There is no correlation between the size of a community speaking a language and the size of the phonological inventory of that language

by VLADIMIR PERICLIEV

Abstract

In the target article, Trudgill assumes, based on the inspection of some Austronesian/Polynesian languages, that large community size favours medium-sized phonological inventories, whereas small community size favours either small phonological inventories or large inventories, and he then undertakes to explain these "facts". A crosslinguistic empirical test, however, reveals conclusively that such assumptions are invalid and therefore Trudgill's explanation is fallacious in explaining a phenomenon that does not exist.

Keywords: phoneme inventories, social structure

1. Hypotheses

In the target article, Peter Trudgill makes (among others) the following two interrelated FACTUAL claims about the existence of a correlation between the size of a community speaking a language and the size of the phonological inventory of that language (cf. his Conclusions (iv) and (v), respectively):

Claim I: Large community size favours medium-sized phonological inventories.
Claim II: Small (=non-large) community size favours either small phonological inventories or large inventories (but not medium-sized ones).

The claims assume that the universe of all inventories is split exhaustively into three categories: "small", "medium", and "large", and the universe of all linguistic communities is split exhaustively into two categories: "small" and "large (= non-small)". The claims can then be represented as follows:

	All communities:	All inventories:
Claim I:	large	medium
Claim II:	small (=non-large)	small or large (non-medium)

meaning that languages spoken by large communities will favour medium-sized inventories and, vice versa, languages having medium-sized inventories will favour large communities (Claim I), and languages spoken by small communities will favour either small or large inventories and, vice versa, languages having either small or large inventories will favour small communities (Claim II). This posits a one-to-one correspondence (= equivalence, two-way

implication) between large communities and medium-sized inventories on the one hand, and small communities and small/large inventories on the other. Put still differently, the claims amount to saying that the space of all languages is partitioned in two: (i) languages that BOTH have medium-sized inventories AND are spoken by large communities; and (ii) languages that BOTH have small/large inventories AND are spoken by small communities.

In his article, Trudgill is primarily concerned with Austronesian/Polynesian languages, and rightly notes (p. 316) that "[i]n the absence of a large-scale database of evidence on this topic, taken from different language families in different parts of the world, any conclusions to be produced here [in his article] can be only suggestive and tentative".

The goal of this comment is to report a crosslinguistic test of the two claims above. I should note at this stage that, judging from the context and the examples Trudgill gives, by "phonological inventory" he means the CONSONANTAL inventories of languages (rather than inventories including both consonants and vowels), and I will therefore focus on consonantal inventories, using the latter complex term for greater clarity of exposition.

2. Sampling and data collection

As a preliminary step of the testing, we need to collect data on the size of consonantal inventories of languages and the size of the populations speaking those languages. I therefore compiled such a database on the basis of the 451 languages and their phonological inventories as described in the UPSID-451 sample (Maddieson & Precoda 1991, Maddieson 1991, cf. also Maddieson 1984). The number of consonants for each of these 451 languages was computed, and then the corresponding sizes of the linguistic communities speaking the languages were added, based on the information contained in the electronic version of *Ethnologue*. For 23 languages (coming from different language families/areas of the UPSID sample) I could not provide information about the number of their speakers (because the languages are extinct, there is no estimate about speakers, or they do not figure in *Ethnologue*). This left us with 428-odd languages containing the relevant information. The resultant sample clearly inherits UPSID's representativeness and, just as UPSID itself, can be used for making valid statistical inferences about the (consonantal) segment inventories of the languages of the world.

3. Definitions

In order to be able to examine Claims I and II, we need to have a specific NUMERICAL INTERPRETATION of the meanings of the terms participating in the claims, especially as regards the meanings of "small inventory", "medium-sized inventory", and "large inventory", and of "large community" and "small

community". Trudgill does not give an exact numerical interpretation of these terms. This makes the claims insufficiently specific and thus not amenable to direct testing. Indeed, is, e.g., a "small community" one comprising 100, 1000, 10,000, 25,000, or more speakers? Do 17 consonants constitute a "small" or a "medium-sized" inventory? Or do 26 consonants constitute a "medium-sized" or a "large" inventory? Depending on the specific numerical interpretations of these terms the outcomes of a test may vary.

There are two different approaches to handling this indeterminacy of terms in the inspected propositions: (A) We select some "reasonable" interpretations of the terms and test the claims with these interpretations; (B) we try to determine whether there are interpretations of the terms for which the claims are actually true. In Section 4 I follow approach A, and in Section 5 approach B.

4. Approach A: Method and results

Let us first look for a "reasonable" numerical interpretation of "medium-sized inventory" (a "small inventory" will then be one which is smaller than the "medium-sized" one, and a "large inventory" one which is larger than the "medium-sized" one). The mean of consonantal inventories is about 22, and one standard deviation of the mean is 9, giving the interval 22 $\pm$9 as a reasonable interpretation of a "typical" or "average" consonantal inventory. The interval 22 $\pm$9 comprises 78 % of all languages in our 428 language sample. Varying this interval slightly, I presume, will also yield intuitively plausible average consonantal inventories.

Insofar as the demarcation between a "large community" and a "small community" is concerned, it seems to me that any number from, say, 1,000 to 100,000 can split the languages into ones spoken by small vs. large populations in a way that is intuitively satisfactory.

We are now ready to test Claims I and II empirically for DIFFERENT reasonable interpretations of their component terms, by selecting some combinations of such interpretations. Let us call the interval 22 $\pm$9 a (reasonable) INVENTORY SIZE DEMARCATOR and any integer within the interval 1,000–100,000 a (reasonable) COMMUNITY SIZE DEMARCATOR. Selecting a specific value for each demarcator turns the indeterminate Claims I and II into completely definite statements amenable to testing. Thus, e.g., selecting the Community size demarcator 5,000 and the Inventory size demarcator 13–31 turns Claims I and II respectively into:

Claim I′: Community sizes $>$ 5,000 speakers (large ones) favour inventories between 13 and 31 consonants inclusive (i.e., medium-sized ones); and

Claim II′: Community sizes $\leq$ 5,000 speakers (small ones) favour either less than 13 consonants (small inventories) or more than 31 consonants (large inventories).

Table 1. *Fourteen tests of Claims I and II under different "reasonable" interpretations of the terms "small" vs. "large" community size and "small" vs. "medium-sized" vs. "large" consonantal inventory*

Test No.	Community size demarcator	Inventory size demarcator	Validity of Claim I		Validity of Claim II	
1	1,000	13–31	262/405	65%	23/166	14%
2	5,000	13–31	226/387	58%	41/202	20%
3	25,000	13–31	171/370	46%	58/257	23%
4	50,000	13–31	149/364	41%	64/279	23%
5	100,000	13–31	130/359	36%	69/298	23%
6	1,000	18–26	162/365	44%	63/266	24%
7	5,000	18–26	141/332	43%	96/287	33%
8	25,000	18–26	114/287	40%	141/314	45%
9	50,000	18–26	102/271	38%	157/326	48%
10	100,000	18–26	89/260	34%	168/339	50%
11	10,000	19–25	113/288	39%	140/315	44%
12	15,000	19–25	110/279	39%	149/318	47%
13	35,000	19–25	90/254	35%	174/338	51%
14	80,000	19–25	76/240	32%	188/352	53%

Table 1 summarizes the results of a number of tests of Claims I and II performed computationally, using as a database our 428 language sample (cf. Section 2). The tests use different interpretations of the component terms of Claims I and II, which were chosen randomly from the intuitive values agreed upon above. The meanings of columns in Table 1 are self-explanatory. Columns 4 and 5 give the positive examples and all relevant examples (separated by slash) and the percentage of validity of the respective claim.[1]

This suite of random tests strongly suggests that Trudgill's Claims I and II are not valid. Both claims are generally valid below or around the threshold of 50 %, and there are therefore no linguistic PREFERENCES of the types suggested by Trudgill. Only in tests Nos. 1 and 2 does Claim I slightly exceed the 50 % threshold, with values of 65 % and 58 % respectively. However, these results do not provide any (even minor) support for Trudgill's claims either. There are two reasons for this. First, these language distributions are simply to be expected by chance, and therefore present no significant clustering indicative of a

1. As usual, I count as positive examples the languages which both are spoken by large communities and have medium-sized inventories (Claim I) or both are spoken by small communities and have small or large inventories (Claim II). Counterexamples are the languages satisfying only one of the two conjoined conditions (e.g., a language spoken by a large community but having a small inventory will be a counterexample to either claim). The number of "relevant examples" is equal to the sum of positive examples and the counterexamples.

real linguistic preference.[2] And, secondly, in these two tests the corresponding two companion Claims II, which are also crucial to Trudgill's argumentation, are hopelessly unsupported (with validity of 14 % and 20 % respectively).

It could be (correctly) objected that a definite conclusion as to the validity of tested claims is premature since our suite of tests is not exhaustive by far and involves only a dozen of combinations of term interpretations, while there is a vast number of logically possible ones, leaving the theoretical possibility for finding term interpretations for which the claims are valid.

5. Approach B: Method and results

To remedy this, let us now pursue the second approach B by conducting a graphical test of Trudgill's claims. Let us plot the languages in the sample in a *xy* scatter diagram, where the *x*-axis displays the size of consonantal inventories and the *y*-axis the number of speakers. Each language in the sample will be represented by a point, whose position in this diagram will be determined by the number of consonants it has and by the number of people the language is spoken by.

What would Trudgill's claims amount to viewed from this graphical perspective?

Figure 1 is a graphical representation of an artificial language data set I generated computationally,[3] consisting of 428 languages (as our original sample), which conforms 100 % to Trudgill's Claims I and II. (Note that the *y*-axis is an

2. E.g., let us look at test No. 1, Claim I (Table 1), which is supported in 262 languages (or $^{262}/_{405}$ = 65 % validity) as are spoken by more than 1,000 people and also have inventories from 13 to 31. As already mentioned, 78 % of all languages in the sample (or 335 languages out of 428) have consonantal inventories in the interval 22 ±9, i.e., from 13 to 31. I computed that 336 languages are spoken by more than 1,000 people, giving also 78 % of all languages in the sample. Now, the probability of the occurrence of one language which both has more than 1,000 speakers and has from 13 to 31 consonants will be equal to the product of these two percentages, i.e., $0.78 \times 0.78 = 0.61$ (or 61 %). Now we ask: How many languages having both properties (= positive examples) are to be expected to occur by chance in the whole sample of 428 languages? To find this, we should multiply 428 by the probability 0.61, giving $428 \times 0.61 = 261$ languages. Thus, we expect 261 languages as chance support and we find almost the same number, 262, in the actual test.

3. For those interested, I give the basics of the algorithm. First, the Community size demarcator and the Inventory size demarcator were chosen to be 10^4 = 10,000 speakers and 20–30 consonants, respectively. Then, for any language in our 428 language sample, the following was done: if it conforms to either claim, it is listed with unchanged descriptors in the artificial sample; if it does not conform to any of the two claims, the size of its consonantal inventory is randomly changed to conform to one of the claims (this category was chosen which was closer to the original consonantal number, e.g., for a language spoken by a small community but having a medium-sized inventory of 21 consonants, a random number would be chosen below (not above) this interval. In effect, the algorithm generates a "corrected version" of the original sample.

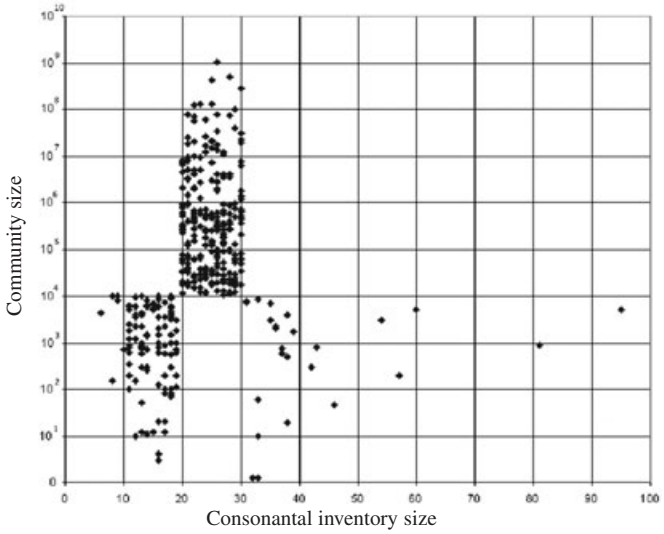

Figure 1. *Distribution of an artificial 428 language sample conforming 100 % to Claims I and II*

exponential scale with a base of 10, i.e., 10^1 = 10 speakers, 10^2 = 100 speakers, 10^3 = 1,000 speakers, etc.)

As Figure 1 clearly shows, and this would be the case for ANY conforming language sample, Claims I and II could be true[4] only if all (or at least, most) languages cluster in three regions: an upper middle region (corresponding to Claim I: large communities correlate with medium-sized inventories), and a lower left region or a lower right region (corresponding to Claim II: small communities correlate with either small or large inventories, respectively).

In the particular example, both claims are true under the following numerical interpretation of the claims' component terms: Community size demarcator 10^4 = 10,000 and Inventory size demarcator 20–30 (which is equivalent to saying that: a "small community" is smaller than or equal to 10,000 speakers, a "large community" is larger than 10,000 speakers; a "small inventory" is smaller than 20 consonants, a "medium-sized inventory" lies in the interval 20–30 consonants, and a "large inventory" is greater than 30 consonants).

4. I say "could be true", and not "will be true" because clustering in the mentioned regions is a necessary, but not also a sufficient condition for Claims I and II to be true. Thus, some clusterings may arise due simply to chance, and therefore a candidate cluster should additionally be shown to be statistically significant. See also Footnote 2.

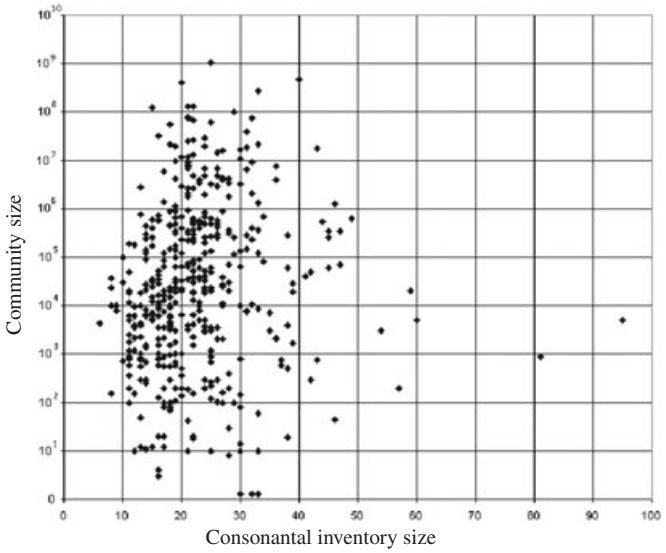

Figure 2. *Distribution of the actual 428 language sample*

Now we can look at the actual distribution of our original 428 language sample and compare the result to that of the artificial (and conforming) sample, given in Figure 1. The actual distribution is given in Figure 2. The graph in Figure 2 is markedly different from that in Figure 1. There is not a trace of any of the three distinct clusters in Figure 1, meaning that Claims I and II are not really supported under ANY INTERPRETATION of their component terms. Our graphical testing thus enforces the preliminary impression of the falsehood of the claims, arrived at by other means in Section 4.

6. Conclusions

We can safely conclude then 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. (Neither does such a correlation exist when whole inventories, i.e., consonants AND VOWELS, are involved, as preliminary tests suggest to me.)

Finally, let me emphasize that my considerations in this note in no way undermine Trudgill's more general concern for the search for significant correlations between societal types and linguistic structure. I believe, however, that such undoubtedly fruitful pursuits, of a mostly theoretical nature, had better be complemented by extensive data collection and the severe testing of the theoretical hypotheses. Of course, saying this I am not denying the role of rea-

soning, as distinct from direct empirical testing, in both hypothesis formation and in hypothesis testing.[5] Thus, the analysis of Claims I and II (performed in Section 1) is quite suggestive of their doubtful nature. Indeed, once we have recognized, owing to that analysis and without any recourse to data inspection, that a one-to-one correspondence is assumed between large communities and medium-sized inventories (Claim I) and between small communities and small (or large) inventories (Claim II), it immediately becomes conspicuous that even if we accept Trudgill's argumentation in the direction size of community "causes" size of inventory, there can hardly be a causal connection in the other direction, viz. size of inventory "causes" size of community. The reason is simply that the size of a phonological inventory of a language – irrespectively of whether this size is related to communicative efficiency (as assumed by Trudgill) or not, or whether it is related to whatever other linguistic property – can hardly be expected to significantly affect the birth and death rates of the population speaking that language.

Received: 25 November 2003
Revised: 8 January 2004

Bălgarska Akademija na Naukite

Correspondence address: Mathematical Linguistics, Institut po matematika i informatika, bl. 8, Bălgarska Akademija na Naukite, 1113 Sofia, Bulgaria; e-mail: peri@math.bas.bg

Acknowledgement: I am grateful to my daughter Violetta Periclieva, a student of Bulgarian Philology at Sofia University, for collecting from the electronic version of *Ethnologue* the information on the number of speakers for the 451 languages contained in UPSID.

References

Maddieson, Ian (1984). *Patterns of Sounds*. Cambridge: Cambridge University Press.

— (1991). Testing the universality of phonological generalizations with a phonetically specified segment database: Results and limitations. *Phonetica* 48: 193–206.

Maddieson, Ian & Kristin Precoda (1991). Updating UPSID. *UCLA Working Papers in Phonetics* 74: 104–114.

Pericliev, Vladimir (1990). On heuristic procedures in linguistics. *Studia Linguistica* 43-2: 59–69.

5. For a discussion and illustration of various heuristic reasoning methods see, e.g., Pericliev (1990).