elpublishing

Language Documentation and Description

ISSN 2756-1224

This article appears in: *Language Documentation and Description, vol 16*. Editor: Peter K. Austin

Swadesh lists are not long enough: Drawing phonological generalizations from limited data

RIKKER DOCKUM & CLAIRE BOWERN

Cite this article: Rikker Dockum & Claire Bowern (2018). Swadesh lists are not long enough: Drawing phonological generalizations from limited data. In Peter K. Austin (ed.) *Language Documentation and Description, vol 16*. London: EL Publishing. pp. 35-54

Link to this article: http://www.elpublishing.org/PID/168

This electronic version first published: August 2019

Swadesh lists are not long enough: Drawing phonological generalizations from limited data

Rikker Dockum and Claire Bowern

Yale University

Abstract

This paper presents the results of experiments on the minimally sufficient wordlist size for drawing phonological generalizations about languages. Given a limited lexicon for an under-documented language, are conclusions that can be drawn from those data representative of the language as a whole? Linguistics necessarily involves generalizing from limited data, as documentation can never completely capture the full complexity of a linguistic system. We performed a series of sampling experiments on 36 Australian languages in the Chirila database (Bowern 2016) with lexicons ranging from 2,000 to 10,000 items. The purpose was to identify the smallest wordlist size to achieve: (1) full phonemic coverage for each language, and (2) accurate phonemic distribution compared to the full dataset. We hypothesize that when these two criteria are met they represent a minimally complete sample of a language for basic phonological typology. The results show coverage is consistently achieved at an average lexicon size of approximately 400 items, regardless of the original lexicon size sampled from. These results hold broad significance, given the predominance of word lists smaller than 400 items. For fieldwork, this study also provides a guideline for designing documentation tasks in the face of limited time and resources. These results also help to make empirically-grounded decisions about which datasets are suitable for use for which research tasks.

Dockum, Rikker and Claire Bowern. 2018. Swadesh lists are not long enough: Drawing phonological generalizations from limited data. *Language Documentation and Description* 16, 35-54.

1. Introduction

Linguistics requires generalizing from limited data. Documentation, no matter how detailed, cannot capture the full complexity of a language. At the same time, we have at our fingertips ready access to more linguistic data than at any time in the past. This gives rise to two questions: given limited resources, how much data is minimally sufficient for a given research question? Secondly, as more and more data becomes available, what tasks is a given dataset suitable for?

Here we present the results of experiments that address these questions with respect to phonology in Australian languages.[1] Drawing from the Chirila database (Bowern 2016), lexicons with more than 2,000 entries were selected to represent 'complete' natural languages. These 36 lexicons ranged between 2,000 and 10,000 entries. A series of sampling experiments were performed on each test language to determine the minimal wordlist size necessary to satisfy two criteria: (1) full phonemic coverage for the language, and (2) faithful phonemic distribution to the full dataset. We hypothesize that when these two criteria are met, the sample represents a phonologically complete subset of the lexicon, suitable for basic phonological analysis and generalization. We conversely hypothesize that failing to meet the criteria indicates an unsuitable dataset (though such a dataset may be suitable for other research purposes). From another perspective, this study seeks to identify the point of diminishing returns from having more data. This is conceptually related to 'stopping rules' in mathematics (Hill 2009), and their application in clinical trials and surveys, where they are used to determine when additional data will no longer significantly change the results of a study (Meinert 2012). However, analyses performed with insufficient data are probably the more common problem in linguistics.

The use of basic vocabulary wordlists is commonplace in language documentation fieldwork. The most well-known is perhaps the Swadesh list, with versions of varying sizes published over the years: 215 (Swadesh 1952), 200 (Swadesh 1955), and 100 (Swadesh 1972). Swadesh designed his lists for glottochronology and lexicostatistics, but standard wordlists are common in language documentation generally. Lists designed for automated comparison tend to fall around the same size range as Swadesh's or smaller: for example, 241 items (Cross 1964), 200 (Matisoff 1978), 128 (Lohr 2000), 100 (Wilson 1969), or 40 (McMahon & McMahon 2005; Wichmann,

[1] Thanks to Ethan Campbell-Taylor, who contributed to the early stages of this project. This research was funded by NSF grants BCS-0844550 and BCS-1423711.

Holman & Brown 2016). Lists focused on areal or general language documentation tend to be larger: 436 items in the SIL Southeast Asia Wordlist (SIL MSEAG 2002), 1408 in the Comparative Bantu Wordlist (Guthrie 1967–71), 1707 in the SIL Comparative African Wordlist (Snider & Roberts 2004), and 1310 in the Intercontinental Dictionary Series template (Borin, Comrie & Saxena 2013).

Once gathered, this lexical data can be used for myriad analyses, from manual comparison to large scale statistical studies. One of the oldest and most common uses is in comparative historical linguistics for language classification. The question of wordlist size has previously been tested in the domain of statistical methods for detecting historical connectedness between languages. Ringe (1992: 55–64) found no added statistical benefit for quantitative historical comparison with a Swadesh 200 wordlist over a Swadesh 100. His experiment took English and Latin as the test case, matching only on initial consonants. Ringe concluded that the numbers of words involved in detecting such relationships is so small that wordlist changes can easily affect the percentage of matches, and mostly serve to increase noise. However, Kessler (2001: 65–66) points out that in a binomial test like that used by Ringe (1992), as the number of observations increases, the threshold for statistical significance decreases proportionally. Therefore, matches that fell below significance with 100 words may become statistically significant with a 200-item list. Kessler goes on to use χ^2 tests with 28 languages pairs to show that, when randomly sampling 50 and 100 items from the Swadesh 200, the larger sample improves p values for related languages with no artificial boost to the unrelated ones. The larger sample identified statistically significant connections between five language pairs, versus only two pairs in the smaller sample (Kessler 2001: 67–69). Thus, there is both theoretical and empirical cause to expect a larger wordlist to achieve a better result. However, none of these prior studies explicitly addressed the question of diminishing returns.

2. Methods and Data

2.1 Data

Data for this project come from the Chirila database of Australian languages (Bowern 2016), a comparative lexical database containing material from most of the languages of Australia. A subset of the data — from 165 languages — has been phonemically normalized, and can thus be reliably used for inferring phonological patterns in those languages. However, in the full Chirila dataset, the number of items in those lists varies extensively,

from under 100 to over 10,000. Preliminary data were examined in Gasser & Bowern (2014), but that work did not control for the overall wordlist length in drawing conclusions. Instead, they imposed an arbitrary cutoff of 400 items. Figure 1 shows the wordlist holdings in the full Chirila database, sorted by length, as of 1st January 2018.

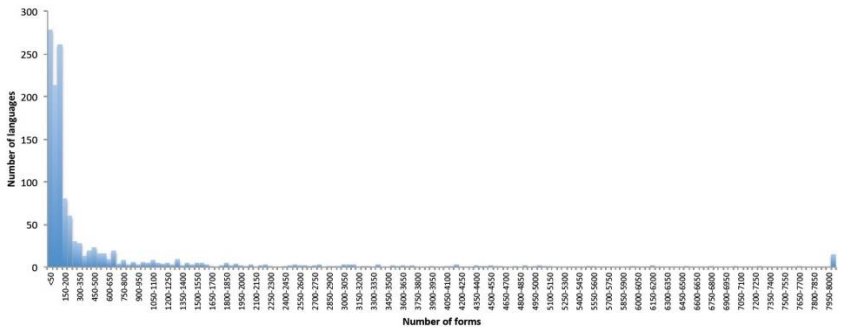

Figure 1 Chirila wordlist holdings, sorted by length.

The input data from the full Chirila dataset is less clean than one would wish for a study of this type. As discussed in more detail in Bowern (2016), sources for the Chirila database vary extensively in data sanity — some wordlists, for example, were retyped and information categorized (and are thus easy to adapt for different analytical projects), while others were imported directly from digital originals, where the source file did not impose rigid data structures. For example, in at least 50 cases, the only digital wordlist data that we had access to were wordlists or dictionaries in Microsoft Word. In other cases, the materials were organized in a database format (such as the Toolbox backslash-coded text file format), but the information was not consistently categorized. Some wordlists have material in the head lexical field other than the phonological form of the headword, such as conjugation information, dialect annotations, or in some cases, disambiguation of glosses. This material is in the process of being moved out of the lexical head fields but must be done manually.[2] These additional annotations, however, potentially introduce

[2] It is possible, of course, to 'scrub' a data field automatically, for example by assuming that only the first word in the field is relevant information. While this allows us to have clean data, it loses the other information in the field, which may be needed for other projects.

new 'phonemes' into the data, particularly if they contain English words. Furthermore, lexicons treat loanwords differently, some adapting them to the phonology of the language, others representing them as English or Kriol words. The 165 phonemicized lists represent the cleanest subset of the data. As an extra precaution, we introduce the concept of a 'marginal' phoneme to capture items which are present in our dataset but are likely to represent noise. This is discussed in Section 4 below.

The 165 phonemically normalized lexicons were sorted by lexicon size, and all with 2,000 or more items, the size large enough for our tests, were selected for inclusion. This gave a working dataset of lexicons from 36 languages.[3] The largest had more than 10,000 items, providing a wide range of sizes and allowing us to test sensitivity to the size of the full lexicon. The languages, wordlist size, and source information are provided in the Appendix.

2.2 Sampling

For each of the 36 lexicons, we initially sampled at six sizes per lexicon, with n = 50, 100, 250, 500, 1,000, or 2,000. These pilot results showed that a regular threshold for achieving our criteria appeared to fall between 250 and 500 items. Thus, we subsequently added more fine-grained sample sizes in that range, sampling at 300, 350, 400, and 450 items. For each sample size, we generated samples using four methods: taking the first n items of each lexicon, the last n items, randomly sampling n items, and randomly sampling an item from n equally sized bins. Thus, each combination of a sample size and a sampling method constitutes a test condition. The first n and last n methods were included with the expectation that they would perform poorly, in order to establish a baseline for comparison. We also conducted pilot tests on lists limited by part of speech or semantic field; this is discussed further in Section 4.

We created a script in Python (version 3.4.3; Python Software Foundation 2018) to generate the subsets. For each full lexicon and its subsets, the script then performs phonemic segmentation, compiles a phoneme inventory, and calculates phoneme frequency. The script then compares the properties of

[3] One language, Wubuy, with a list of 4600 items, was excluded, leaving 36 languages from an original set of 37. Wubuy was excluded due to the extensive use of archiphonemes in headwords, which makes it not comparable to the other phonemicized wordlists, which have a clearer mapping between underlying phonological categories and 'surface' representation.

each subset to that of the full lexicon. Finally, we ran this process 1,000 times and calculated final scores for each test condition. Data was further processed and visualized in R (R Core Team 2018).

2.3 Scoring

We scored each subset test condition on two criteria: (1) phonemic coverage, and (2) fidelity of phonemic distribution. For the first criterion, phonemic coverage is scored for each run as a value of either 0 and 1, where a 1 means every phoneme in the full lexicon was attested in the sample. After discounting marginal phonemes, the overall coverage score for each test condition was calculated as a continuous value between 0.0 and 1.0, representing the number of times out of 1,000 runs that every phoneme was observed in individual runs.

We scored our second criterion, phoneme distribution similarity, using the residual sum of squares (RSS). RSS is calculated by squaring the difference between the frequency of each phoneme in a given subset test condition and the frequency of that phoneme in the full lexicon, and then summing all the squares. A score closer to zero represents higher similarity. We calculated the frequency of each phoneme as the proportion of lexical items that a phoneme appears in (where a phoneme is counted at most once per entry).

3. Results

The following sections present the results of the simulations. For ease of interpretation, visualization of our results requires changes to the height of the y-axis in some figures. As this detail can easily become confusing, the y-axis height is included in some figure titles. Figures presented together under a single figure number always have the same y-axis height. The first of these is Figure 2. Using Figure 2a as an example, this set of four graphs represents the results for all languages under the first n items sampling method. Each bar shows a different sample size, and the y-axis gives the sum RSS for all 36 languages at that sample size. The fill color of each bar illustrates phoneme coverage, where each of the 36 languages contributes some color to the bar according to its phoneme coverage score. In this case, it happened that for every language tested phoneme coverage was uniformly 0 or 1 in all 1,000 runs, thus giving the appearance of a binary result in the non-random sampling methods, despite being calculated as continuous values.

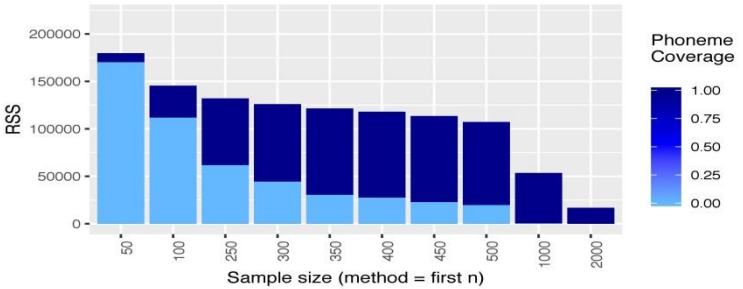

Figure 2a - Sampling first *n* items

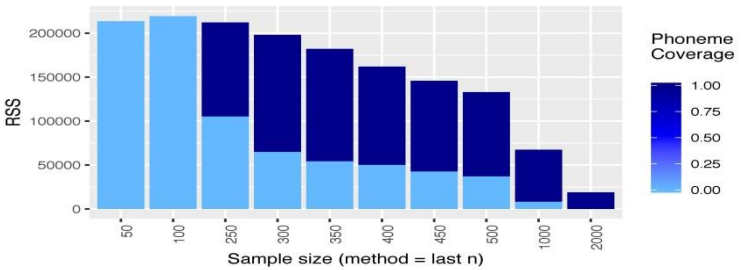

Figure 2b - Sampling last *n* items

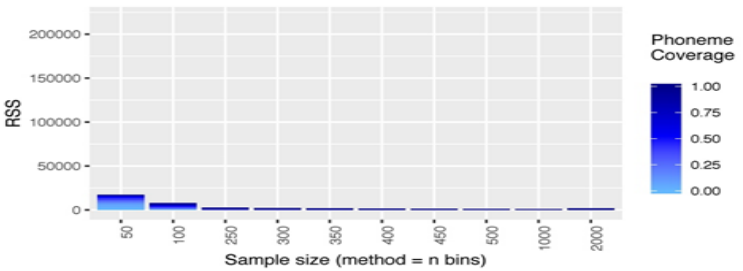

Figure 2c - Sampling from *n* bins

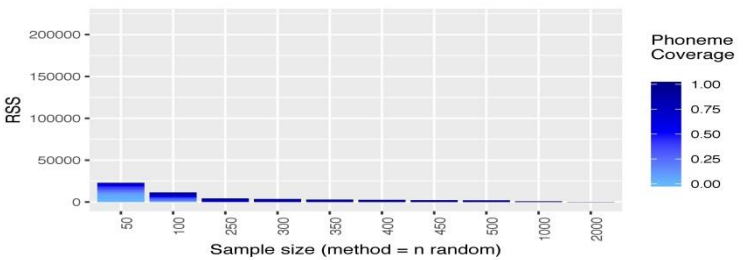

Figure 2d - Sampling *n* random items

Figure 2: RSS scores summed for all languages, sorted by sample size

3.1 Non-random sampling methods

Of the four sampling methods, the two non-random ones (first *n* and last *n*) performed very poorly, as expected. These results are seen in Figures 2a and 2b. Two things are notable: first, that their RSS scores are much worse, and second, that the subset size must be extremely large before even phoneme coverage is reliably achieved for most languages, as compared to the two random sampling methods. This is, of course, expected; samples that come from the start or end of an alphabetically sorted dictionary are very likely to be skewed towards certain phonemes.

Even ignoring RSS performance, with first *n* and last *n* there is no point below $n = 1000$ where phonemic coverage is reliably achieved. The first *n* and last *n* sampling methods are inadequate for virtually every sampling size, except perhaps $n = 2000$. And of course, since the lower bound of lexicon size of the 36-lexicon sample is 2,000 items, this 'subset' represents virtually the entire lexicon for many of the languages sampled, so this sample size is not very informative. Clearly, any sample that only draws from the beginning or end of the lexicon utterly fails to be representative of the lexicon as a whole. Even the smallest subset ($n = 50$) using the random sampling methods, while still the worst performing from its method, is more than an order of magnitude better than the non-random methods. If the gap were not so enormous between the RSS scores of the random and non-random sampling methods, we might want to examine more closely whether particular languages are skewing these distributions more than others. Given these results, however, this seems unnecessary.

3.2 Random sampling methods

We turn now to methods which randomly sample points in the wordlists. Results are presented in Figures 2(c), 2(d), and Figure 3.

The graphs in Figure 3(a) and (b) present the same data as in 2(c) and (d), except that the *y*-axis has been scaled down by a factor of 10, from 200,000 to 20,000. This better illustrates the performance of each sample size on both metrics. Phonemic coverage remains poor for sample sizes 50 and 100, but improves starting at at $n = 250$; for phoneme distribution, there is a steep decline in RSS scores (indicating better fidelity to the full lexicon) that also starts at $n = 250$. Past 250 items, RSS performance begins to plateau. Consider the results in Figure 4.

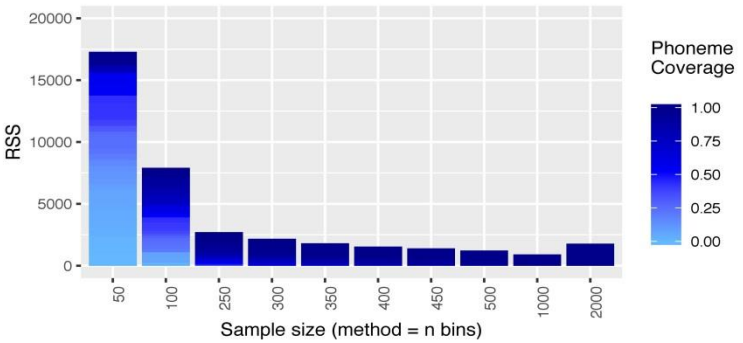

Figure 3a - Sampling from *n* bins

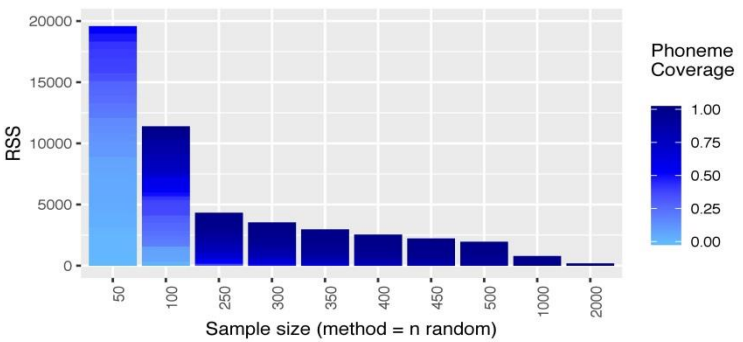

Figure 3b - Sampling *n* random items

Figure 3: RSS scores summed for all languages, sorted by sample size (y axis rescaled).

In Figure 4, each bar represents a language. Since we are now dealing with results for individual languages, we again have scaled down the y-axis, from 20,000 in Figure 3 to 1,000 in Figure 4. In each graph, the bars are ordered from left to right by the size of the full lexicon. This perspective shows that our scoring criteria are not sensitive to the source lexicon size. At each sampling size, performance is roughly the same across all 36 languages.[4]

The six sample sizes in Figure 4 are the original six sizes that we sampled in pilot tests. While the sharpest improvements in RSS score occur after the smallest two sample sizes, there are also improvements in scores across all languages when going from 250 to 500. Past 500 items, scores continue to improve for some languages, but plateau for others. In order to identify the threshold of diminishing returns more accurately, we also sampled at 300, 350, 400, and 450 items, to drill down into the space between the 250 and 500 items. Results for these samples are shown in Figure 5.

In the set of graphs in Figure 5, once again the y-axis is scaled down, now to just 150. This makes it easier to see that while we are making gains both in RSS score and phonemic coverage as we move up in sample sizes.

Taking the average of 1,000 runs for each language, we found that only once we sample at 400 items do we consistently meet both of our criteria. This appears to be a common threshold across all 36 languages. This threshold is insensitive to the full lexicon size. Smaller samples regularly fail to observe some phonemes or else exhibit skewed phoneme distributions.

[4] It is unclear why, in 4(f), with $n = 2000$, the RSS scores are worse than $n = 1000$. It is likely a sampling artifact.

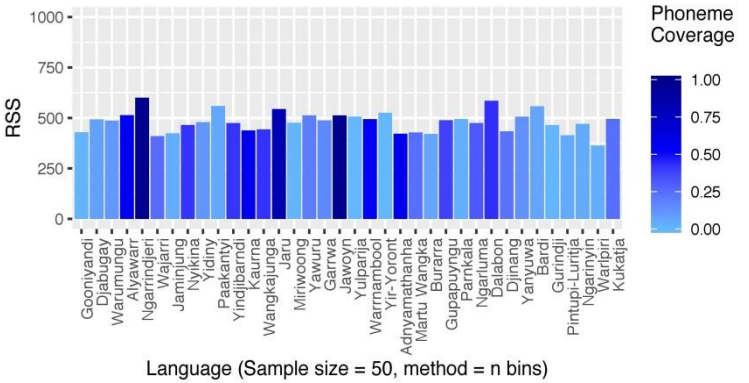

Figure 4a - n = 50

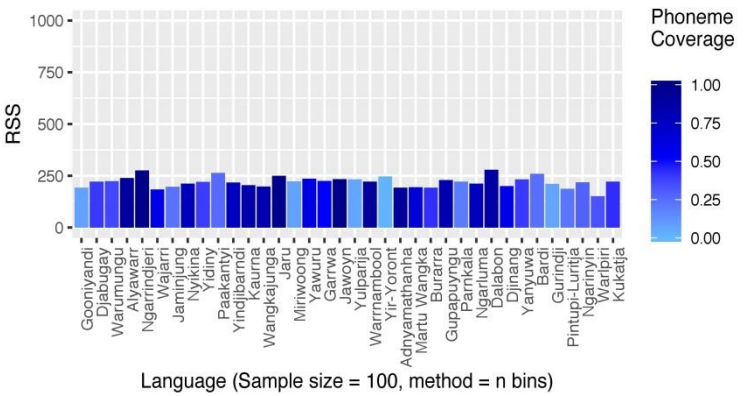

Figure 4b - n = 100

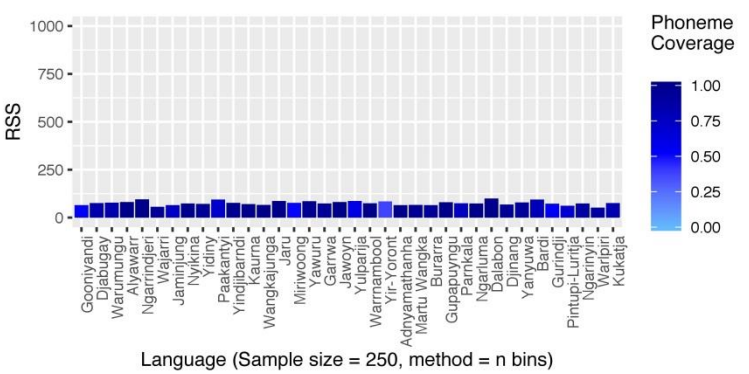

Figure 4c - n = 250

Figure 4(i): RSS score by language with y-axis scaled to 1000 (method = n bins)

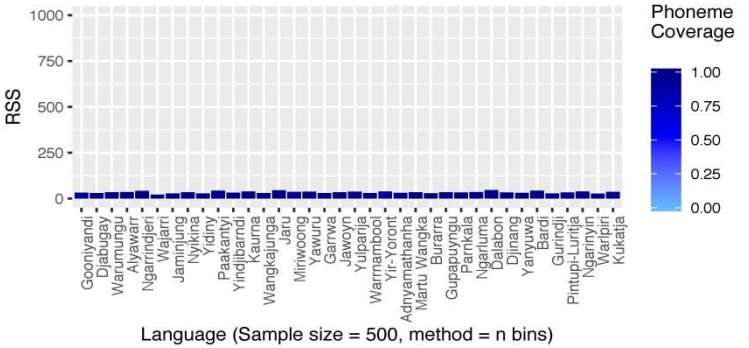

Figure 4d - n = 500

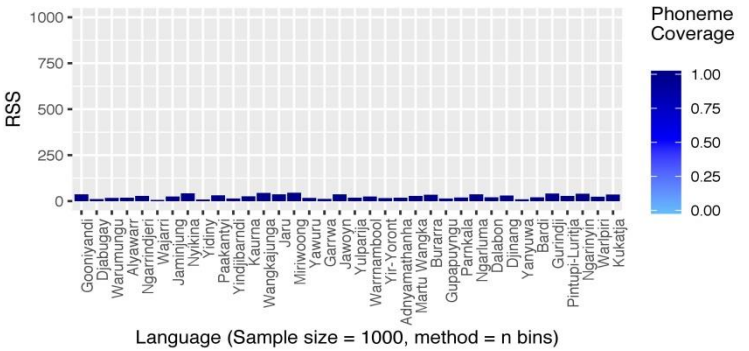

Figure 4e - n = 1000

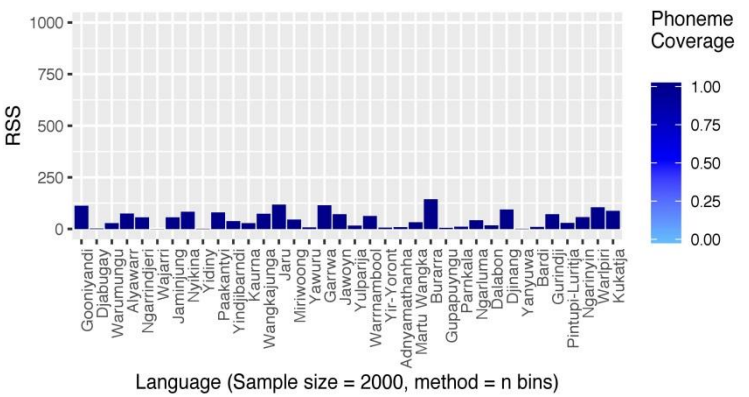

Figure 4f - n = 2000

Figure 4(ii): RSS score by language with y-axis scaled to 1000 (method = n bins)

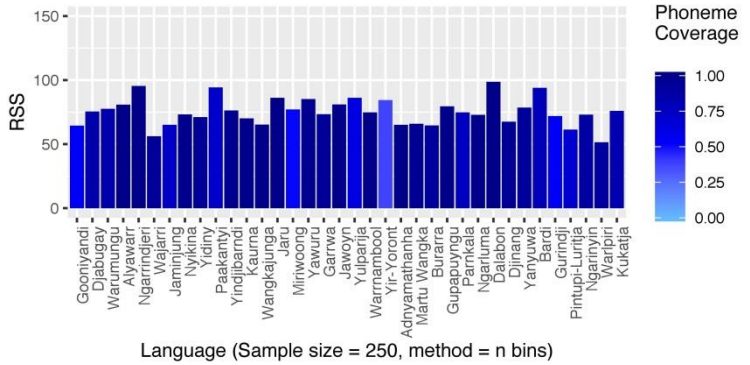

Figure 5a - n = 250

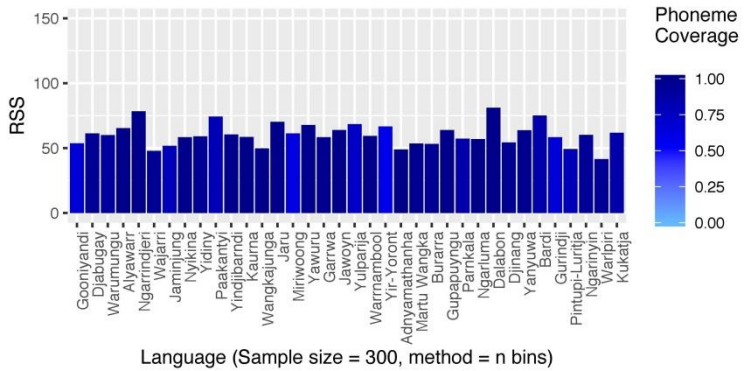

Figure 5b - n = 300

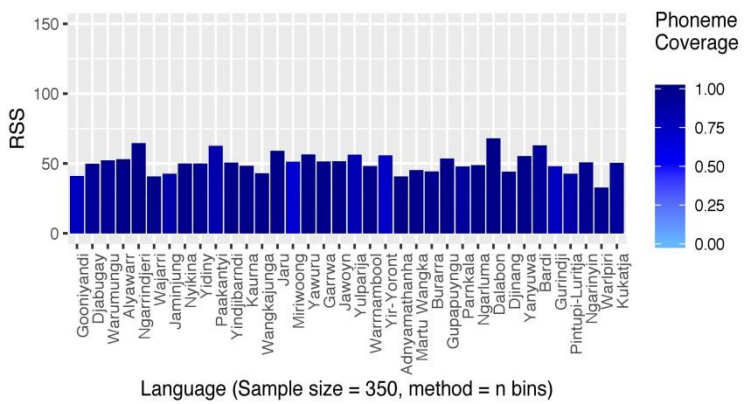

Figure 5c - n = 350

Figure 5(i): RSS score by language with y-axis scaled to 150 (method = n bins)

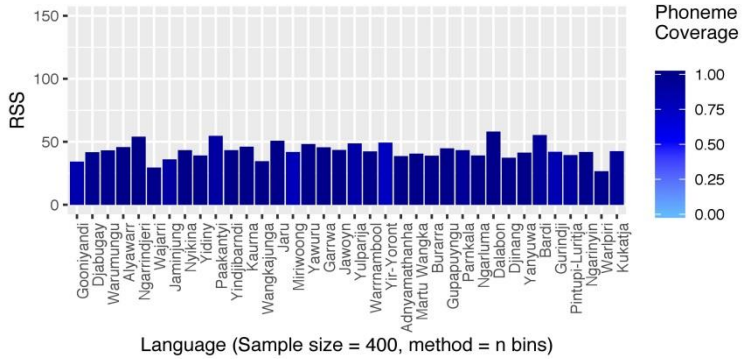

Figure 5d - n = 400

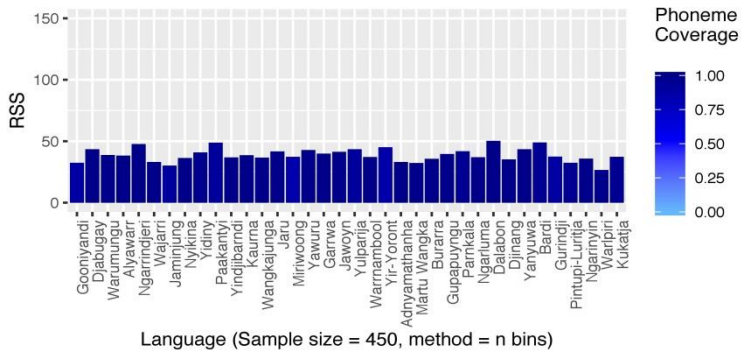

Figure 5e - n = 450

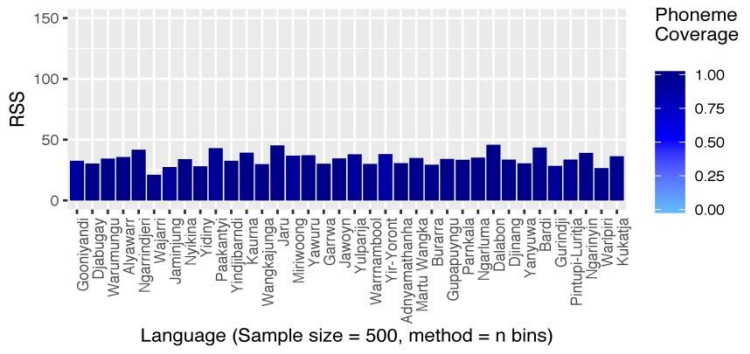

Figure 5f - n = 500

Figure 5(ii): RSS score by language with y-axis scaled to 150 (method = n bins)

4. Discussion

4.1 'Marginal' phonemes and inventory size

As mentioned in earlier sections, we dealt with likely noise in the dataset by classifying low-frequency phonemes as 'marginal'. In calculating phoneme coverage, marginal phonemes did not count for or against the score. Marginal phonemes might be present in the data for a variety of reasons, including data artifacts or loanwords. However, quantifying what constitutes a marginal phoneme is not trivial. It might be defined as either a fixed value (ideally if all wordlists were of similar length) or proportional to the size of the wordlist. For this study, we classified phonemes as marginal if they occurred in fewer than 0.5% of items in the lexicon. For a wordlist of 10,000 items, this means a phoneme that appears fewer than 50 times is marginal. By setting the bar for marginality relatively low, some genuine phonemes may be discounted, but it ensures that whenever we observe failure to achieve coverage, it should represent a genuine deficiency in the subset.

These results apply to languages with 'average' phoneme inventory sizes. As per Gasser & Bowern (2014), the languages in the survey vary in inventory size from 16 segments to 38, with the median being 25.[5] We note that these results are robust to variation in inventory size at this level; i.e., while we see some variation in performance (as evidenced by the range of RSS scores in the results), those results are close to uniform at the 250–500 level, despite differences in inventory sizes. While we would expect that languages with much larger phoneme inventories would require longer wordlists in order to show adequate coverage by the metrics used here, our result is applicable to the majority of languages of the world.

4.2 Implications for analysis

These findings hold significance for linguists and linguistics beyond our narrow question. At the core of this study is the issue of the reliability of studies that examine various properties of languages across many wordlists of very limited size. Working with sufficiently representative data is imperative for being able to draw reliable conclusions across the languages of the world.

---

[5] Per Maddieson's survey in WALS (wals.info), the 'more typical consonant inventory size is in the low twenties, with the mean for the 562 languages being 22.7, the modal value 22 and the median 21.' The typical number of vowels is 5-6; so we might say that the average number of phonemes is somewhere around 27, which is about the same as what we find in our dataset for Australia.

Even if the properties studied here — phoneme coverage and phoneme distribution — are not considered particularly interesting ones to test, they should be taken as an intentionally 'low bar' for success. They are entry-level data on languages. If the wordlists that large studies are based around fail on these simple and accessible metrics, then the existence of more subtle analytical questions only serves to strengthen the case for a necessary higher threshold for minimal wordlist size.

Our 400 item threshold is not intended as an argument that we never need more than 400 items, nor that we should discard smaller wordlists. Studies such as Ringe (1992) and Kessler (2001) use a specific research question as a test case, that of detecting connectedness via shared onset consonants, in order to make claims about the suitability of different wordlist sizes. We, on the other hand, have taken the approach of selecting two properties that are readily observable in a wordlist, and determining just how much data is needed to find those same properties in our sample. This study demonstrates that when we deal with wordlists below this 400-item threshold, we are likely to be incorrectly representing basic facts about the phonology of the language. That is a problem, particularly when those phonological analyses are then used as input to larger claims.

The points made here also hold for wordlists that are restricted by some syntactic or semantic property, such as part of speech or semantic field. In pilot tests examining subsets consisting of, e.g. only verbs or nouns, or only flora and fauna terms, coverage in our sample was poor enough that we did not pursue this line further. A more thorough study would require significant additional tagging in the lexical database, but we argue that pilot tests, in light of our findings with random sampling, sufficiently confirm that semantically or syntactically restricted wordlists will not be representative of general phonological properties either.

In the domain of quantitative historical linguistics, for which so many of these wordlists were designed, even if Ringe (1992) is correct in concluding that a Swadesh 200 list is no better than a Swadesh 100 list for calculating remote language relatedness, our findings suggest that neither may be truly sufficient. Either list size is likely to have missed certain phonemes entirely. To give one example, a Swadesh 100 list for English and German contains no instances of initial /j/:/j/ correspondences at all (Kessler 2001, 66). However, it is also important to note that the results of both Ringe and Kessler may rely on the notion of 'core' vocabulary or conceptual universality in the items that make up Swadesh lists. A future direction planned for this project is to test pseudo-Swadesh lists by extracting them from larger lexicons in Chirila, in order to see how closely their phonological properties compare to our comparably sized random samples. Work by Blasi et al. (2016) suggests that basic vocabulary wordlists contain a significantly non-random distribution of phonemes in a language.

4.3 Implications for language documentation

In language documentation, we have a great need for better coverage of the thousands of undocumented and under-documented languages, even as many are being lost. One of our goals for this paper is to inform the work of field linguists. In planning a language survey project, for instance, where depth will have to make concessions in order to achieve breadth, researchers should have a scientifically principled convention of what constitutes a minimally sufficient dataset.

Given limited time, how can field linguists serve their immediate research aims, while also maximizing the future potential for their work in larger projects? The answer of course depends on the purpose of the fieldwork. For comparative purposes, an optimal wordlist would be one that combines basic vocabulary with other vocabulary that is more likely to show loans, as well as a mixture of parts of speech, avoiding only high frequency items, etc. Perhaps the best recommendation is to gather the Swadesh 200 list (to allow for comparison with other languages with those words), and then another 200–300 items drawn from such categories as: flora and fauna, local material culture, or high frequency vocabulary that is not on the Swadesh list, to balance out the list. There is no independent empirical reason to prefer the Swadesh list over other similarly sized basic vocabulary lists, however. One of any number of standard lists or regionally tailored lists should also prove sufficient.

A field-wide conventional minimum wordlist size of at least 400 items per site is one area where we would see immediate gains in the results that come out of survey-style documentation. This will have downstream benefits as those data make their way into databases and corpora. And obviously, more than 400 items still remains ideal.

5. Conclusions

In this ‘Big Data’ era, linguistic datasets and databases of all kinds are proliferating, and quantitative linguistic work is becoming easier and more common. At the same time, ‘Big’ means different things in different areas of linguistics. A corpus of a million words of natural language is relatively small, but a lexical database with a million entries is among the largest that currently exist. Databases are often aggregated from disparate sources and documentation traditions. Breadth and depth of coverage are in constant competition. Large scale databases are regularly used to study questions that the data were not originally gathered to answer, and while this is one of their key features, convenient data may not always be data that is well-suited to the problem at hand. Work that aims to generalize across large sets of languages comprising many small datasets must be informed about the limitations and assumptions that come with available data.

We certainly would not want to restrict ourselves to quantitative work only on languages with at least a 10,000 word dictionary, as we would miss important generalizations from the vast majority of languages that do not have one that large. Equally, though, if we want to make good analyses, it is important to identify the cutoff for less well-resourced languages. Our findings suggest that Swadesh lists are not enough for many purposes. Moreover, we argue that a movement toward more explicit metascientific analysis and discussion of the match between dataset and research question should become the norm in the linguistics literature. This type of methodological transparency goes hand in hand with arguments about open datasets and improved data citation standards (Berez-Kroeker et al. 2017). Measures like these contribute to resolving the replication crisis (Vanpaemel et al. 2015) that affects the social sciences and natural sciences alike.

References

Berez-Kroeker, Andrea L., Lauren Gawne, Smythe Kung Susan, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chellih, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice & Anthony C. Woodbury. 2017. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1), 1–18. doi:10.1515/ling-2017-0032.

Blasi, Damián E, Søren Wichmann, Harald Hammarström, Peter F Stadler & Morten H. Christiansen. 2016. Sound–meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences* 113(39), 10818–10823.

Borin, Lars, Bernard Comrie & Anju Saxena. 2013. The intercontinental dictionary series - a rich and principled database for language comparison. In Lars Borin and Anju Saxena (eds.) *Approaches to Measuring Linguistic Differences*, 285–302. Berlin: De Gruyter Mouton.

Bowern, Claire. 2016. Chirila: Contemporary and Historical Resources for the Indigenous Languages of Australia. *Language Documentation & Conservation* 10, 1–44. http://hdl.handle.net/10125/24685.

Cross, Ephraim. 1964. Lexicostatistics has not yet attained the status of a science. *Proceedings of the International Congress of Linguists*, 480-489.

Gasser, Emily & Claire Bowern. 2014. Revisiting phonological generalizations in Australian languages. *Proceedings of the 2013 Annual Meeting on Phonology*. doi:10.3765/amp.v1i1.17.

Guthrie, Malcolm. 1967–71. *Comparative Bantu: an introduction to the comparative linguistics and prehistory of the Bantu languages*. 4 Vols. Amersham: Gregg International Publishers.

Hill, Theodore P. 2009. Knowing when to stop: How to gamble if you must. The mathematics of optimal stopping. *American Scientist* 97(2), 126–133. http://www.jstor.org/stable/27859299.

Kessler, Brett. 2001. *The Significance of Word Lists*. Stanford: CSLI Publications.

Lohr, Marisa. 2000. New approaches to lexicostatistics and glottochronology. In Colin Renfrew, April McMahon & Larry Trask (eds.) *Time Depth in Historical Linguistics*, 209–222. Cambridge: McDonald Institute for Archaeological Research.

Matisoff, James A. 1978. *Variational Semantics in Tibeto-Burman: Organic Approach to Linguistic Comparison*. Oxford: Institute for the Study of Human Issues.

McMahon, April & Robert McMahon. 2005. *Language Classification by Numbers*. Oxford: Oxford University Press.

Meinert, C. L. 2012. *Clinical Trials Handbook: Design and Conduct.* London: John Wiley & Sons. doi:10.1002/9781118422878.ch126.

Python Software Foundation. 2018. *Python Language Reference, Version 3.4.3*. http://www.python.org.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.

Ringe, Donald A. 1992. On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society* 82(1), 1–110.

SIL MSEAG. 2002. Southeast Asia 436 Word List.

Snider, Keith & James Roberts. 2004. SIL comparative African wordlist (SILCAWL).' *Journal of West African Languages* 31(2), 73–122.

Swadesh, Morris. 1952. Lexico-statistic dating of prehistoric ethnic contacts: With special reference to North American Indians and Eskimos. *Proceedings of the American Philosophical Society* 96(4), 452–463.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21(2), 121–137.

Swadesh, Morris. 1972. What is glottochronology? In *The Origin and Diversification of Language*, 271–284. London: Routledge.

Vanpaemel, W., M. Vermorgen, L. Deriemaecker & G Storms. 2015. Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra* 1(1). doi:10.1525/collabra.13.

Wichmann, Søren, Eric W. Holman & Cecil H. Brown. 2016. The ASJP Database (Version 17). http://asjp.clld.org/.

Wilson, Darryl. 1969. The Binandere language family. *Occasional Papers a.18*, 65–86. Canberra: Pacific Linguistics.

Appendix: Languages used in this study

Language	Wordlist Size	Language	Wordlist Size
Adnyamathanha	2827	Ngarinyin	7586
Alyawarr	2043	Ngarluma	3499
Bardi	4795	Ngarrindjeri	2087
Burarra	3153	Nyikina	2153
Dalabon	3553	Paakantyi	2187
Djabugay	2015	Parnkala	3434
Djinang	3557	Pintupi-Luritja	6034
Garrwa	2626	Wajarri	2115
Gooniyandi	2013	Wangkajunga	2386
Gupapuyngu	3375	Warlpiri	9193
Gurindji	4951	Warrnambool	2779
Jaminjung	2143	Warumungu	2025
Jaru	2407	Yanyuwa	4254
Jawoyn	2724	Yawuru	2561
Kaurna	2357	Yidiny	2172
Kukatja	10139	Yindjibarndi	2324
Martu Wangka	3012	Yir-Yoront	2823
Miriwoong	2451	Yulparija	2761