting several phoneme-inventory-related outcomes (*N*=969).

Looking within families, we see that the slope of the log(phonemes)~log(population) line is further predicted to vary across genera with a standard deviation of 0.0088 (again: more than twice the original estimate's standard deviation, and nearly as large as the estimate itself). This further underscores the fact that genealogical information is crucial to any prediction of phoneme inventory size, and predictions based only on speaker population ignore this important source of variability. To reiterate, the models suggest that the phoneme~population trend varies family to family and even genus to genus, fluctuating around zero, suggesting that any apparent relationship between population and phoneme inventory size does not generalize to language writ large.

Examining further the results in Table 1, we see that four of the seven models yielded estimates of the fixed effect that achieved statistical significance (these were the models predicting total phonemes, all vowels, monophthongs, and quality-only vowel contrasts). This can be concluded from the magnitude of the *t*-values for the fixed-effect estimates, since for large samples the *t*-distribution effectively converges on the normal distribution, and two-tailed $p<0.05$ significance for fixed effects can be informally assessed by looking for *t*-values whose absolute values are greater than 2 (Baayen et al. 2008).[14] Given the large sample size (*N*=969), estimating statistical significance from these *t*-values is justified.[15] However, when interpreting

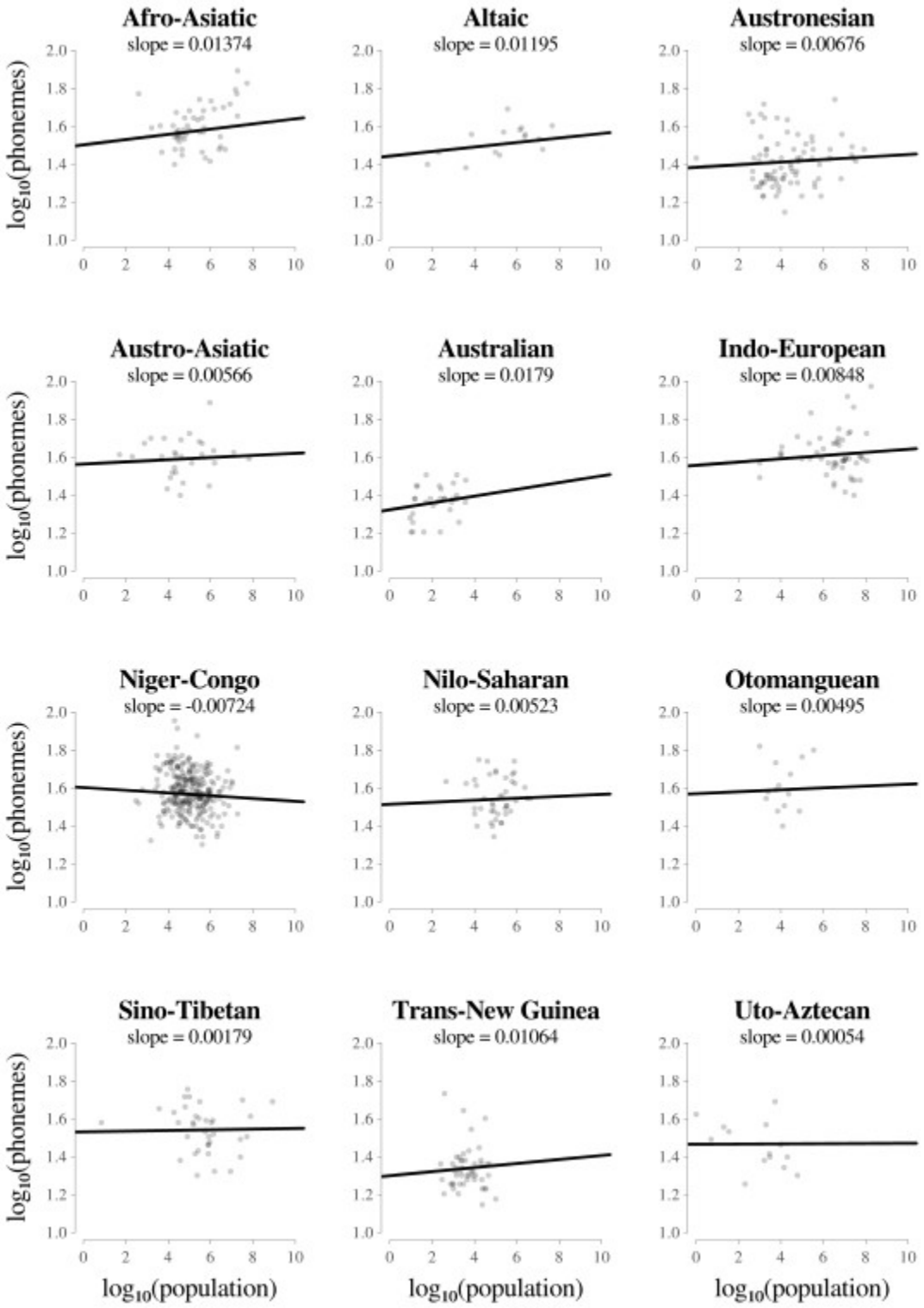

FIGURE 2: Family-level fitted lines from the mixed model predicting total phonemes, for the 12 largest families in PHOIBLE.

these models, it is important to remember that the traditional criterion of statistical significance ($p<0.05$) is not the only criterion for determining whether a model accurately reflects reality, despite the tendency for researchers across many fields to rely on *p*-values as the arbiter of whether a finding is interesting and therefore publishable.[16] In particular, some assessment of effect size is critical to an appropriate interpretation of model validity. With this in mind, it should be noted that in the four models in which log(population) achieved statistical significance as a fixed effect predictor, the estimated effect ranged from 0.009 to 0.017, meaning that across all languages, the models predict an increase of only 1.02-1.04 phonemes for each tenfold increase in speaker population. Given that speaker populations in the data range from 1 to 840 million (spanning nine orders of magnitude), the variability in phoneme inventory size that can be ascribed to differences in speaker populations is ultimately rather small (i.e. the model predicts a difference of only 9 phonemes between the languages with the smallest and largest speaker populations). Compare this to Figure 3, which shows the means and standard errors of the phoneme inventory sizes within each decade of speaker population, and it becomes clear that a change of 9 phonemes is well within the range of variability within any one order of magnitude (the standard errors of each decade range from 9 to 23). Recall also that the predicted phoneme~population relationship is tempered by the variability of slopes across families and genera (discussed above), so that in some families the relationship is slightly stronger, but in many families the relationship is flat or even decreasing.

5. CONCLUSION. In this paper, we highlight some methodological issues in previous studies of the relationship between properties of language and extralinguistic factors. Specifically, in reviewing the literature on the relationship between speaker population size and various measures of phonological complexity, we note that there are methodological shortcomings in both sampling techniques and in statistical modeling. To address these shortcomings, we conducted our own study using a larger and more representative sample of phoneme inventories drawn from the PHOIBLE knowledge base, and applied a more rigorous statistical analysis. Our results indicate that correlations between population size and phoneme inventory size are quite small when compared to differences among language family groups, and that the phoneme~population relationship varies between modestly positive, zero, and modestly negative

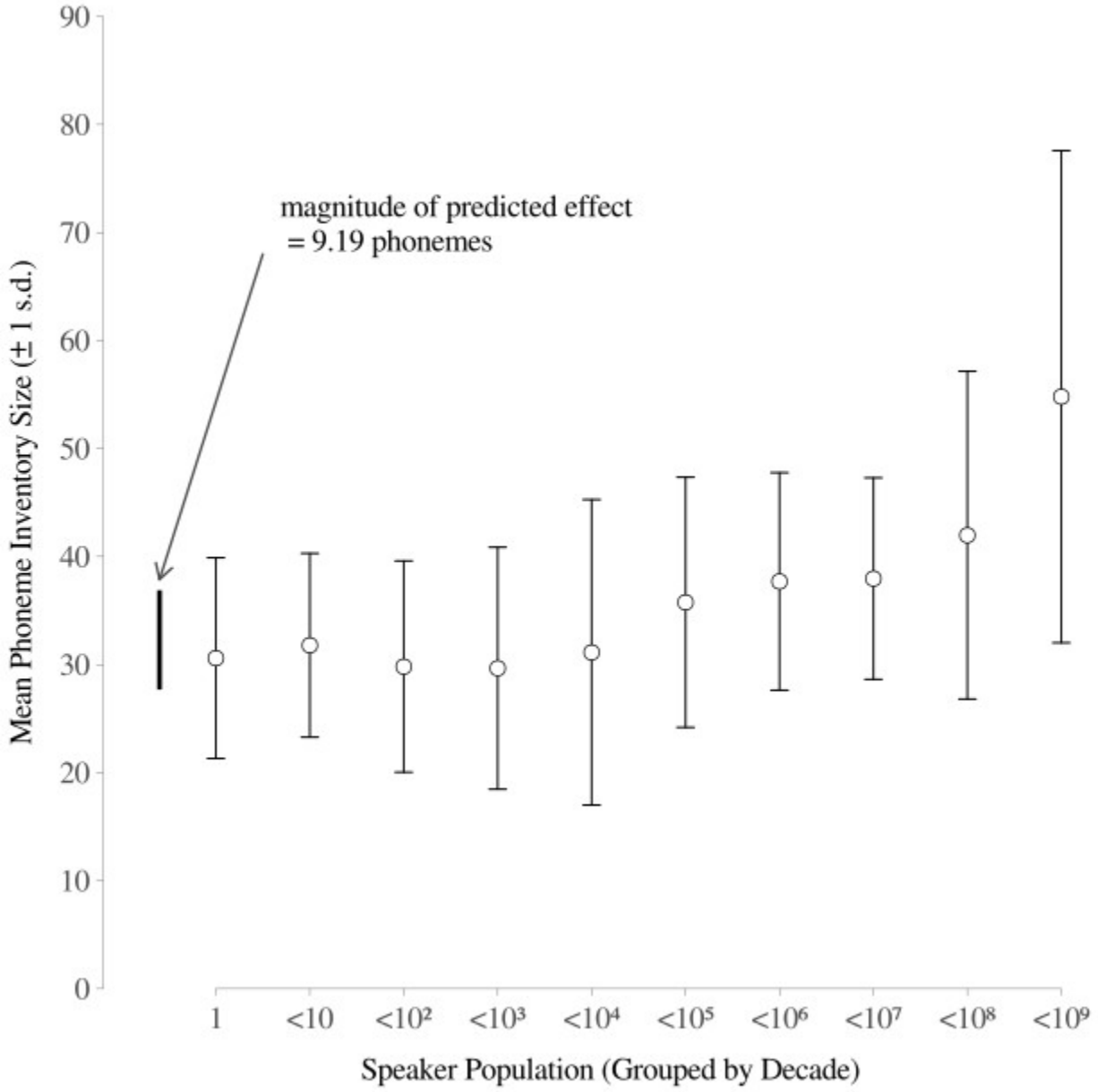

FIGURE 3: Means and standard errors for total phonemes in the PHOIBLE data, grouped by population (in order of magnitude bins).

depending on language family and genus. We conclude that the correlations seen between population and phoneme inventories are likely to be artifactual.

These results accord with intuitions we believe to be shared by many linguists: that similarity in population size is unlikely to be a source of phonemic similarity (either in total inventory size

or in segmental composition). Indeed, why should we expect population size to have any correlation with phonological structure at all, since populations can be decimated by war or disease in a matter of months, whereas phonological change seems to be much more gradual? Relatedly, why would we expect CURRENT population figures to relate to phonological structures that in most cases have been unchanged for hundreds of years? (cf. Donohue & Nichols 2011 on this point). Peter Trudgill makes a similar point quite succinctly in a recent paper: 'My suggestion [in earlier papers] was very much that the five social factors [community size, language contact, social network density, degree of shared information in the community, and social stability] could be expected, in combination, to have various kinds of influence on phoneme inventory size; it will never, I suggest, be sufficient to look at population figures alone' (Trudgill 2011:156, cf. Trudgill 2004a, Trudgill 2004b).

Of course, an intuition (even a broadly shared one) does not discount the possibility that some unexpected extralinguistic factor like population size MIGHT be influencing languages in subtle ways, and indeed, statistically significant correlations between phoneme inventories and populations were found in our data. However, we believe that the magnitude of the relationship is not substantial enough to be of interest when viewed in light of the variation within and across genealogical groupings. Thus we find no compelling reason to consider population size as a potential causal factor in the development of phonological systems, and thus no reason to postulate explanations, mechanisms, or reasons why such a pattern exists. We acknowledge that we are treading dangerously near to the scientific faux pas of interpreting a null result, and urge our readers to weigh our conclusions with appropriate caution and skepticism. We welcome other researchers to rexamine our data and statistical models (data and R code available on request). Of course, our data set is not immune to issues of bias any more than any other typological database is; however, its relatively large size does help mitigate bias stemming from over- or under-representation of specific language families. This can be seen in Figure 4, which plots the languages per family in PHOIBLE vs the language sample used by Hay and Bauer 2007, with the languages per family included in the Ethnologue (Lewis 2009) as reference. Notable in this figure is the unusual spike in Hay and Bauer's data set for the Indo-European family (ieur), and the relatively poor representation of Niger-Congo (ncon) at the far right of the graph (discussed above). Compare this with the line representing PHOIBLE, which more closely tracks the shape

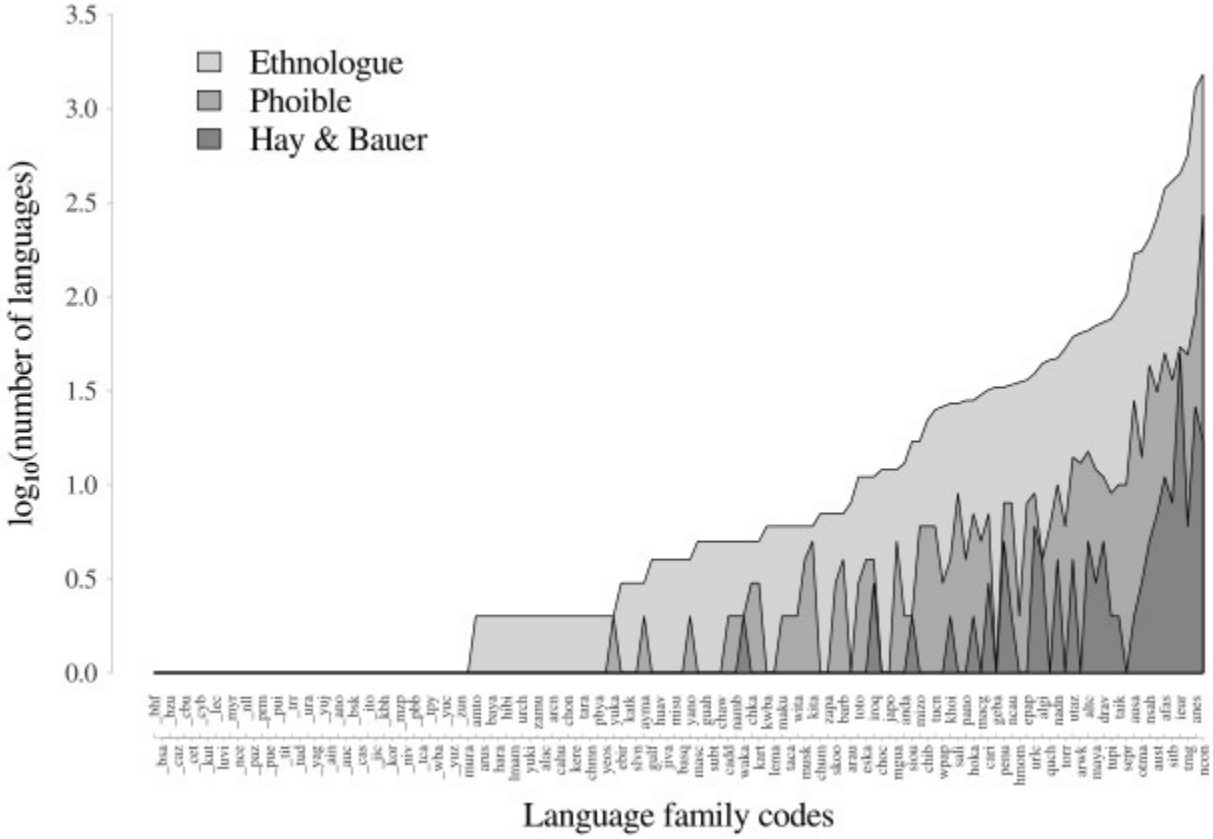

FIGURE 4: Languages per family in Ethnologue, PHOIBLE, and Hay & Bauer 2007. Family codes beginning with an underscore are isolates; all other family codes are from Multitree 2009.

of the Ethnologue line, particularly with respect to those two families.

Turning to larger questions, this study is situated within a growing branch of linguistic inquiry that examines how to quantify linguistic complexity and asks whether complexity is correlated with extralinguistic factors (be they environmental, biological, or societal). It is important to note that we are not asserting any particular interpretation of our data with regard to complexity; on the contrary, we tend to agree with the comments of Newmeyer 2011 that there seems to be no principled, theory-neutral way of measuring the grammatical complexity of language in a way that allows meaningful comparison across languages. In other words, agreeing on what counts as complexity is non-trivial, and even if we agree that, for example, more phonemes implies greater phonological complexity and the presence of noun classes implies

greater morphological complexity, there does not seem to be a principled way to compare complexity in one area of the grammar to complexity in another. In this light, our study should be seen not as a stake in the ground with regard to the determiners of phonological complexity, but rather as an example of how modern, data-rich approaches to linguistic questions can help steer us away from speculations that turn out to be rabbit-holes, while at the same time opening up new possibilities in the questions that we as linguists can address.

REFERENCES.

ATKINSON, QUENTIN. 2011. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332.346–349.

BAAYEN, R. HARALD; D. J. DAVIDSON; and DOUGLAS BATES. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59.390–412.

BAKKER, DIK. 2011. Language sampling. *The Oxford handbook of linguistic typology*, ed. by Jae Jung Song. Oxford: Oxford University Press.

BAKKER, PETER. 2004. Phoneme inventories, language contact, and grammatical complexity: A critique of Trudgill. *Linguistic Typology* 8.368–375.

BAUER, LAURIE. 2007. *The linguistics student's handbook*. Oxford: Oxford University Press.

CHANARD, CHRISTIAN. 2006. Systèmes alphabétiques des langues africaines. Online: http://sumale.vjf.cnrs.fr/phono/.

CROTHERS, JOHN; JAMES LORENTZ; DONALD SHERMAN; and MARILYN VIHMAN. 1979. *Handbook of phonological data from a sample of the world's languages: A report of the Stanford Phonology Archive*. Palo Alto, CA: Department of Linguistics, Stanford University.

CYSOUW, MICHAEL. 2005. Quantitative methods in typology. *Quantitative linguistics: An international handbook*, ed. by Reinhard Köhler, Gabriel Altmann, and Rajmond Genrikhovich Piotrovskii, 554–578. Handbooks of Linguistics and Communication Science 27. Berlin: W. de Gruyter.

Cysouw, Michael; Dan Dediu; and Steven Moran. 2012. Comment on "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa." Science 335.657–657.

DONOHUE, MARK, and JOHANNA NICHOLS. 2011. Does phoneme inventory size correlate with population size? *Linguistic Typology* 15.161–170.

DRYER, MATTHEW. 1989. Large linguistic areas and language sampling. *Studies in Language* 13.257–292.

DRYER, MATTHEW, and MARTIN HASPELMATH (eds.) 2011. *The world atlas of language structures online*. Munich: Max Planck Digital Library. Online: http://wals.info/.

FOUGHT, JOHN; ROBERT MUNROE; CARMEN FOUGHT; and ERIN GOOD. 2004. Sonority and climate in a

world sample of languages: Findings and prospects. *Cross-Cultural Research* 38.27–51.

GELMAN, ANDREW, and JENNIFER HILL. 2007. *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

GORDON, RAYMOND G. (ed.) 2005. *Ethnologue: Languages of the world*. 15th ed. Dallas, TX: SIL International. Online: http://www.ethnologue.com/.

HAMMARSTRÖM, HARALD. 2007. *Handbook of descriptive language knowledge: A full-scale reference guide for typologists*. LINCOM Handbooks in Linguistics 22. Munich: Lincom Europa.

HARTELL, RHONDA. 1993. *Alphabets des langues africaines*. Dakar, SN: UNESCO, Bureau Régional de Dakar.

HASPELMATH, MARTIN; MATTHEW DRYER; DAVID GIL; and BERNARD COMRIE (eds.) 2005. *The world atlas of language structures*. Oxford: Oxford University Press.

HAUDRICOURT, ANDRÉ. 1961. Richesse en phonèmes et richesse en locuteurs. *L'Homme* 1.5–10.

HAY, JENNIFER, and LAURIE BAUER. 2007. Phoneme inventory size and population size. *Language* 83.388–400.

LEWIS, M. PAUL (ed.) 2009. *Ethnologue: Languages of the world*. 16th ed. Dallas, TX: SIL International. Online: http://www.ethnologue.com/.

LUPYAN, GARY, and RICK DALE. 2010. Language structure is partly determined by social structure. *PLoS ONE* 5.e8559.

MADDIESON, IAN. 1984. *Patterns of sounds*. Cambridge Studies in Speech Science and Communication. New York: Cambridge University Press.

MADDIESON, IAN. 2011a. Consonant inventories. In Dryer & Haspelmath.

MADDIESON, IAN. 2011b. Vowel quality inventories. In Dryer & Haspelmath.

MADDIESON, IAN. 2011c. Tone. In Dryer & Haspelmath.

MADDIESON, IAN, and KRISTIN PRECODA. 1990. Updating UPSID. *UCLA Working Papers in Phonetics* 74.104–111.

MORAN, STEVEN. 2012. Phonetics information base and lexicon. Seattle, WA: University of Washington doctoral dissertation.

MORAN, STEVEN, and RICHARD WRIGHT. 2009. *Phonetics information base and lexicon (PHOIBLE)*. Seattle, WA. Online: http://phoible.org/.

MUNROE, ROBERT; JOHN FOUGHT; and RONALD MACAULAY. 2009. Warm climates and sonority classes: Not simply more vowels and fewer consonants. *Cross-Cultural Research* 43.123–133.

NETTLE, DANIEL. 1996. Language diversity in west Africa: An ecological approach. *Journal of Anthropological Archaeology* 15.403–438.

NETTLE, DANIEL. 1999a. *Linguistic diversity*. Oxford: Oxford University Press.

NETTLE, DANIEL. 1999b. Is the rate of linguistic change constant? *Lingua* 108.119–136.

NETTLE, DANIEL. 1999c. Using social impact theory to simulate language change. *Lingua* 108.95–117.

NEWMEYER, F. J. 2011. Can one language be “more complex” than another? Lecture presented at the UW colloquium series, Seattle, WA.

NORDHOFF, SEBASTIAN; HARALD HAMMARSTRÖM; and MARTIN HASPELMATH. 2012. *Glottolog/Langdoc*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Online: http://glottolog.livingsources.org.

NOWAK, ANDRZEJ; JACEK SZAMREJ; and BIBB LATANÉ. 1990. From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review* 97.362–376.

PERICLIEV, VLADIMIR. 2004. There is no correlation between the size of a community speaking a language and the size of the phonological inventory of that language. *Linguistic Typology* 8.376–383.

PERKINS, REVERE. 1992. *Deixis, grammar, and culture*. Typological Studies in Language 24. Amsterdam: John Benjamins.

RIJKHOFF, JAN, and DIK BAKKER. 1998. Language sampling. *Linguistic Typology* 2.263–314.

SAPIR, EDWARD. 1912. Language and environment. *American Anthropologist* 14.226–242.

SNIJDERS, TOM, and ROEL BOSKER. 1999. *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

SPROAT, RICHARD. 2011. Phonemic diversity and the out-of-Africa theory. *Linguistic Typology* 15.199–206.

STEVENS, JAMES. 1986. *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

TRUDGILL, PETER. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society* 3.215–246.

TRUDGILL, PETER. 1989. Interlanguage, interdialect and typological change. *Variation in second language acquisition: Psycholinguistic issues*, ed. by Susan M. Gass, Carolyn Madden, Dennis R. Preston, and Larry Selinker, 243–253. Clevedon, UK: Multilingual Matters.

TRUDGILL, PETER. 1996. Dialect typology: Isolation, social network and phonological structure. *Towards a social science of language: Papers in honor of William Labov*, ed. by Gregory R. Guy, Crawford Feagin, Deborah Schiffrin, and John Baugh, 1:3–21. Amsterdam: John Benjamins.

TRUDGILL, PETER. 1997. Typology and sociolinguistics: Linguistic structure, social structure and explanatory comparative dialectology. *Folia Linguistica* 31.349–360.

TRUDGILL, PETER. 2002. Linguistic and social typology. *The handbook of language variation and change*, 707–728. Oxford: Blackwell.

TRUDGILL, PETER. 2004a. Linguistic and social typology: The Austronesian migrations and phoneme inventories. *Linguistic Typology* 8.305–320.

TRUDGILL, PETER. 2004b. On the complexity of simplification. *Linguistic Typology* 8.384–388.

TRUDGILL, PETER. 2011. Social structure and phoneme inventories. *Linguistic Typology* 15.155–160.

VAN DER LAAN, MARK, and SHERRI ROSE. 2010. Statistics ready for a revolution. *Amstat News*. Online: http://magazine.amstat.org/blog/2010/09/01/statrevolution/.

WICHMANN, SØREN, and ERIC HOLMAN. 2009. Population size and rates of language change. *Human Biology* 81.259–274.

WICHMANN, SØREN; TARAKA RAMA; and ERIC HOLMAN. 2011. Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology* 15.177–197.

WICHMANN, SØREN; DIETRICH STAUFFER; CHRISTIAN SCHULZE; and ERIC HOLMAN. 2008. Do language change rates depend on population size? *Advances in Complex Systems* 11.357.

2009. *Multitree: A digital library of language relationships*. Ypsilanti, MI: Institute for Language Information and Technology (LINGUIST List), Eastern Michigan University. Online: http://multitree.org/.

1 As one example, the Dogon in Mali lexically distinguish about 30 local grasshopper species; *Kraussaria angulifera* is an especially tasty variety when salted and roasted (Jeffrey Heath, p.c.).

2 This section highlights a few currents in the ongoing discourse that exemplify the kinds of claims made about the relationship between socio-environmental factors and linguistic structures. For a more thorough review of this literature, see Moran 2012:chap. 7.

3 Indeed, Trudgill himself notes that some small, isolated communities like the !Xũ speakers have extremely large phonemic inventories, leading him to a more convoluted set of conclusions about the relationship between phoneme inventory size, population size, and language contact (see Trudgill 2004a:317, for the full, revised typology of cases).

4 As Bauer states: '...the sample here is not a random one (and this may have implications for the uses to which the list can put) but a sample of opportunity, which means that well-described languages and major languages stand a far better chance of figuring here than poorly-described languages and minor languages' (2007:221).

5 For example, the language chosen to represent the Bantoid languages in UPSID is Swahili, which is rather atypical as far as Bantoid languages go in that it lacks lexically contrastive tone (cf. the 17 Bantoid languages in Hartell 1993 and Chanard 2006, of which 15 are described as tonal).

6 As Maddieson notes, 'Availability and quality of phonological descriptions are factors in determining which language to include from within a group' (1984:5–6).

7 Bakker 2011 points out that the bibliographic bias can also be inflicted by the linguistic theory used in language documentation. That is, creating a sample not only requires language documentation, but comparable analyses. Note that Bakker's estimates might be a bit too high. Cf. Hammarström 2007 and MPI-EVA's Glottolog (Nordhoff et al. 2012).

8 For example, Atkinson 2011:346 begins, 'The number of phonemes — perceptually distinct units of sound that differentiate words — in a language is positively correlated with the size of its speaker population in such a way that small populations have fewer phonemes.' Note that Atkinson also replicates Hay and Bauer's 2007 finding using the WALS data set and gets similar results, but Atkinson combines data from Maddieson's 2011a, 2011b, 2011c WALS chapters on consonant inventories, vowel quality inventories, and tone. This methodology is

flawed because inventory size is discretized in WALS, and because Atkinson erroneously weights consonant, vowel, and tone inventories equally in his model of phonemic diversity (see Sproat 2011 and Cysouw et al. 2012 for more thorough critiques).

9 Hay and Bauer chose seven languages as the minimum for this cutoff — a number chosen solely because it preserved the needed degrees of freedom in their model — which led to the inclusion of five families as predictors: Altaic, Austronesian, Indo-European, Niger-Congo, and Penutian. Note however, that they relied on the original grammars for genealogical classifications, and if the list of languages in their data set (from Bauer 2007) is reclassified using genealogical data from Ethnologue, the families that meet their seven-language criterion change to include Afro-Asiatic, Australian, and Sino-Tibetan, with Altaic and Penutian both reduced to only five representative languages. Note that there is also a risk of genealogical bias in their data, since genealogically stratified samples can change drastically depending on the genealogical classification used (cf. Rijkhoff & Bakker 1998).

10 It is possible that language contact could compromise the independence of data points, and as such it would be desirous to impose geographical as well as genealogical controls or constraints on the language sample analyzed. However, there is no way that we know of to systematically quantify the extent of contact between two (or more) languages, especially given the extreme variability in the availability of historical records and the asymmetries of language contact effects on the participant languages. As such, any non-independence due to language contact must be accepted as noise in the data, and assumed (or at least hoped) to be negligible when compared to other predictors or covariates.

11 Another critique, though somewhat tangential, regards the reproducibility of their data: languages are listed in Bauer 2007 by name, not by ISO 639-3 code, and many of the languages in their data set are in fact macrolanguages or sub-genera (e.g. Berber, Malagasy, Malay). This, when combined with the reliance on primary source classifications for language genealogies (see Note 9) makes confirming their population figures and replicating their results rather difficult.

12 http://phoible.org/

13 It is also more difficult to estimate effect size using a Spearman analysis, since it models only the ordering of the datapoints, ignoring the relative distances between them (rather than

fitting the data set as a whole with a single easily describable function).

14 The extremely high *t*-values for fixed-effect intercepts are expected (but often uninformative) in analyses like this.

15 More formal assessments of significance could be made through sampling from the posterior distribution of a Markov-chain Monte Carlo simulation, but for these models that is unfeasible due to the slope-intercept correlations within the random effects. Correlation between intercepts and slopes of the same grouping factor is common, and its main risk is simply difficulty in interpreting the intercept (Gelman & Hill 2007:288). This can usually be overcome by normalizing the predictor variable so that its distribution is centered around zero, but in random-slope models the choice to do this is not trivial (see Snijders & Bosker 1999:80–81, for discussion). Regardless, since we are primarily interested in the relationship between phonemes and population (rather than what a 'typical' language looks like), we are much more interested in slopes rather than intercepts anyway, so difficulty in interpreting the intercepts is not a problem for our analysis.

16 This problem is nicely summed up by van der Laan and Rose (2010): 'We know that for large enough sample sizes, every study — including ones in which the null hypothesis of no effect is true — will declare a statistically significant effect.'