THE QUANTITATIVE PHONETIC ANALYSIS OF LANGUAGES WITH SMALL ACCESSIBLE POPULATIONS

Márton Sóskuthy[1], Una Chow[2], Gloria Mellesmoen[3], Emily Sadlier-Brown[4]

[1–4]Department of Linguistics, University of British Columbia
[1]marton.soskuthy@ubc.ca, [2]una.chow@ubc.ca

ABSTRACT

Many of the world's languages – especially Indigenous languages – have small accessible populations. Such languages present challenges for quantitative phonetic analysis, which can in turn hamper the dissemination of results. This is due to phoneticians' overreliance on null hypothesis significance testing. Even data from a single speaker can provide valuable insights if a suitable approach is used. We present a case study that focuses on the VOT of ejectives in Gitksan based on a single speaker. We show how issues that stem from a small speaker pool can be remedied using a Bayesian approach that incorporates prior information from other languages. We also discuss further alternatives to null hypothesis significance testing such as foregrounding descriptive statistics and exploratory methods, and a flexible interpretation of what constitutes the statistical population for linguistic studies.

Keywords: Indigenous languages, small population, phonetic fieldwork, quantitative analysis

1. INTRODUCTION

According to the *Ethnologue* database [1], over 50% of the world's languages have fewer than 10,000 speakers, over 25% have fewer than 1,000 and over 12% have fewer than 100. In general, the smaller the population size of a language, the more difficult it is to access and recruit fluent speakers for phonetic study – though ease of access also depends on other factors such as geographical isolation. This paper uses the label *languages with small accessible populations* (LSAP) to refer to languages with a very small speaker pool and/or difficulties accessing speakers.

Indigenous languages are often LSAPs, including First Nations languages in British Columbia, Canada (the focus of some of our own work). In a 2018 survey representing 177 of 203 First Nations communities in BC (with a combined population of 137,653 individuals), only 3% of individuals spoke the language of their community fluently; among these speakers, 52% were aged 65+ and only 2.8% were under the age of 24 [2]. Government policies (including the residential school system) have disrupted language transmission in Indigenous communities, leading to low numbers of speakers. For example, there are fewer than 30 L1 speakers of Comox-Sliammon (Salish), and therefore a small population to recruit from. The representation of these languages in the literature is limited by the circumstances and total number of language users.

The phonetic study of Indigenous LSAPs is of utmost importance. It can feed into language teaching and revitalisation efforts, and it can enrich Indigenous peoples' experience with their language. These languages can also yield invaluable linguistic insights and broaden perspectives on patterns in speech, which are often heavily biased towards well-documented languages with large and easily accessible speaker populations [3,4].

However, the small size of the speaker samples that represent these languages constitutes a significant barrier to phonetic analysis. Researchers with restricted sample sizes cannot be sure how well their sample represents the language at large, which adds uncertainty to their conclusions. This is an issue regardless of the field of enquiry, but it is brought into especially stark focus in quantitative studies that rely on statistical analysis – a common approach in modern instrumental phonetics. As a result, phoneticians working on LSAPs often face significant challenges in data analysis, which in turn hampers dissemination of the results.

We argue that many of the difficulties surrounding work on LSAPs stem from a widespread overreliance on hypothesis testing using *p*-values (commonly referred to as *null hypothesis significance testing* or NHST). It has long been recognised that NHST is but a single tool in a much broader statistical toolkit [5]. In recent years, the literature on statistics in cognitive science has marked a move towards alternative approaches such as parameter estimation with uncertainty and exploratory data analysis [6,7]. We suggest that these tools allow phoneticians to better utilise the information in data from LSAPs. This paper serves two goals: (i) to provide an example of how such an approach can be implemented in practice and (ii) to convince analysts, reviewers and editors that quantitative studies of LSAPs are worthwhile.

In what follows, we first outline the main statistical problem in more detail. We then present a case study of voice onset time (VOT) in ejective stops in a single speaker of Gitksan, a Tsimshianic language spoken in

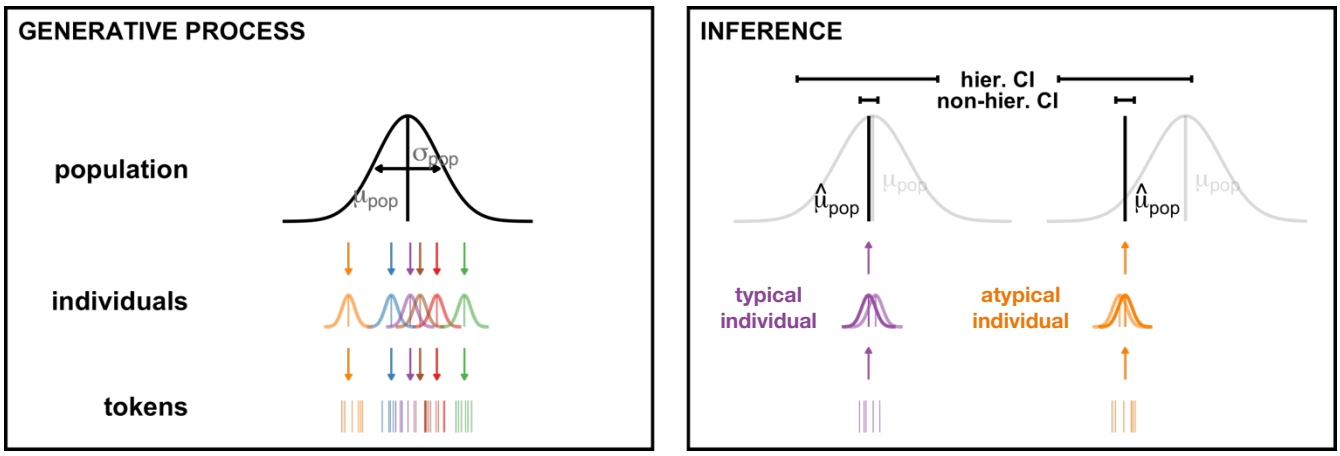

Figure 1: A graphical illustration of the main issue with statistical inference for LSAPs. The left-hand panel shows a schematic model of how phonetic data are generated by taking random individuals from a population and then sampling random tokens from those individuals. The right-hand panel shows what inferences can be made when phonetic data are only available from a single typical (left) or atypical (right) individual (hier. = "from a hierarchical model").

north-western British Columbia. We show that a Bayesian analysis utilising prior information from other languages can provide estimates of ejective VOT in Gitksan with a meaningful measure of uncertainty that allows for phonetic inferences. We also make suggestions as to what other tools could be used in the analysis of LSAPs and how reviewers and editors could approach evaluating work on these languages.

2. THE STATISTICAL CHALLENGE

Data from a limited number of speakers tend to strike analysts, reviewers and editors as insufficient. This occurs even when the small sample of speakers in a given study constitutes the entire extant speaker population. But why? Most of the time, we are interested in what is typical in a given language. In statistics, typicality is usually defined in terms of some central tendency (e.g. the mean) of the *statistical population* [8], which is often not the same as the accessible population. To give an example, in a medical study we might want to test a treatment that should work for all humans alive and yet to be born. This is our statistical population. However, it is theoretically impossible to test the treatment on humans who have not been born, and it is practically impossible to test it on all humans who are currently alive. Not all of the statistical population is accessible, and therefore the *statistical* and *accessible population* are not the same.

The same considerations arise in studies of LSAPs. The statistical population is often understood as a reasonably large set of speakers from the recent history of the language. Many of these speakers may not be accessible either because they are no longer alive or for practical reasons. Since typicality is defined in terms of the statistical population, the accessible population may or may not reveal typical patterns in the language.

Fig. 1 is a schematic illustration of the issue of typicality. The left-hand panel shows a common conceptualisation of how phonetic data are generated (and one that is implicitly assumed by anyone who uses hierarchical models, also known as mixed effects models [9]). Displayed horizontally is some phonetic quantity, such as VOT. The normal distribution at the top of the panel represents the statistical population, with the mean (μ_{pop}) and standard deviation (σ_{pop}) shown by the vertical line and horizontal arrows. Individuals sampled from this population are often fairly close to the population mean in terms of their phonetic behaviour (e.g. the purple and brown distributions in the middle), but sometimes they are quite far from it (e.g. the orange distribution on the left). What phoneticians observe, however, are not these individual-level distributions, but the actual tokens produced by the individuals, plotted at the bottom. These tokens show variation as well, and can sometimes deviate from the underlying individual-level patterns.

The right-hand panel shows two possible scenarios for statistical inference, one based on a single typical speaker and the other on a single atypical speaker. The analyst observes only the tokens plotted at the bottom. Based on these tokens, they estimate the individual-level distribution (darker shade), which may be different from the original distribution that generated the tokens (lighter shade). This is due to token-level variation and biases in sampling. Since only a single speaker is available in each scenario, the best estimate for the population-level mean is simply the mean of that individual. In the typical-speaker scenario on the

left, this estimate is close to the actual population-level mean, but in the atypical-speaker scenario, it is far from it. If there is only a single speaker or a handful of speakers in an LSAP study, the probability that the atypical scenario will arise is uncomfortably high. This is the reason for the intuition that a small speaker sample is not sufficient. Of course, the sample will necessarily be small in LSAPs due to the small size of the accessible population.

Quantitative researchers must accept some uncertainty in their estimates. In fact, a large part of statistical practice focuses on quantifying this uncertainty. LSAPs present two particular challenges in this context. First, the uncertainty of population-level values estimated from a small sample of speakers can be very large. This is illustrated by the wide confidence intervals (CI) at the top of the right-hand panel of Fig. 1, based on hierarchical models that account for both token-level and individual-level variation. As a result of this uncertainty, analyses of data from LSAPs using *p*-values as a binary criterion for robustness will rarely yield reliable effects. Second, when the speaker sample is extremely small (e.g. a single individual), it is impossible to reliably estimate variation across individuals and consequently impossible to come up with reliable estimates of uncertainty in population-level parameters. A common but mistaken reaction to this situation is to fall back on a non-hierarchical model, ignoring across-individual variation entirely. The estimates of uncertainty from such non-hierarchical models are unduly low, as illustrated by the second row of CIs in Fig. 1. While the CI from the hierarchical model is wide enough to include the true population mean even in the atypical-speaker scenario, the narrow non-hierarchical CI creates false confidence in a highly misleading estimate.

In section 3, we flesh out one possible alternative in the form of a Bayesian analysis. This approach avoids the issues above by (i) focusing the analyst's attention on graded measures of uncertainty instead of a simple binary decision criterion like the *p*-value; and (ii) approximating across-individual variation using external sources, which allows for the estimation of uncertainty around population-level parameters. In section 4, we suggest further solutions that question the assumptions made at the beginning of this section (e.g. what the statistical population is, or whether uncertainty must be quantified numerically).

3. CASE STUDY: GITKSAN EJECTIVE VOT

We analyse data from a single speaker of Gitksan to estimate the positive VOT associated with Gitksan ejective stops. The data and code for this study are available as part of the online supplementary materials at `osf.io/d7m34`. Isolated Gitksan words were elicited and recorded from a fluent L1 speaker (male, aged 65+) in a quiet room at the University of British Columbia. Word-initial prevocalic [t' c' k^w' q'] and their following vowels were labeled in Praat [10]. VOT values of 120 stop-tokens (4 places of articulation x 5 vowel contexts x 2 words x 3 repetitions) were calculated from the textgrids using a Praat script.

We aim to get a sense of what VOT values are plausible for this language, that is, to perform estimation with uncertainty. As noted above, a hierarchical model is needed to account for potential differences in typicality across speakers. However, there are difficulties in fitting such a model to this data, since it is not possible to estimate across-individual variation (corresponding to σ_{pop} in Fig. 1) from a single speaker. Our solution is to use a Bayesian hierarchical model (fitted with the `brms` package [11]) and exploit our ability to specify priors (see below). The model is essentially the same one that we would use for a larger speaker pool. The outcome variable is VOT, the only fixed effect is the intercept (corresponding to mean VOT) and there are random intercepts by speaker, as shown by the following `lme4`-style [12] formula:

(1) VOT ~ 1 + (1 | speaker).

Priors are distributions that embody prior knowledge or beliefs about parameters, and must be specified as part of a Bayesian analysis. The model returns a *posterior* distribution that results from updating the priors based on new evidence from the data. For the most part, we use *regularising* priors (details in online materials), which provide information about plausible values for parameters, but let the data determine the specific estimates. However, since across-speaker variation cannot be estimated from the data, this parameter is provided to our model as an *informative* prior obtained from three languages where by-speaker VOT figures for ejectives were available: Turkish Kabardian [13] (9 speakers), Cochabamba Quechua [14] (8 speakers) and Witsuwit'en [15] (11 speakers). It is a gamma distribution with a mean of 16.39 ms (shape = 24.53, scale = 0.67) estimated through a separate Bayesian model (details in online materials). This prior tells our model how much speakers tend to deviate from mean ejective VOT in other languages (but *not* what typical VOT values are). This then enables the construction of uncertainty estimates that allow for the possibility that our speaker is atypical.

The estimated population-level VOT for ejectives in Gitksan is summarised in the posterior distribution in Fig. 2. The mean of this distribution is 68 ms, which is the model's best guess for the population-level value, and is identical to the mean VOT for our single speaker (cf. Fig. 1 and the discussion in section 2). The rest of the distribution shows our uncertainty about

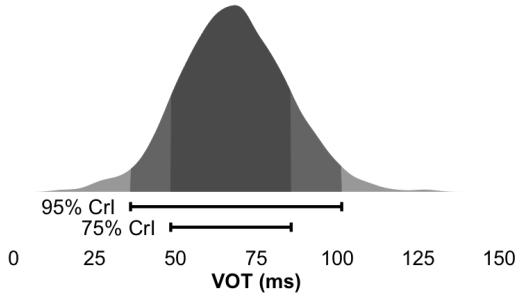

Figure 2: The posterior distribution of average VOT based on a Bayesian mixed model fitted to a speaker of Gitksan with across-speaker variation estimated from other languages. The whiskers show 95% and 75% credible intervals (corresponding to different levels of certainty).

this value. The 95% credible interval (CrI) is [36 ms, 101 ms]: we can be almost certain that the population-level value lies in this range. While null hypothesis significance testing usually limits itself to these two statistics, there is more information in the posterior. For instance, the 75% CrI is [49 ms, 86 ms]: we can be reasonably confident that ejective VOT in Gitksan is between these two values. Looking at the lower end of the distribution, it is extremely unlikely that the population mean is lower than 25 ms, placing it firmly in the long-lag VOT range. This contrasts with languages such as Witsuwit'en, where ejective VOT can be as low as 20 ms for some speakers [15]. The posterior can also be used as a prior for further studies of Gitksan ejective VOT with data from other speakers.

Using data from other languages to allow for hierarchical modelling is a compromise. However, it is better than (i) using a non-hierarchical model with unacceptably overconfident estimates (95% CrI of [62 ms, 75 ms]); or (ii) leaving the data unreported, which would lead to a loss of valuable information.

4. DISCUSSION

Sections 2 and 3 provided an exposition of the key statistical challenge of LSAPs and described a potential solution based on Bayesian modelling. Although estimating a single mean VOT value is simpler than most quantitative problems, the method generalises to more complex scenarios. The key is to be able to estimate across-speaker variation for relevant parameters from other languages. For instance, if the main research question is about differences in ejective VOT between two different places of articulation, one can look at how much this effect tends to vary in other languages.

As noted above, statistics provides a toolbox [5]. The method above is just a single tool. There are many other ways to approach the issue of small accessible populations. For instance, data description, exploration and visualisation are often under-utilised in phonetic studies. In some cases, the data are reduced to a single coefficient (or even just a *p*-value) from a regression model. We encourage analysts to put a heavier emphasis on communicating descriptive information in a format that is easy to digest (e.g. in the form of informative and well-designed graphs [16]). There are also many techniques for data exploration such as principal component analysis [17] and random forests [18] that provide alternative ways to summarise data. Providing a comprehensive but easily digestible summary can be more informative than running a hypothesis test that is doomed to fail or mislead due to a low sample size. While these descriptive methods do not offer measures of uncertainty, it is straightforward to verbally outline the challenges of parameter estimation with data from LSAPs.

Another tool that can be extremely useful in this context is the sharing of data and analysis code [19]. This facilitates both the evaluation of the analysis and helps future cumulative research efforts. We note that in our own attempt to find baseline figures on across-speaker variation in ejectives, we have not yet found a single publication providing raw data alongside the published results.

Finally, the statistical population for a given analysis is not predetermined by convention but depends on the goals and beneficiaries of the research. To give a concrete example, when descriptive work is carried out to facilitate language revitalisation, the community itself may prefer data from a small number of elders. In such a case, the statistical population may just be a single speaker, obviating the need for hierarchical modelling.

With these considerations in mind, how should a reviewer or editor approach the task of evaluating work on LSAPs? To be clear, we do not advocate for a blank cheque for this type of research. Quantitative work on LSAPs must be rigorous and must also show awareness of the dangers of a small speaker pool. The key to success for an analysis of LSAP data is to clearly identify research goals that can be attained with limited data; make the best use of the data through the judicious application of quantitative methods; and highlight any potential limitations of the conclusions. If these criteria are fulfilled, we see no reason to criticise work on LSAPs based solely on the number of speakers.

To summarise, NHST is overused in many scientific contexts, and is far too restrictive for work on LSAPs. Descriptive statistics, exploratory methods, visualisation and Bayesian modelling can all provide valuable alternatives. In addition, while all research requires careful thought about the scientific context and intended beneficiaries, this is even more important in work on LSAPs.

7. REFERENCES

[1] Eberhard, D. M., Simons, G. F., Fennig, C. D. 2022. *Ethnologue: Languages of the World (25th ed.)*. SIL International. http://www.ethnologue.com.

[2] Dunlop, B., Gessner, S., Herbert, T., Parker, A. 2018. *Report on the Status of B.C. First Nations Language (3rd ed.)*. First Peoples' Cultural Council.

[3] Evans, N., Levinson, S. C. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behav. Brain Sci.* 32(5), 429–448.

[4] Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., Majid, A. 2022. Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences* 26(12), 1153–1170.

[5] Gigerenzer, G. 2004. Mindless statistics. *J. Socio.-Econ.* 33(5), 587–-606.

[6] Cumming, G. 2014. The new statistics why and how. *Psychol. Sci.* 25(1), 7–29.

[7] Kruschke, J. K., Liddell, T. M. 2018. The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychon. B. Rev.* 25(1), 178–206.

[8] Yates, D. S., Moore, D. S., Starnes, D. S. 2003. *The Practice of Statistics (2nd ed.)*. Freeman.

[9] Baayen, R. H., Davidson, D. J., Bates, D. M. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59(4), 390–412.

[10] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3.03, retrieved 17 December 2022 from http://www.praat.org/

[11] Bürkner, P. C. 2017. brms: An R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* 80, 1–28. doi: 10.18637/jss.v080.i01.

[12] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67(1), 1–48.

[13] Gordon, M., Applebaum, A. 2006. Phonetic structures of Turkish Kabardian. *J. Int. Phon. Assoc.* 36(2), 159–186.

[14] Gallagher, G., Whang, J. 2014. An acoustic study of trans-vocalic ejective pairs in Cochabamba Quechua. *J. Int. Phon. Assoc.* 44(2), 133–154.

[15] Hargus, S., 2011. *Witsuwit'en Grammar: Phonetics, Phonology, Morphology*. UBC Press.

[16] Healy, K. 2018. *Data Visualization: A Practical Introduction*. Princeton University Press.

[17] Baayen, R. H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press.

[18] Breiman, L. 2001. Random forests. *Machine Learning* 45(1), 5–32.

[19] Foster, E. D., Deardorff, A. 2017. Open Science Framework (OSF). *Journal of the Medical Library Association: JMLA* 105(2), 203.