

Title: Phoneme Inventory Size and Population Size

Authors:

Jennifer Hay (corresponding author)
Department of Linguistics
University of Canterbury
Private Bag 4800
Christchurch
New Zealand
Phone: +64 3 364-2242
Email: jen.hay@canterbury.ac.nz

Laurie Bauer
School of Linguistics and Applied Language Studies
Victoria University of Wellington
P.O. Box 600
Wellington
New Zealand
Phone: +64 4 463-5619
Email: laurie.bauer@vuw.ac.nz

Acknowledgements:

We are grateful to Harald Baayen, Ann Bradlow, Brian Joseph, Christian Langstrof, Peter Trudgill, Paul Warren and the Language reviewers for their comments and suggestions, and to Vladimir Pericliev for generously sharing his data with us.

Abstract:

This short report investigates the relationship between population size and phoneme inventory size, and finds a surprisingly robust correlation between the two. The more speakers a language has, the bigger its phoneme inventory is likely to be. We show that this holds for both vowel inventories and consonant inventories. It is not an artefact of language family.

July 2006. Slightly Revised Version to appear in Language

Phoneme Inventory Size and Population Size

Introduction

To our knowledge, no-one has ever reported a statistical correlation between the number of speakers of a language, and how many phonemes that language has. This, of course, is not surprising: why would anyone look for such an association in the first place? It is certainly not an association one would necessarily expect. However in the process of proof-reading a manuscript that contained information about series of languages, including their populations and vowel inventories, it struck us that there appeared to be some link. We couldn't resist checking this apparent link more systematically. In this short report we show that there is, indeed, an association between phoneme inventory and population size. We do not have well-developed theoretical arguments to offer about why this should be. However the correlation seemed intriguing enough that it was worth simply publishing the result, and leaving it up to readers to draw their own conclusions.

Materials

Bauer (in prep) is a handbook designed to provide a range of useful information for linguistics students. One part of the handbook is a list of some 250 languages with summary information about each, including its language family, where it is spoken, how many speakers it has and typological features such as the relative order of subject verb and object, or of noun and adjective, and so on. One piece of information that was collected for these purposes was the number of vowels each language is said to have.

When we noticed an apparent association between population size and vowel inventory size, we decided to supplement this information with information about the consonant inventory as well. Since not all of the original materials were still to hand, it proved difficult to collect all the relevant data here, and we were not able to collect information on consonants for as many languages as we had for vowels. We also decided to exclude from our analysis any language which did not have any living speakers. We therefore ended up with full population, vowel and consonant information for a set of 216 languages.

While this is not a random sample, we think it is a reasonably representative one. An ideal approach to such a sample might be to do random selection from some list such as the *Ethnologue* (Grimes 1988); however this would prove impractical in terms of gathering the required additional information, and would not have been appropriate for the original purpose of the selection in Bauer (in prep). For the purposes of the book, languages were chosen to provide a geographical and genetic spread of languages, while including languages which students might be expected to or want to know something about.

Thus ‘big’ languages like English, Hindi and Mandarin were chosen, and ‘small’ but linguistically well known languages like Basque, Diyari and Hixkaryana were also selected. Languages which were not well-described in works easily available in accessible libraries stood very little chance of being selected. Thus the set of languages selected is somewhat biased towards ‘big’ languages and towards Indo-European and Pacific languages (because we are in New Zealand), but covers a range of languages from around the world. While the selection was not random, we cannot identify anything in the selection process that would have introduced an artefactual correlation between population size and phoneme inventory.

The phoneme inventory counts are taken directly from other linguists’ analyses. No additional analysis of languages’ phoneme inventories has been conducted by us. Thus, when we described the vowel information as representing how many vowels a language is ‘said to have’, the wording was deliberately careful. However much we may believe in the non-uniqueness of phonemic solutions (Chao 1934), it comes as a surprise to see just how different two descriptions of the same language can be. For this reason, we have taken care to consider various subsets of the phoneme inventories separately, so that we could apply appropriate caution to any effect carried by parts of the inventory which are particularly prone to variation across analysts.

We distinguished between ‘basic monophthongs’ which differ in quality only, and ‘extra monophthongs’ which consisted of non-quality distinctions, such as length and nasalisation. The basic monophthong counts are likely to be much more consistent across analysts. As Maddieson (2005:14) points out, the languages for which length and nasalized forms are listed as separate phonemes cannot necessarily be relied upon, as when analysts are considering whether such a distinction is phonemic “the considerations which would lead to making one choice or the other are often finely

balanced and lead different scholars to different conclusions”. Maddieson excludes such forms from his analysis. We list them separately, and regard the counts with appropriate caution. Diphthongs were also listed separately, and here the analysis is even more open to interpretation, since diphthongs may be analysed as independent phonemes, as sequences of vowel and glide or as sequences of non-identical vowels.

In terms of consonants, a distinction was made between obstruents (including plosives, affricates, implosives, ejectives, clicks, fricatives) and sonorants (including nasals, liquids and glides), but again some caution is required. The numbers here were also often different from one analysis to another.

Analysis

Before conducting any statistical analysis, we inspected the consonant and monophthong inventory sizes for outliers, and removed two extreme outlier languages (i.e. showing values more than 4 standard deviations above the mean). These were !Xu (for total consonants) and Acooli (for total monophthongs). We also took the log of the population size, in order to minimize the effect of outliers.

The left panel of Figure 1 shows the positive correlation between the log population of speakers of a language and their basic monophthong inventory, i.e. including quality distinctions only. The right panel repeats the correlation, this time also including additional phonemic vowel differences such as length and nasalization. The correlation involving ‘basic monophthongs’ is much tighter; if there turns out to be some causal relationship between population size and number of vowels, then this tighter correlation for ‘basic monophthongs’ might be attributed to the greater consistency here across analysts.

Any correlation regarding diphthong inventory may be very unreliable, due to the inherent role of the interpretation of the analyst. However we note, in passing, that population size is also well correlated with the number of diphthongs listed by analysts ($\rho = .28$, $p < .0001$).

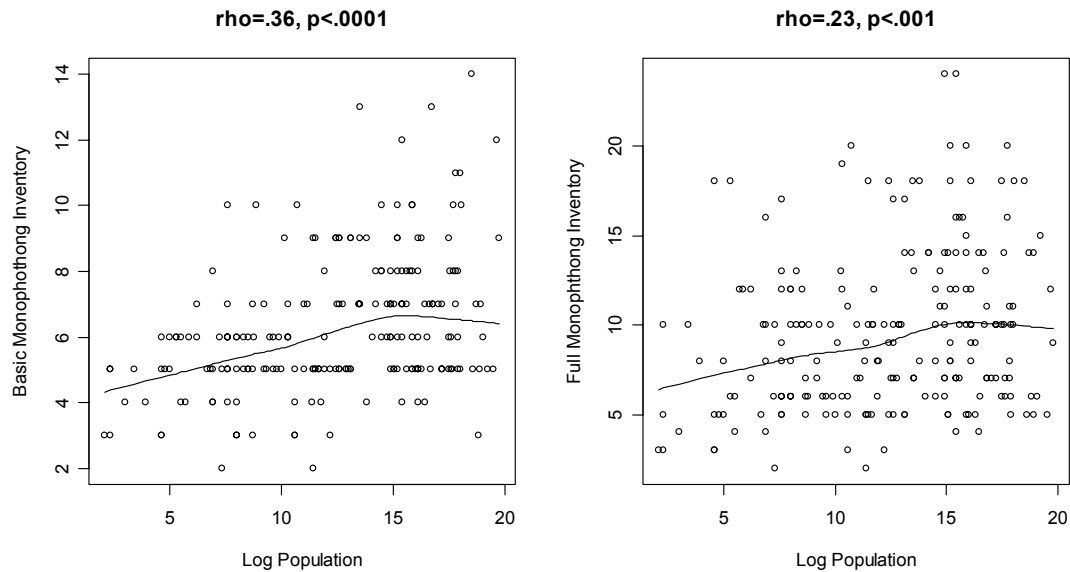


Figure 1: The association between population size and vowel inventory. Each point represents a language. The left panel shows the inventory of basic monophthongs, distinguished by quality. The right panel shows the inventory including other distinctions analyzed as phonemic, such as nasalisation and length. The line shows a non-parametric scatterplot smoother fit through the points (Cleveland 1979).

Figure 2 shows the correlation of population size with the size of a language's obstruent inventory, sonorant inventory, overall consonant inventory, and overall phoneme inventory. All of these return significant correlations. As described above, we have removed two languages from the sample because they fall more than 4 standard deviations from the mean. This was because we were worried that they would exert undue influence on the statistics. A reviewer was worried about the effects of removing these languages, and so we have also checked all of the above correlations with the two languages include. All correlations remain significant.

It is important to note that what we are seeing here is an overall statistical tendency. The graphs in Figures 1 and 2 all show a reasonable amount of scatter, reflecting the fact that there are individual languages which go against this general trend. Faroese, for example, has just 45,000 speakers, but 21 obstruents, 18 sonorants, 6 monophthongal vowel qualities plus a length distinction, and 8 diphthongs, many of which also show a long/short distinction (see Thráinsson et al. 2004).

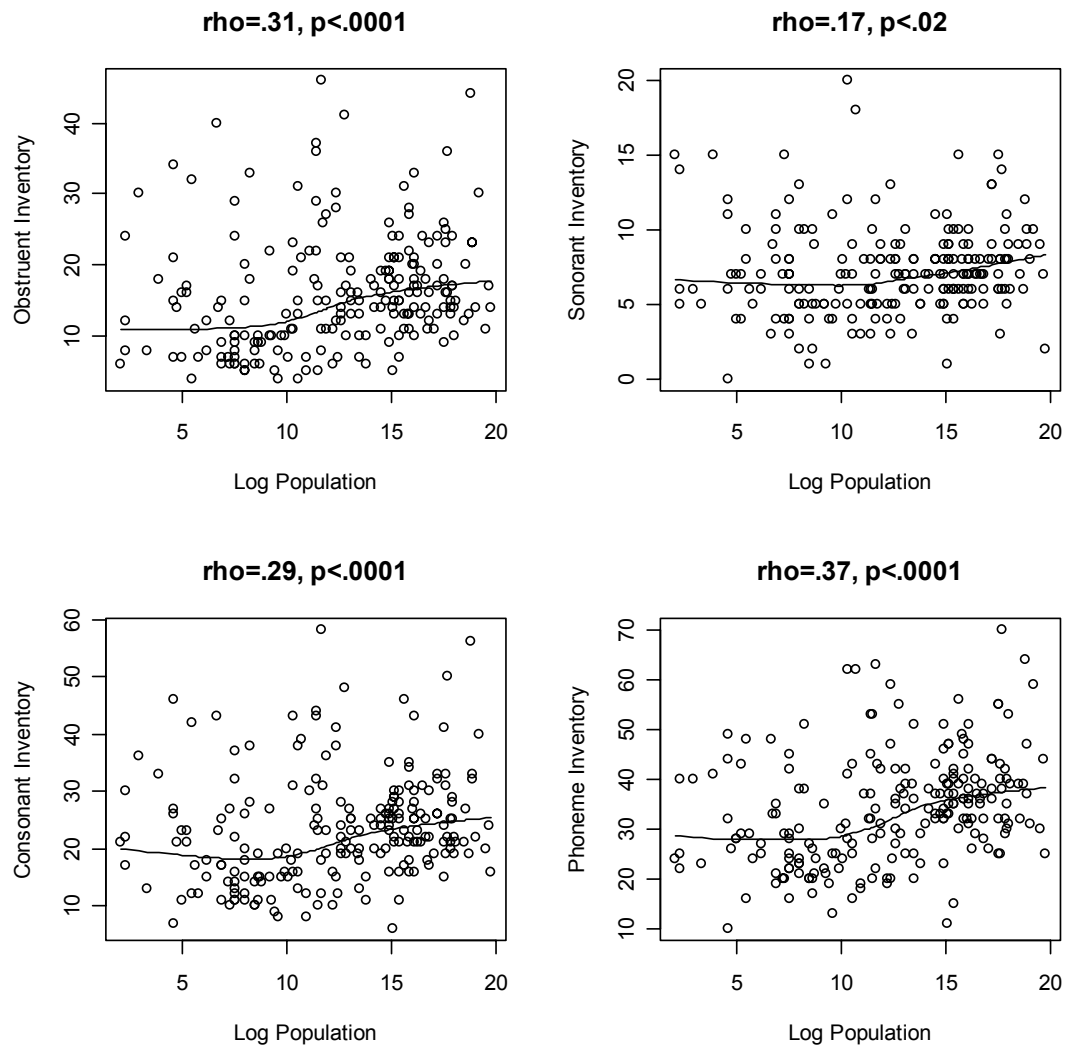


Figure 2: The association between population size and inventories of obstruents (top left); sonorants (top right); all consonants (bottom left) and all phonemes (bottom right). Each point represents a language. The line shows a non-parametric scatterplot smoother fit through the points (Cleveland 1979).

That vowel inventory and consonant inventory are both correlated with population size is quite remarkable. This is especially so because consonant inventory and vowel inventory do not correlate with one another at all in this data-set ($\rho = -.01, p = .86$).

Maddieson (2005) also reports that there is no correlation between vowel and consonant inventory size in his sample of 559 languages. Despite the fact that there is no link between vowel inventory size and consonant inventory size, both are significantly correlated with the size of the population of speakers.

We were suspicious that some of this trend might be carried by an association between language families and population size. Australian languages, for example, tend to have small vowel inventories, and small populations. Indo-European languages tend to have larger vowel inventories, and large populations. The fact that vowel and consonant inventory are not correlated at all, but that both correlate with population suggests that language family is not the whole story here. That is, if the relationship between population size and both the consonant and vowel inventories were solely an artefact of language family, we might expect the consonant and vowel inventories to correlate with one another, and they do not. Nonetheless, it seemed important to assess the role that language family may be playing.

Each language in our database has been coded for the language family it belonged to. Given how controversial some language families are, it might seem that our analysis here could be skewed by the particular choice of labels for language families. The classification of the various languages was based on the data available in the sources for Bauer (in prep). Since some of these sources were much older than others, there was some variation in the names of language families, and in the attribution of individual languages to particular families. However, since we were not concerned with the minutiae of the classifications, but with the top-level classifications, these variations were relatively easily eliminated on inspection and we believe the classification we used to be fairly robust.

In order to assess the significance of population while factoring in language family, we fit an ordinary least squares linear regression model, predicting the total phoneme inventory. We included language family as an independent variable, identifying all families for which we had seven or more languages represented, and classifying all other families as “other”. The threshold of seven languages was chosen so as to represent a reasonable number of language families in the model (seven language families met this threshold), while still retaining an appropriate number of degrees of freedom in our model. The model statistics are shown in tables 1 and 2.

Table 1: Analysis of Variance for ordinary least squares model predicting phoneme inventory size.

Factor	d.f.	Partial SS	MS	F	P
Family group	6	3534.326	589.0543	6.4	<.0001
Log population	1	699.9648	699.9648	7.61	0.0063
REGRESSION	7	6415.775	916.5392	9.96	<.0001
ERROR	206	18952.06	92.00028		

Table 2: Coefficients for ordinary least squares model predicting phoneme inventory size

	Value	Std. Error	t	Pr(> t)
Intercept	32.0872	4.1607	7.712	0.0000
family=Altaic	-4.5173	4.6882	-0.9635	0.3364
family=Austronesian	-13.4739	3.5936	-3.7494	0.0002
family=Indo-European	1.689	3.2953	0.5125	0.6088
family=Niger-Congo	-0.8249	3.7932	-0.2175	0.8281
family=other	-5.0686	3.3043	-1.534	0.1266
family=Penutian	-5.2322	4.6232	-1.1317	0.2591
Log Population	0.4718	0.1854	2.544	0.0117

Residual standard error: 9.512 on 206 degrees of freedom
Adjusted R-Squared: 0.24

Figure 3 plots the predictions of the model (overall model $r^2=.24$). Language family does indeed have a significant influence, with Indo-European languages having the largest phoneme inventories, and Austronesian languages having smaller phoneme inventories (top panel). However in addition to language family, the log population of speakers is a separate, significant predictor. This is shown in the bottom panel of figure 3, which plots the predicted effect of population size, while holding language family constant.

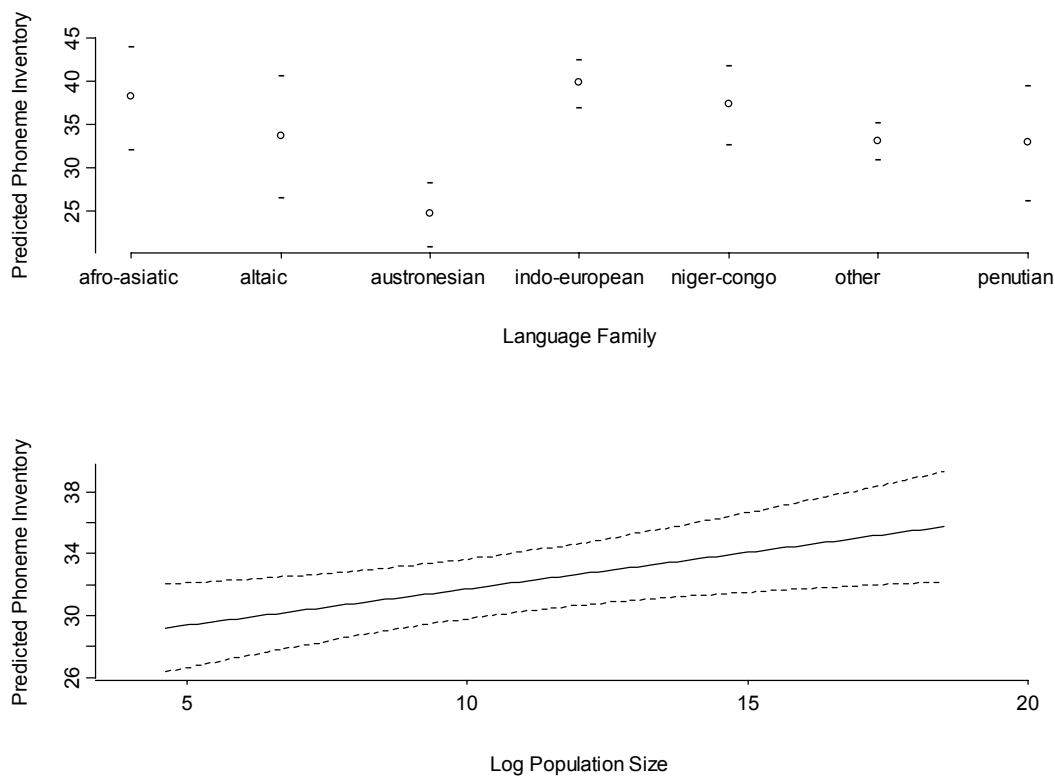


Figure 3: Predictions of model predicting phoneme inventory size. The effects of Language Family (top panel) and Population Size (bottom panel). Dashed lines show 95% confidence intervals.

Our sample of languages is, of course, not random. It is a function of the particular languages for which the relevant information happened to be readily available to us. It would be very difficult to create a fully random sample. This would involve taking all known languages, randomly sampling them, and then setting out to find population size and phoneme inventory for each selected language. In some cases this would require extensive research, including fieldwork.

In order to assess the degree to which the specific selection of languages under investigation is responsible for the significance of the regression model, we conducted bootstrap validation of the model, using the ‘validate’ function from Harrell’s Design library in R (see description in Harrell 2001). The validation technique involves refitting the model over random subsamples of our data. About 63% of the languages

are included in each random sample, together with replacement. That is, some languages are included in each sample more than once, so that the random sample remains the same size as the original sample. An automatic backwards step-down variable selection procedure is employed. This is repeated 200 times. In all 200 iterations, language family was retained as a significant predictor. Log population was retained in 180. This provides good evidence that the significance of the model is not due to the inclusion of any specific languages in our sample. The average R-squared value across these 200 models is .22.

This seems to provide some reassuring evidence that the observed correlation with population size is not due to the specific collection of languages included in our overall sample, nor is it due to undue influence of language family. However there is still a worrying element regarding the role of language family in the model reported above. 44% of our languages fall into the ‘other’ group. That is, 44% of the languages belong to a language family that is represented by fewer than 7 languages in our sample. The model presented in Tables 1 and 2 certainly demonstrates that the observed difference is not due to differences between large family groups in the sample. For example, there is a difference between Indo-European (many speakers, many phonemes) and Austronesian languages (few speakers, few phonemes), but this is not carrying the effect. But the potential effect of other, smaller, family groups may still be influencing the correlation.

As a final check, then, we decided to check how this correlation holds across family groups. That is, we removed the potential for undue influence by specific language families, by reducing each language family to one data point¹. We did this by calculating the mean population size for each language family, and the mean number of phonemes. There are 42 language families in the sample.

Table 3 shows the Spearman’s correlation between mean population size and mean inventory size, for different parts of the phoneme inventory. There is a significant

¹ We owe thanks to Harald Baayen for suggesting this approach.

effect in most parts of the phoneme inventory. One exception is the sonorants – this is the subset of the consonant inventory which showed the weakest effect in Figure 2. The other is the full monophthong count including the ‘extra’ monophthongs. This parallels the weaker effect within this subset of phonemes we have already observed in Figure 1, and can be ascribed to the high degree of variability across analysts in terms of what might count as an ‘extra’ monophthong.

Figure 4 shows plots for two of these correlations – the correlation for the basic monophthong inventory (left panel), and the correlation for the full consonant inventory (right panel). The points are plotted with the names of the language families, to give some sense of how the different language families are distributed across the space.

In these correlations, each language family is reduced to a single point. This analysis therefore eliminates the possibility that the observed correlation between population size and inventory is being carried by one or two over-represented language families in our sample.

Table 3: Spearman’s correlation between mean language family population, and mean inventory size.

Correlation with:	R^2	P<
All phonemes	.46	.003
Basic Monophthongs	.47	.002
All Monophthongs	.2	.2
Diphthongs	.53	.001
Plosives	.33	.05
Fricatives	.53	.001
Sonorants	.24	.13
Obstruents	.43	.005
All Consonants	.45	.005

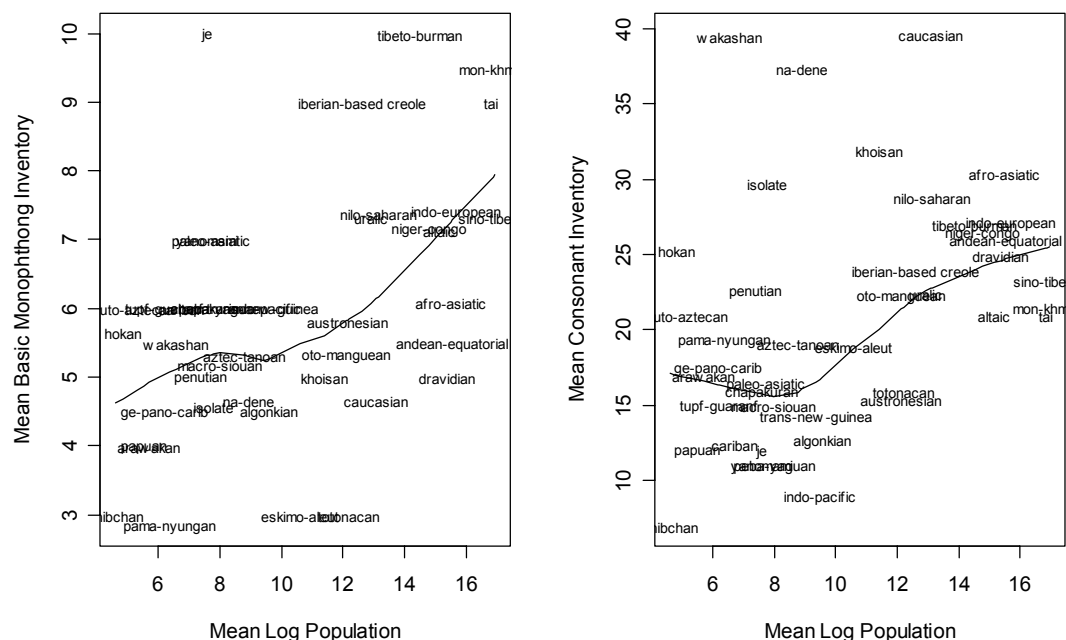


Figure 4: Mean log population and mean basic monophthong inventory (left panel) and consonant inventory (right panel). Each point represents a language family.

Again, within this sample there is absolutely no correlation between the mean number of vowels a family has and the mean number of consonants ($-.03, p=.87$). This makes the fact that both correlate with mean population even more remarkable.

A reviewer of this paper suggested that it could be possible that the number of ‘phonemes’ in a language tends to increase as the language is studied, and that languages spoken by more speakers tend to receive more attention. This is an intriguing suggestion, which would provide a sociological explanation for our observed effect. However this portion of our analysis has reduced each language family to one data point – investigating the mean population size and phoneme inventory by language family. This eliminates the possibility that a few highly studied language families (such as Indo-European) might be driving the effect in this

way. The fact that the correlation is robust both across languages and language families suggests strongly that there is something here that requires explanation.

Discussion

There is a surprisingly strong relationship between the size of a language's phoneme inventory and the number of speakers of that language. While such a correlation has not (to our knowledge) been reported before, some researchers have, in fact, speculated that there may be a link between the size of a community and the phoneme inventory.

For example Haudricourt (1961, as cited in Trudgill 2002) has argued that small inventories are the result of "impoverishment" which occurs in situations characterized by monolingualism, and isolation and /or by "non-egalitarian bilingualism". Haudricourt suggests that in certain environments a particular group may be sufficiently dominant that they have no motivation to articulate clearly. Such people are able to confuse two phonemes, or omit to produce a phoneme without fear of "mocking". "This is why we find fewer consonants in the language of the Iroquois who terrorised their neighbours, or in the languages of the people of Tahiti and Hawaii who combine island isolation with significant demographic development as compared to other less favoured archipelagos" (Haudricourt 1961 as cited in Trudgill 2002; Trudgill's translation). Trudgill appears relatively unconvinced by Haudricourt's interpretation, but does agree with the general prediction that isolated communities may have smaller inventories. This would be so because "initial small community size ... would have led in turn to tight social networks, which would have implied large amounts of shared background information - a situation in which communication with relatively low level of phonological redundancy would have been relatively tolerable" (Trudgill 2002: 720). The factors Trudgill (1995: 356) believes may lead to small phoneme inventory size include isolation from contact with other languages, initial small community size, tight social networks and large amounts of shared background information. This hypothesis suggests that, in a fuller investigation, one should perhaps also attempt to take into account the degree to which each language is isolated from contact with other languages. This may well be partially correlated with population size, and responsible for some of what we have

found. However this would be a much more major undertaking, and beyond our much more modest goals.

In addition to small community size potentially leading to small phoneme inventories, Trudgill claims that small communities can also lead to very large inventories. This is because of “the ability of such communities to encourage continued adherence to norms from one generation to another, however complex they may be” (Trudgill 2004a: 317). Thus, Trudgill claims, the combined effects of isolation, network structure and language contact should lead languages with small populations to have either very small or very large inventories, and languages with larger populations to favour “medium-sized populations” (2004a: 17). While we have clearly found some evidence of smaller populations favouring smaller inventories, there is no evidence in this data set that they also favour larger inventories.

In a commentary on Trudgill’s (2004a) paper, Pericliev (2004) investigates the relationship between consonant inventory and population size in a set of 417 languages. He examines the data in various ways, and concludes (2004: 382) that “there is no correlation of the kind suggested by Trudgill between the size of a community speaking a language and the size of the consonantal inventory of that language”. As his purpose is to provide counter-evidence for Trudgill’s theory that larger populations should lead to medium-sized inventories, Pericliev does not actually test for a straightforward linear relationship between population size and inventory. However he has kindly shared his data with us, and we have tested this correlation. It is highly significant (spearman’s $\rho = .21$, $p < .0001$). That the overall correlation between phoneme inventory and population size is significant in this larger sample provides strong evidence that the observed correlation is not an artefact of our sampling procedure. Pericliev’s sample was collected entirely independently and includes a different (larger) sample of languages. Yet it also contains the same correlation that we have observed.

In defending his thesis against Pericliev (and other commentaries), Trudgill argues that the effects of population size, network structure and language contact situation need to be considered together, and so there would be “no reason at all to expect to find a simple correlation between the numbers of speakers in a language and the

number of phonemes in that language” (2004b: 386). While we agree with Trudgill that there is no obvious reason to expect such a correlation, the data we discuss in this paper certainly suggests that such a correlation exists.

One possible explanation for the correlation may come from issues relating to learnability. It would be a large leap to assume that speakers of languages with smaller populations are exposed to a narrower range of speakers (and/or dialects) than speakers of languages with larger populations. After all, each individual speaker of a language certainly doesn’t necessarily interact with every other speaker. It is a tempting leap, though, because if this were true, it would suggest an explanation in terms of the robustness and learnability of categories based on this different exposure.

Experiments designed to teach non-native phoneme distinctions show better learning, and considerably better long-term retention if multiple voices are used in training (Lively et al. 1993, 1994; Logan et al. 1991). Results on listener adaptation to foreign-accented English also demonstrate that “exposure to talker variability also facilitates rapid, talker independent perceptual learning of a foreign accent which involves a wide range of acoustic-phonetic features” (Bradlow and Bent 2005: 2884).

Such results tempt one to speculate that exposure to less variability would lead to less robustness of phonemic categories. Exposure to variability is important, as “variability causes the need for abstraction” (Pierrehumbert, Beckman and Ladd 2001).

The learning of phonemes involves abstraction over learned distributions of speech sounds (see, e.g. Pierrehumbert 2000). The more exposure to more different speakers, the denser these distributions presumably are. Work on the acquisition of phoneme categories (Maye and Gerken 2000, Maye, Werker and Gerken 2002, Maye and Weiss 2003) shows that infants use distributional information in the signal in order to discern phoneme boundaries. That is, when an infant (or adult) is exposed to tokens from a particular phonetic space in a uni-modal distribution, they tend to learn this as a single category. When a distribution over the same phonetic space is bimodal, it is learned as two categories. Increased exposure to large number of speakers would lead to denser distributions and so (presumably) make learning of this kind more robust. With sufficient exposure, categories could be easily learned which would be difficult with more limited, less varied, exposure.

Variability facilitates the learning of categories, and repeated prolonged exposure also sharpens the boundaries of these categories. Lee et al. (1999) show that phoneme boundaries sharpen considerably right up until late childhood. Pierrehumbert (2001) builds an effect of "entrenchment" into her exemplar-theoretic model in order to simulate this kind of effect: categories become sharper with repeated exposure.

The effects of the size of a population of speakers are also revealed by multi-agent modelling work. For example Bart de Boer (2000, 2001) has done work constructing computer-based models which simulate the emergence and transmission of vowel systems. The models work by simulating a population of speakers which can produce and perceive vowels. After a series of iterations in which the agents attempt to imitate one another's productions, vowel systems emerge in the population which resemble human vowel systems. While investigating the properties of the parameters of his model, de Boer (2000) manipulates the size of the population, finding that in his modelling "the success of all population sizes is comparable, but the vowel system size of small populations is smaller than that of large ones, reflecting the lower stability". The effect of interaction amongst a large number of speakers is to increase the stability of systems with many vowels. Of course, to suggest that this is responsible for our correlation, supposes that the number of speakers of a language is somehow correlated with the number of different speakers an individual is exposed to over the course of their lifetime. While this may be true, it is not necessarily so. And the degree to which de Boer's prediction would follow over to consonants is not so clear: most of this kind of work has focused on vowel systems.

Regardless of the explanation for the correlation, its existence raises a number of further empirical questions. One is the question of whether the relationship between population size and phoneme inventory is also relevant across dialects of individual languages. Are dialects spoken by fewer speakers also likely to have fewer phonemic categories? In addition, if there is a causal relationship between population size and phoneme inventory size, then we might also expect to see covariation of population size and phoneme inventories over time². Neither the population of speakers of a language nor its phoneme inventory are constant. Some languages undergo drastic changes in their population of speakers due to catastrophic factors

² Our thanks to Brian Joseph for this observation.

such as disease or genocide. The results presented here raise the question of whether such changes in population have a tendency to lead to changes in phoneme inventory size. Such changes need not be cotermporal of course – changes in phoneme inventory size may lag considerably behind population trends. It would, of course, be a surprising finding if fluctuations in population size were indeed paralleled in the phonemic system, and we would not necessarily want to commit to such a prediction. But our results point to this it as an intriguing question.

Conclusion

We have reported a positive correlation between how many phonemes a language has, and how many speakers it has. This correlation exists both within the vowel inventory and within the consonant inventory. This is not an artefact of language family. We do not know what the underlying causes of this correlation are. But it is certainly intriguing, and we hope that this short discussion note will generate some discussion of the possible causes of such a relationship.

References

- Bauer, Laurie (in prep.) *The Linguistics Student's Handbook*. To be published by Edinburgh University Press.
- Bradlow, Ann R. and Bent, Tessa (2003) Listener adaptation to foreign accented English. In M. J. Sole, D. Recasens, & J. Romero (Eds.), *Proceedings of the XVth International Congress of Phonetic Sciences*, Barcelona, Spain, Pp. 2881-2884.
- Chao, Yuen-Ren (1934) The non-uniqueness of phonemic solutions of phonetic systems. Reprinted in Martin Joos (ed.) (1957) *Readings in Linguistics I*. Chicago and London: Chicago University Press, 38-54.
- Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of American Statistical Association*, 74, 829-836.
- De Boer, Bart (2001) *The Origins of Vowel Systems*. Oxford University Press, Oxford.
- De Boer, Bart (2000) Self organization in vowel systems, *Journal of Phonetics* 28 (4), 441-465.
- Grimes, Barbara F. (1988). *Ethnologue*. Dallas, TX: SIL.
- Harrell, F. E (2001) *Regression Modelling Strategies*. Springer, Berlin.
- Haudricourt, André (1961) Richesse en phonèmes et richesse en locuteurs. *L'Homme* 1: 5-10.
- Lee, S. , Potaminos, A. and Narayanan, S. (1999) Acoustics of children's speech: Developmental changes in temporal and spectral parameters. *Journal of the Acoustical Society of America* 105, 1455-1468.
- Lively, S. E, Pisoni, D.B, Yamada, R.A, Tohkura, Y. and Yamada, T. (1994) Training Japanese listeners to identify English /r/ and /l/: III. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America*, 96, 2076-2087.
- Lively, S.E., Logan, J.S, and Pisoni, D.B (1993) Training Japanese listeners to identify English /r/ and /l/: II. The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, 94, 1242-1255.
- Logan, J.S., Lively, S.E, and Pisoni, D.B (1991) Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, 89, 874-886.
- Maddieson, Ian (2005) Vowel Inventories. In Haspelmath, Martin, Matthew S. Dryer, David Gil, and Bernard Comrie (eds) *The World Atlas of Language Structures*. Oxford University Press, Oxford, pp 14-15.
- Maye, Jessica & Daniel Weiss (2003) Statistical cues facilitate infants' discrimination of difficult phonetic contrasts. *Proceedings of the 27th Annual Boston University Conference on Language Development*: 508-518.

Maye, Jessica, Janet Werker & LouAnn Gerken (2002) Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82 (3): B101-B111. 8-518.

Maye, Jessica & LouAnn Gerken (2000) Learning phoneme categories without minimal pairs. *Proceedings of the 24th Annual Boston University Conference on Language Development*: 522-533.

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

Pierrehumbert, Janet (2000) What people know about sounds of language. In *Studies in the Linguistic Sciences* 29(2), 111-120.

Pierrehumbert, Janet (2001) Exemplar dynamics: Word frequency, lenition, and contrast. In Bybee, J. and P. Hopper (eds) *Frequency effects and the emergence of linguistic structure*. John Benjamins, Amsterdam, 137-157.

Pierrehumbert, Janet, Beckman, Mary E and Ladd, D. Robert (2001) Conceptual Foundations of Phonology as a Laboratory Science, in Burton-Roberts, N., Carr, P. and Docherty, G. (eds) *Phonological Knowledge*, Oxford, UK: Oxford University Press, 273-304.

Thráinsson, Höskuldur, Hjalmar P. Petersen, Jógvan Í Lon Jacobsen & Zakaris Svabo Hansen (2004) *Faroese*. Tórshavn: Føroya Fróðskaparfelag.

Trudgill, Peter (1995) Dialect Typology: Isolation, Social Network and Phonological Structure. In Gregory Guy, Crawford Feagin, Deborah Schifffrin and John Baugh (eds) *Towards a Social Science of Language. Vol 1: Variation and Change in Language and Society*. John Benjamins Publishing Company, Amsterdam/Philadelphia, pp 3-22.

Trudgill, Peter (2001) Linguistic and Social Typology. In J. K. Chambers, Peter Trudgill and Natalie Schilling-Estes (eds) *The Handbook of Language Variation and Change*. Blackwell Publishers Ltd, Massachusetts/Oxford. 707-728

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.

Appendix: Languages in Database:

!Xu, Abkhaz, Acoli, Afrikaans, Akan, Albanian, Amele, Amharic, Amoy, Apalai, Arabana-Wangganguru, Arabic, Armenian, Arrernte, Basque, Bengali, Berber, Blackfoot, Breton, Bulgarian, Burmese, Burushaski, Canela-Kraho, Cantonese, Cashinahua, Catalan, Cherokee, Cheyenne, Chipewyan, Chukchee, Cree, Croatian, Crow, Czech, Dakota, Dan, Dani, Danish, Daur, Dinka, Diyari, Dutch, Dyirbal, Efik, English, Erromangan, Estonian, Evenki, Ewe, Faroese, Farsi, Fijian, Finnish, Fore, French, Friesian, Fula, Gaelic, Georgian, German, Gilyak, Greek (modern), Guarani, Gujarati, Haida, Hausa, Hawaiian, Hebrew, Hindi, Hixkaryana, Hopi, Hungarian, Icelandic, Igbo, Ijo, Illocano, Indonesian, Irish, Italian, Japanese, Javanese,

Kabardian, Kamba, Kambara, Kannada, Kanuri, Karen, Kashmiri, Ket, Khmer, Kilivila, Kiowa, Kirghiz, Klamath, Kobon, Koiari, Korean, Kota, Kpelle, Kurdish, Kwakwala, Laotian, Latvian, Lenakel, Lithuanian, Luiseno, Maasai, Madi, Maidu, Malagasy, Malay, Malayalam, Maltese, Mam, Mandarin, Maori, Marathi, Margi, Mari, Mazateco, Meithei, Mende, Miwok, Mixtec, Mongolian, Nahuatl, Nama, Navajo, Nez Perce, Ngiti, Nootka, Norwegian, Ojibwa, Oneida, Oromo, Ostyak, Panjabi, Papiamentu, Pashto, Pima, Piraha, Pitjantjatjara, Polish, Pomo, Portuguese, Provencal, Quechua, Quiche, Quileute, Rapanui, Romanian, Romany, Rotuman, Russian, Saami, Samoan, Sanskrit, Sanuma, Seneca, Serbian, Shona, Shoshone, Sindhi, Sinhalese, Slovakian, Slovenian, Somali, Sorbian, Sotho, Spanish, Sundanese, Swahili, Swedish, Tagalog, Tahitian, Tamil, Telugu, Tetun, Thai, Tibetan, Tigrinya, Tiwa, Tiwi, Tlingit, Toba Batak, Tokelauan, Tol, Tongan, Totonaco, Trukese, Tswana, Tukang Besi, Turkish, Tuvaluan, Tzeltal, Ukrainian, Ulithian, Urubu-Kaapor, Vietnamese, Wai Wai, Warekena, Wari, Warlpiri, Washoe, Welsh, West Greenlandic, Wintu, Wolof, Yagua, Yiddish, Yimas, Yoruba, Zapotec, Zoque, Zulu, Zuni