2025. Proc Ling Soc Amer 10(1). 5913. https://doi.org/10.3765/plsa.v10i1.59013

Minimal wordlist size for a phonological profile: New evidence from Kra-Dai languages

Kanyarin Boonkongchuen, Rikker Dockum*

Abstract. This study replicates, and extends to Kra-Dai languages, earlier work on minimal wordlist size needed to make a phonological profile of a language. Previous work on Australian languages recommended approximately 400 randomly sampled words to comprise a minimally complete profile in terms of reliably capturing every phoneme, and with accurate distribution. We survey 55 Kra-Dai languages to show that a longer minimal list is necessary, which we attribute to typological differences like larger consonant/vowel inventories. Given the widespread use of short wordlists in fieldwork, these results hold significance for designing language documentation surveys as well as projects that use legacy wordlist data.

Keywords. Phonology; Kra-Dai; Language Documentation; Swadesh Lists

1. Introduction. Much language documentation, especially legacy documentation work, takes the form of surveys of wordlists. These wordlists are often short because they use concepts list like the Swadesh 100 (Swadesh 1955) or 207 (Swadesh 1952) wordlists. And although linguistics as a field is moving towards stronger norms of participatory, community-focused documentation, we still have a large number of limited lexicons available for understudied languages as legacy data. This can be seen in large databases such as the Automated Similarity Judgement Program (ASJP) database (Holman et al. 2008; Wichmann et al. 2022) or the Rosetta Project (Good & Hendryx-Parker 2006). They both include a large number of languages within their databases, but the wordlist for most languages, especially understudied ones, remains fairly short.

How do we decide whether a wordlist is minimally complete? If we want to be able to use legacy wordlists as data in new studies, we want them to be a good representation of the language we are studying. Baird et al. (2022) called this specific problem the 'Bird-Himmelmann problem'. This is derived from when Steven Bird in the Resource Network for Linguistic Diversity mailing list questioned whether any documentation work for an understudied language had met Himmelmann's definition for documentation, which was 'to provide a comprehensive record of the linguistic practices characteristic of a given speech community.' It has become increasingly important to quantify this question and judge the completeness of a wordlist because fieldwork can be an expensive and time-consuming task, and both money and time are finite resources. However, many studies still rely on data that was collected previously by someone else. This highlights the value and the importance of designing language documentation tasks to consider future research applications, and try their best to compile wordlists that meet a definition of minimally complete. This helps ensure that any conclusions that are drawn from the data won't be undermined by limitations of the initial data. This is also why the idea of quantifying a *minimally complete* wordlist is useful and necessary.

One example of why it is desirable to have quality wordlists can be seen in the World Atlas of Language Structures (WALS, Dryer & Haspelmath 2013). WALS is a database that stores the different phonological, grammatical, and lexical properties of each language. They draw from

* Thank you to Prof. Benjamin R. Mitchell for advice on the experimental design. Authors: Kanyarin Boonkongchuen, Swarthmore College (kanyarinboonkongchuen@gmail.com) & Rikker Dockum, Swarthmore College.

already published descriptive material such as reference grammars to compile the different linguistic properties that the WALS database keeps track of. This means that they rely on the original fieldwork studies to be accurate. WALS is a very useful resource for linguists. Because it has data from so many languages, for example, it can tell you if a language has a small, average, or large inventory. This can be very useful information for many linguists and it relies on the accuracy of the initial data so that we can create a fair comparison. Language documentation and fieldwork play a very important part in linguists' work, so questions such as those posed by the Bird-Himmelmann problem should be considered seriously.

This paper extends the work done by Dockum & Bowern (2019), which showed that for 36 Australian languages in the Chirila database (Bowern 2016) with lexicons ranging from 2,000 to 10,000 items, approximately 400 randomly sampled words are needed to able to reliably extract from any given wordlist a minimally complete phonological profile of a language. They defined a minimally complete phonological profile as a lexicon with (1) full coverage of every phoneme in that language, and (2) phonemic distribution that is statistically comparable to that of the full lexicon.

A limitation of Dockum & Bowern (2019) is that only one region of the world was studied, and many of the languages in the study have a similar typological profile. This study replicates the findings of their study, expands upon the methodology, and extends the results to languages in the Kra-Dai language family. One reason that the previous conclusions might not generalize to all languages is that Australian languages on average have been shown to have a relatively small phonemic inventory compared with all the world languages even though they exhibit more variation than is often been assumed (Gasser & Bowern 2013). The World Atlas of Language Structures (WALS, Dryer & Haspelmath 2013) categorizes the world's sound systems based on average numbers of sounds. Australian Aboriginal languages tend to have smaller inventories of both consonants and vowels. For example, a language such as Garrwa which is reported to have 4 vowels would be categorized by WALS as having a 'small vowel inventory'. And a language such as Djabugay which is reported to have 13 consonants would be categorized by WALS as having a 'small consonant inventory'.

Therefore, it was unclear whether the results from Dockum & Bowern (2019) applied more broadly around the world.

2. Background. Currently, there has not been a large amount of research on minimal wordlists. But that is slowly changing as our tools for aggregating and comparing data have improved. However, there is still a limited amount of resources to spend on fieldwork. As previously mentioned, there is the Dockum & Bowern (2019) study that this paper is building upon. Similar work has also been done by Baird et al. (2022) which was briefly mentioned in the introduction for introducing the idea of the Bird-Himmelmann problem.

Baird et al. (2022) used translations of the story 'North Wind and the Sun' in 158 different languages to investigate at which point every phoneme in that language appears at least once. The authors themselves described this as a very low bar but were surprised to find that some of the languages in their study still failed to meet this standard. The text was broken down into tokens with one phoneme per token, and then they investigated at which token were all of the phonemes observed. For this study, vowel length was not counted as two separate phonemes and neither were tones considered. The median number of tokens needed to observe all phonemes was just over a thousand for both methods that were investigated in the paper. If we conservatively ap-

proximate the number of phonemes per word to be 5, then we would need around 200 words to find every phoneme at least once. This was only done to provide a brief comparison to the Dockum & Bowern (2019) paper where the final recommendation was a minimum of 400 words. The coverage metrics from Dockum & Bowern (2019) and Baird et al. (2022) are the same where they are looking for at least one instance of every phoneme. The key difference between the two papers that is relevant for this paper is that the 2019 paper also included an additional metric to measure faithful phoneme distribution.

One important point of note here is that both of these papers suggest that minimally a wordlist should be longer than many conventional survey wordlists that currently exist. One of the most common wordlists is the Swadesh lists or similar regional adaptations (Bowern 2015:39). The original Swadesh list initially started as the Swadesh-200 list (Swadesh 1952) which was then later compressed into the Swadesh-100 list (Swadesh 1955). An even shorter version was also produced by Holman et al. (2008) for the ASJP database with only 40 concepts and will be discussed further later in the paper. The prevalence of these short wordlists contrasts with the recommendations given by Baird et al. (2022) and Dockum & Bowern (2019) who have both shown that for phonological analysis, it is best to have longer wordlists. This doesn't mean that shorter wordlists don't have their place, but that it is important to consider the match between task and dataset when working with legacy data or data gathered by others.

2.1. WORDLIST DATABASES AND THEIR RATIONALES. Although analysis of wordlists is not an abundantly common research topic currently, there still exist many wordlists. Such as the Lexibank database by the Max Planck Institute for Evolutionary Anthropology (List et al. 2022), the Comparative Bantu Online Dictionary (CBOLD) by the University of California in Berkeley, or even the Contemporary and Historical Reconstruction in the Indigenous Languages of Australia (CHIRILA) by Yale University (Bowern 2016), which is the dataset used for Dockum & Bowern (2019). Each of these wordlists was created to match their task and have their rationale for their length and choice of items.

Another database for wordlists would be the Automated Similarity Judgement Program (ASJP) database which currently holds 10,169 unique ISO codes. These wordlists only contain 40 concepts which the authors of ASJP have said to be sufficient for their task. The purpose of ASJP was to automatically classify wordlists into different language families. They initially tested their program with the 100-word Swadesh lists and found good results (Brown et al. 2008). And in their subsequent paper, Holman et al. (2011), they further reduced the number of concepts to 40 concepts. Their metric for this reduction was a concept's loan resistance. They found that by doing this, they were able to make their language family classifications more accurate. This shows that their 40-word concept list is more than sufficient for their purpose. However automatic language family classification is only one type of task. So even though the ASJP concept list is very well suited to their explicit purpose, this does not mean that this concept list is well suited to other linguistic tasks. Dockum & Bowern (2019) have already shown in their 2019 paper that a randomly selected wordlist of 50 words did not reliably find every phoneme in a language.

Meanwhile, the International Dictionary Series (IDS) is a collection of wordlists that currently has a total of 215 languages. The purpose of these wordlists was to allow comparative linguistic studies to be done. In a paper by some of the creators of IDS (Key & Comrie 2023), they explained the rationale behind their data collection guidelines. The IDS has 1310 different

concepts that they use as a guideline for data collection. While not every language will have every single concept and may be left blank, the guidelines specified that words may be added but never removed from the master list. This was to ensure that all IDS wordlists were compatible for comparative studies. The authors acknowledge that gathering enough data for an IDS wordlist is a time-consuming undertaking and they estimated one wordlist should take about 'one person-month of work'. This is in comparison to an ASJP wordlist which was estimated would only take 'less than a day'. A month of dedicated work will be more expensive than a day's worth of work and also be less suited for volunteer work because of the high amount of commitment needed. As for why the IDS chose the 1310 concepts for their master list, the IDS creators adapted work done by Carl Buck from 1929 and 1949. There were no explicit criteria that Buck used when choosing his words other than the goal, which was to 'work out a tentative and skeleton dictionary covering a limited number, perhaps a thousand, of representative groups of synonyms in the principal IE languages' (Key & Comrie 2023). So while no explicit criteria was laid out for how and why which words should be used and how many, Buck was relying on his experience in doing comparative studies in Indo-European to create a master list. While Dockum & Bowern (2019) suggested a minimal wordlist of 400 items, more data is never a detrimental thing because it widens the scope of what is possible with the data. The creators and editors of the IDS go to great lengths to ensure that the data is standardized and the wordlists are of high quality. Not every project will have the luxury or capability to produce a high-quality wordlist that is this long. So the question of how long is long enough still remains.

3. Methodology. The data for this paper are wordlists of 55 different Kra-Dai languages (Dockum, 2024). These have been aggregated from fieldwork done by other researchers, before being cleaned and standardized to a large extent. Some of the transcriptions from older studies were done in pseudo-IPA. For this study, variation between linguists in phonemic notation should not pose a major problem for the research question. It should be reasonable to assume that each author was internally consistent within each list. So we only account for known notation variations in the scripts which will be described below. For this study, the calculation of each metric only happens with data from one specific language at a time. This is why it is fine for the datasets used in this study to have different phonemic notations.

A list of the languages can be found in the appendix. Each language had at least 2000 items which is the same lower threshold used in Dockum & Bowern (2019) to judge a lexicon large enough to stand in for a "full" lexicon for purposes of this research question.

3.1. SUBSET SIZES. Similar to Dockum & Bowern (2019), for each language, a random subset of words is selected and then metrics are computed for that subset. This is done 1500 times for each language so that the results can be averaged over (see 3.2 for justification of this number). Then the wordlist sample subset size is increased and the whole process is repeated. The subset sizes start at 50 and increase in increments of 50 up to 500. After that, then the subset size increases in increments of 100 up to 1200, as well as subset sizes of 1500 and 2000. Thus the subset sizes tested are as follows:

$$50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1500, 2000$$

More granularity was added to some parts of the chosen subset because these subset sizes showed the most drastic change in results. Furthermore, early results showed that the most likely candidate for a minimal wordlist size was in the 200-500 range which was why the granularity of the subsets was increased to 50 in that range. This follows what Dockum & Bowern (2019) did

in their study which first included a preliminary run with a very wide range of subset sizes before reducing their scope in subsequent runs.

3.2. ITERATIONS. At each different subset sample size, we do the sampling 1500 times and take the average of the results of those 1500 runs. We arrived at this number by plotting our metric values against the number of iterations. The metrics have been normalized between 0 to 1 so that they can be easily compared on the same graph. We looked for the point at which the lines on the graph are completely flat, which was at around 1500 iterations which means they have stabilized and any additional iterations did not change the results very much. There can be arguments made for choosing 1000 iterations or even 500 because the metrics do look pretty stable already at that point. But for this dataset, the computations are not very computationally expensive to run so we have chosen to use a more generous number of iterations. But this threshold will change depending on the dataset so it is wise to rerun these calculations every time when working with new datasets.

3.3. SEGMENTATION. Because of the nature of the languages in this dataset and the format they were in, certain decisions need to be made about what counts as individual sounds, as lexical forms in this dataset are not pre-segmented. And because of the scale of the data, these decisions needed to be able to be codified into the scripts, as hand-evaluating every form in the dataset is not feasible. For this study, we decided to use maximal clusters. This means that a word like /*klʉa*/, which means 'salt' in Thai, would have two sounds: /*kl*/ and /*ʉa*/. The first two consonants are clustered together and considered one sound. This is because we would not want to assume that a language has the /*k*/ and /*l*/ sounds just because they appear in combination. Maybe /*l*/ solely appears following /*k*/. We then would not want to consider a subset wordlist incomplete if we never find an /*l*/ in isolation. The same can be applied to vowels and their off-glides. While some sounds will appear in diphthongs, they may not necessarily appear on their own so we should count each diphthong or triphthong as one sound.

The general guidelines for segmentation are as mentioned in the previous paragraph. However, there are still some edge cases because the data in the Kra-Dai dataset has been transcribed using varying IPA and pseudo-IPA conventions from different locations and points in time. These exceptions to the rules above are then just hard-coded into the procedure. Also, each author favored different methods of analysis which resulted in differences in transcription even for the same sounds so all possible combinations had to be considered and accounted for.

For this study, we did take vowel length into account. If a short and long 'a' existed in the full lexicon, we would need to see both a short 'a' and long 'a' within a subset to say that the subset had full coverage. Because we are going for maximal clustering to decide what counts as a sound, the length of the vowel is important because we are including off-glides in vowel clusters.

On the other hand, lexical tone was not a factor that we took into account. So if we found an 'a' with a first tone, we would not also need to see an 'a' with a second tone. Since this seems like it would just straightforwardly increase the number of sounds in a language, we decided not to focus on this aspect of the dataset for this study. However, this could be a topic for future study.

Also as a way to exclude possible human errors such as typos from the dataset, we set a marginal phoneme threshold. Similarly to Dockum & Bowern (2019), this study has set the marginal phoneme threshold to be 0.5%. This means that if a sound only appears in 0.5% of words in the full lexicon, then we exclude it from our calculations as a phoneme to look for. The reason for

this is that if a phoneme appears too infrequently, then there is a high chance that it is a mistake or typo in the data or a faulty transcription. Dockum & Bowern (2019) used 0.5% as the threshold for marginal phonemes and this paper has seen no compelling reason to change it so far.

3.4. METRICS. A wordlist is judged to be completed based on two criteria introduced by Dockum & Bowern (2019). The first criteria is if we find every phoneme within the wordlists and the second is if the frequency of every phoneme is statistically comparable to the full lexicon.

3.4.1. COVERAGE. The first metric is pretty straightforward. If a wordlist observes every phoneme that is present in the full lexicon, this is a wordlist with full coverage. Then we calculate the coverage score of the language using this formula:

$$coverage = \frac{\text{no. of random subsets with full coverage}}{\text{total no. of random subsets}}$$

This will result in a number between 0 and 1 where are higher number is better. A higher coverage number would indicate that more random subsets are finding every phoneme.

3.4.2. MEAN SQUARED ERROR. Quantifying, if phoneme frequencies are statically comparable, is slightly harder than quantifying coverage. This part of the methodology introduces new techniques that build upon the techniques used by Dockum & Bowern (2019). In the 2019 paper, the authors used a metric called Residual Sum of Squares or RSS. To calculate this, first, the difference between the subset's frequency of a phoneme and the full lexicon's frequency was squared. This was done for every phoneme in the language and then this was summed together as the RSS for a language at a given subset size. This can be shown another way in this equation by calculating the RSS for language X:

$$pf = \text{phoneme frequency}$$

$$X = \text{language}$$

$$RSS_X = \sum_{\substack{\text{p = phoneme}}}^{\text{X's phonemes}} (pf^{subset} - pf^{full})^2$$

In Dockum & Bowern (2019), the authors plotted coverage scores and RSS on a chart and employed elbow finding to make a judgment to see at which subset size there stop being significant improvements to the RSS score. The process of elbow finding will be described in more detail later in the section. This is how the 400-word recommendation mentioned earlier was arrived at.

For this paper, the method of calculating coverage has been kept the same as the 2019 paper. However, for the second criterion, we have decided to change the metric from RSS to three new metrics which are Mean Squared Error (MSE), Mean Absolute Error (MAE), and Max Absolute Error (MaxAE). MSE is very similar to RSS but instead of only adding the squared of the difference together, we will also divide it by the number of phonemes.

$$MSE = \frac{RSS}{no.of phoneme}$$

The reason we have decided to change RSS to MSE for this paper is so that we can compare these values between languages. Because RSS does not take into account phoneme inventory at all, a language with a larger phoneme inventory may have a larger RSS score just because of this. Changing this is to MSE it allows us to do a more detailed comparison of languages with this metric.

3.4.3. MEAN ABSOLUTE ERROR. In addition to this, we have also decided to include MAE as another metric. This is because MSE will exaggerate larger differences so an alternative metric may capture different data points that were overlooked by MSE. The formula that we used to calculate this is below.

$$SumAE = \sum_{\substack{\text{X's phonemes} \\ \text{p = phoneme}}} |pf^{subset} - pf^{full}|$$

$$MAE = \frac{SumAE}{no.ofphoneme}$$

3.4.4. MAXIMUM ABSOLUTE ERROR. And finally, the last metric is MaxAE. To calculate this we want to find the phoneme with the largest absolute difference between their frequency in the subset and their frequency in the full lexicon.

$$MaxAE = max(|pf^{subset} - pf^{full}|)$$

The reason for including this metric is to see if one phoneme is grossly under or over-represented in the subset. When taking the average of the error of the phonemes in the two previous metrics, this may hide the fact that one phoneme is under-represented compared to the others. So to rule out this possibility, we have included this metric as one of the metrics to test. The ideal situation is that the MaxAE is only slightly higher than the MAE. This shows that most of the phonemes have similar levels of error instead of one phoneme being responsible for a lot of that error.

The purpose of having more metrics is to make the minimal wordlist judgment more robust. Each metric captures different types of errors which helps to highlight different shortcomings of each wordlist. In Dockum & Bowern (2019), only one metric was used to do elbow finding.

3.5. ELBOW FINDING. Elbow finding is the process of looking at how a dependent variable changes as we change the independent variable and finding the point at which we have the most improvement for the least amount of work or the elbow. In the context of this study, this would be finding the point at which we can be sure that a subset is complete with the least amount of random words elicited. The word elbow comes from the fact that in graphs of this nature there is a sharp point where we see the gradient of change go from steep to flat. This change in gradient indicates that this is the point where good results can be obtained with the least amount of effort. However, elbow finding is subjective but with more metrics, this allows us to see if the elbows align. By also checking if the elbows align, we provide stronger evidence to validate our claims.

4. Results.

4.1. KRA-DAI RESULTS. In this section, we will discuss the results of this study and some of their implications. In Figures 1, 2, 3 and 4, you can see the results of all of the metrics average over every language in the Kra-Dai dataset.

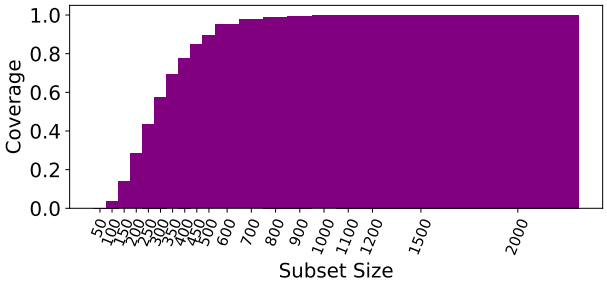

Figure 1. Coverage of the Kra-Dai dataset for subset sizes 50-2000

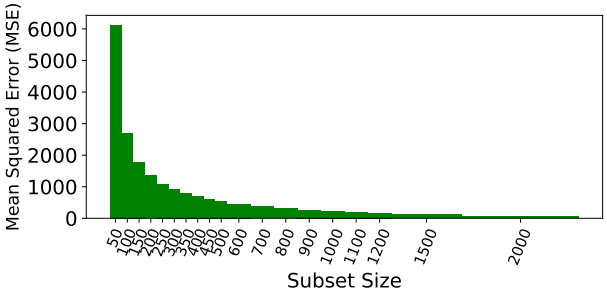

Figure 2. Meanse Squared Error (MSE) of the Kra-Dai dataset for subset sizes 50-2000

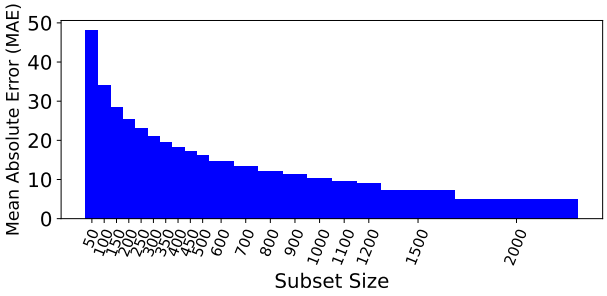

Figure 3. Mean Absolute Error (MAE) of the Kra-Dai dataset for subset sizes 50-2000

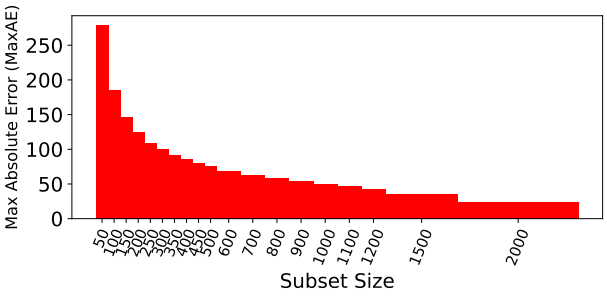

Figure 4. Max Absolute Error (MaxAE) of the Kra-Dai dataset for subset sizes 50-2000

In Figure 1, the subset sizes are shown on the x-axis, and the coverage percentage on the y-axis. How coverage is calculated for each subset can be found in the methodology section. For this visualization, the coverage values of 1500 [1] subsets in each subset sized are averaged to give a number between 0 and 1. 0 means that none of the 1500 subsets contained every phoneme from that language. 1 would mean that all of the 1500 subsets contained every phoneme from that language. So the trend we are hoping to see is that the coverage average increases as we increase the subset size which is shown along the button of the graph. We expect this because the more words that are present within a subset, the higher the possibility that those words will contain every phoneme. We can see that the trend is as we expected and that the coverage average gets very close to 1 at around 700-800 words in a subset.

The next three graphs produced for the Kra-Dai data are the three metrics used to measure distribution of a subset. Similar to the coverage graph, the subset sizes are listed along the x-axis and the specific error metric is on the y-axis. As described in the methodology section, the error metrics used in this paper are Mean Squared Error (MSE), Mean Absolute Error (MAE), and Max Absolute Error (MaxAE). For these metrics, a higher number on a metric means there is more error within a subset. So for these three metrics, we are looking for a value that is as close to 0 as possible. This means the trend we are expecting to see is that the error metric decreases as the subset sizes increases. We expect this to occur because as we increase the number of words in every subset, the frequency at which phonemes appear should become more and more similar to their frequency in real-life usage. As we can observe in Figure 2, 3, and 4, we can see that the trend matches our expectations. All three error metric values decrease as the subset sizes in-

[1] Reference Section 3.2

crease.

The process of elbow finding has been explained in the methodology section and now we have to apply the elbow finding principle to these graphs. A big problem with all but one of these four graphs is that there is not an immediately apparent elbow. Figure 2 would be the only graph where we could make a convincing argument for placing the elbow at around in the 300-450 range.

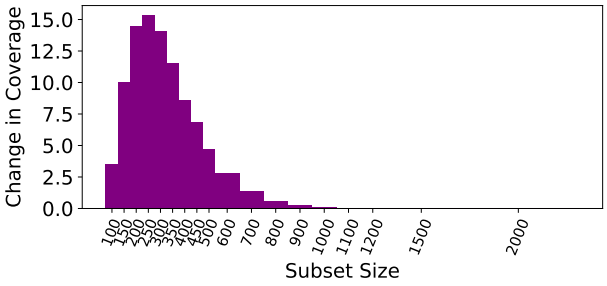

Figure 5. Percent change in coverage in the Kra-Dai dataset for subset sizes 100-2000

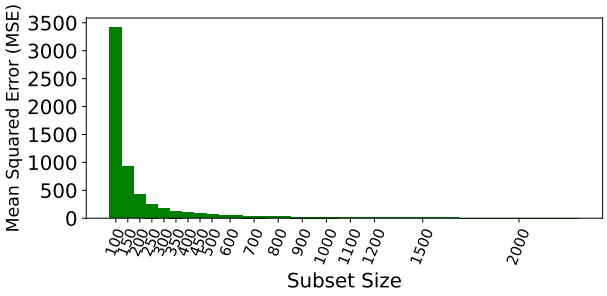

Figure 6. Percent change in Mean Squared Error (MSE) in the Kra-Dai dataset for subset sizes 100-2000

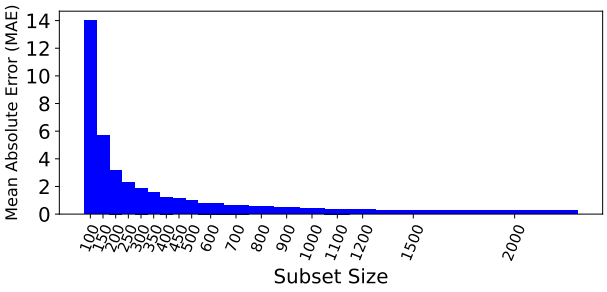

Figure 7. Percent change in Mean Absolute Error (MAE) in the Kra-Dai dataset for subset sizes 100-2000

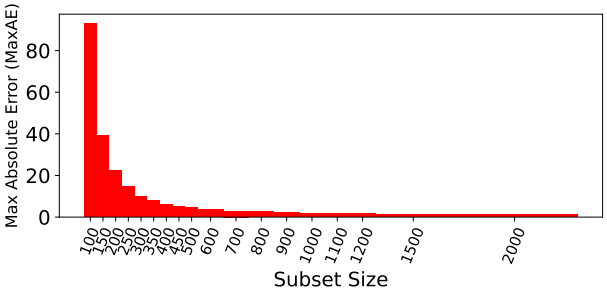

Figure 8. Percent change in Max Absolute Error (MaxAE) in the Kra-Dai dataset for subset sizes 100-2000

4.2. ALTERNATIVE VISUALIZATIONS. Because the current graphs are not as clear cut as we would have hoped for, we then produced alternative visualization to hopefully allow us to draw a more robust conclusion from this data. Instead of just plotting the results on the y-axis, we will instead plot the difference in a metric when moving from one subset size to the next subset size. These graphs can be seen in Figure 5, 6, 7 and 8. These graphs better highlight the amount a metric changes as we increase the subset size. So at the point where we see little to no improvement when increasing the subset size, it means that that is a good place to stop. While having a very large wordlist size will increase the completeness of a wordlist, fieldwork and language documentation can be very expensive in terms of time and money so we would to find a place to stop that will be the least expensive in terms of resources but still yield reliable results.

From these graphs, it becomes apparent that all of the distribution metrics (MSE, MAE, MaxAE) have clear elbows and are around the 300-400 words mark. We can see that after this point, increasing the size of the subset only yields very minuscule improvements which are not

worth doing if resources are limited. If we look at the raw values of the distribution metric such as MAE, we will see that the improvements start to stagnate around MAE of 100. An MAE of 100 means that a phoneme is on average being over or underrepresented by 100%. So to give an example scenario, say a phoneme should appear in 5% of words, in the subset it may appear in 0% of words or 10% of words to get an MAE of 100. The key point here is that it may appear in 0% of words in the subset. This simply means that we are not seeing this phoneme at all which would lower our coverage scores.

It becomes more clear here that the limiting factor is the coverage number in Figure 5. As can be seen in the graph, it takes a lot longer for the change in coverage to get close to zero which happens around the 700-800 range. This means that any increase in wordlist size before that threshold will still cause significant improvements in the chance of finding every phoneme in your language. Having shown that coverage is the limiting factor in this experiment means that coverage will be the metric that carries the most weight when deciding how long wordlists should be.

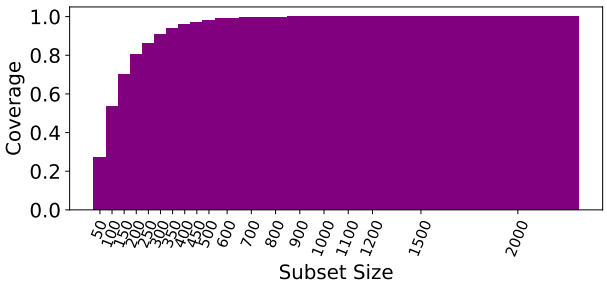

Figure 9. Coverage of the Chirila dataset for subset sizes 50-2000

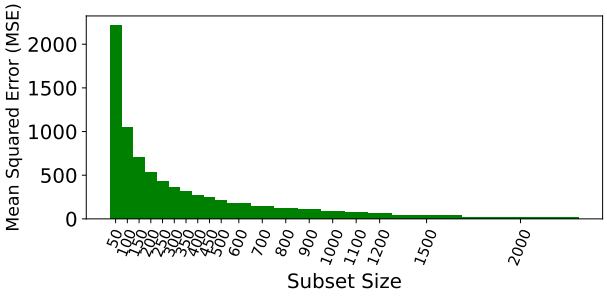

Figure 10. Mean Squared Error (MSE) of the Chirila dataset for subset sizes 50-2000

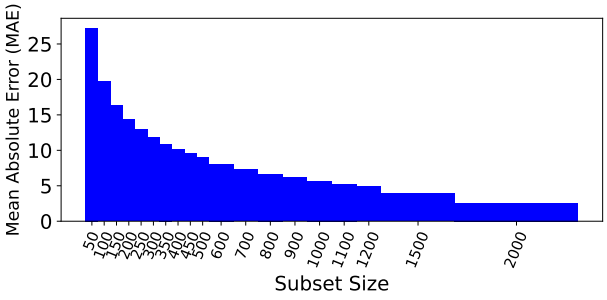

Figure 11. Mean Absolute Error (MAE) of the Chirila dataset for subset sizes 50-2000

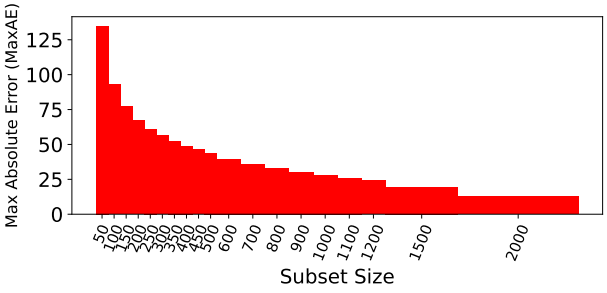

Figure 12. Max Absolute Error (MaxAE) of the Chirila dataset for subset sizes 50-2000

4.3. CHIRILA DATASET RESULTS WITH NEW METHODOLOGY. To provide a comparison to the results from the Kra-Dai data, the data from the Chirila dataset used in the original Dockum & Bowern (2019) paper was run with the new methodology used in this study. The results can be seen in figures 9, 10, 11 and 12. The most notable difference that can be observed in these graphs, in comparison to the ones generated from the Kra-Dai languages dataset, is that the graph showing average coverage is much steeper in the first few subset sizes and reaches an average value of close to 1 much quicker than shown in Figure 1. This comparison can be clearly shown

in Table 1 which compares the point at which each dataset crosses certain benchmarks. Because we have shown earlier that coverage is the limiting factor, we will only be comparing the results for coverage. As is shown in the table, if you randomly sample around 300 words from a language in the Chirila dataset or a similar Australian language, 75% of the time you will have a dataset that contains every phoneme in that language. Then you compare this to a language in the Kra-Dai dataset or a similar language, you will need around 400 words to reach to same level of confidence. This chart may also be used to inform decisions as not every task requires phonologically complete wordlists, such as language family classification done in Holman et al. (2008). It was shown that they were able to get good results on wordlists as short as 40 words. Different tasks require varying degrees of confidence in the wordlist being phonologically complete, so it is possible to use Table 1 as a rough guideline.

Probability of having full coverage	**Number of Words for Kra-Dai**	**Number of Words for Chirila**
75%	400	300
90%	600	400
95%	600	500
98%	800	600
99.5%	900	800

Table 1. Probability of full coverage

4.4. LANGUAGE COMPARISONS. The previous section showed graphs that are averaged over all the languages in the dataset. So far, we have chosen to mainly focus on the aggregated data from every language because we observed that most of the languages from the Kra-Dai dataset produced very similar trends. However, there is still much that can be learned from examining the results of individual languages.

For example, in the Kra-Dai dataset, the language with the smallest phoneme inventory is Yongbei Zhuang. Yongbei Zhuang has a total phoneme inventory of 32 with 13 consonants and 19 vowels. The coverage graphs for this are shown in Figure 13. Because the distribution metric graphs are very similar we am not including them here. The coverage graph is notably more similar to the averaged coverage Chirila dataset graph in Figure 9 than the aggregated graph for the Kra-Dai dataset in Figure 1. Australian languages are well known for having comparatively small phoneme inventories. The language with the largest inventory in the Chirila dataset is Wubuy with 25 consonants and 8 vowels for a total of 33 phonemes with is very close to Yongbei Zhuang's 32 total phonemes. Wubuy's coverage graph can be seen in Figure 16.[2] This shows that phoneme inventory size affects a wordlist size more than its language family. To illustrate this point, we will compare the subset size needed to reach a 95% probability of having full coverage for the languages with the smallest and largest phoneme inventory from the Kra-Dai and Chirila datasets. As stated earlier, Yongbei Zhuang is the Kra-Dai language with the smallest phoneme inventory and Maonan is the one with the highest. For the Chirila dataset, Miriwoong is the lan-

[2] See Dockum & Bowern (2019), footnote 3 for discussion of Wubuy and its use of *archiphonemes*, which makes it less comparable to other wordlists from the Chirila dataset. Wubuy was excluded from that study, but it is retained here.

guage with the smallest phoneme inventory and Wubuy is the one with the highest. The coverage graphs of Miriwoong and Maonan are figures 15 and 14 respectively.

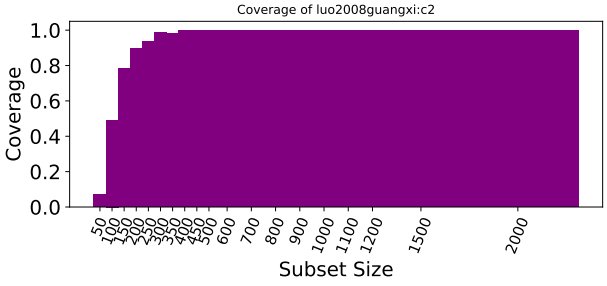

Figure 13. Coverage of Yongbei Zhuang

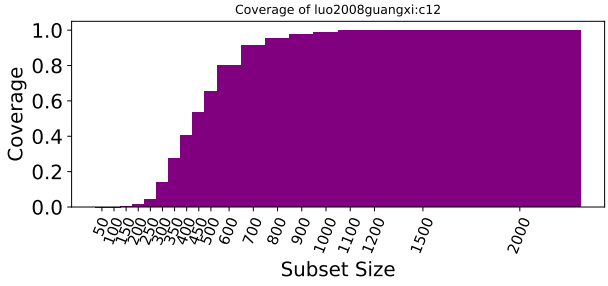

Figure 14. Coverage of Maonan

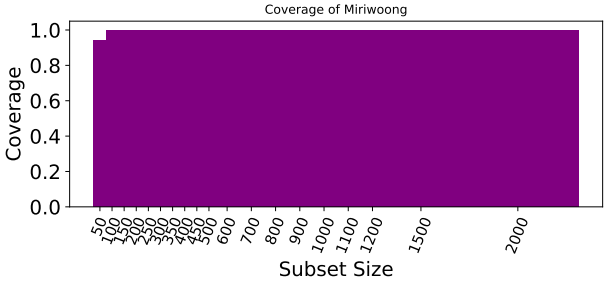

Figure 15. Coverage of Miriwoong

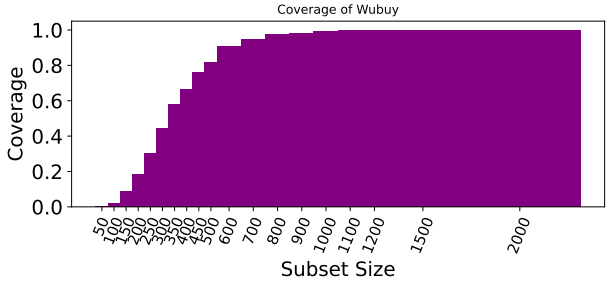

Figure 16. Coverage of Wubuy

So Table 2 shows the statistics of these four languages and how they compare against one another. One interesting point about this comparison is that even though both Yongbei Zhuang and Wubuy have very similar phoneme inventories, Wubuy requires a much longer wordlist to reach the same confidence level as Yongbei Zhuang. So this shows that there are other factors than just phoneme inventory size that can affect the wordlist size required for a phonologically complete wordlist.

	Miriwoong	**Yongbei Zhuang**	**Wubuy**	**Maonan**
Total phonemes	18	32	33	64
Consonants	14	13	25	43
Vowels	4	19	8	21
Subset size for 95% confidence level	100	300	800	800

Table 2. Comparison of different languages coverage

4.5. CORRELATION OF PHONEME INVENTORY AND SUBSET SIZES. The take this idea further, we created a new visualization to show the correlation between phoneme inventory and subset size required for a minimally complete wordlist.

Figure 17 shows all the languages from both the Kra-Dai and Chirila datasets plotted on one graph, with the Chirila dataset being the green dots and the Kra-Dai being the purple dots. The x-axis is the total phoneme inventory of each language and the y-axis is the subset size at which the

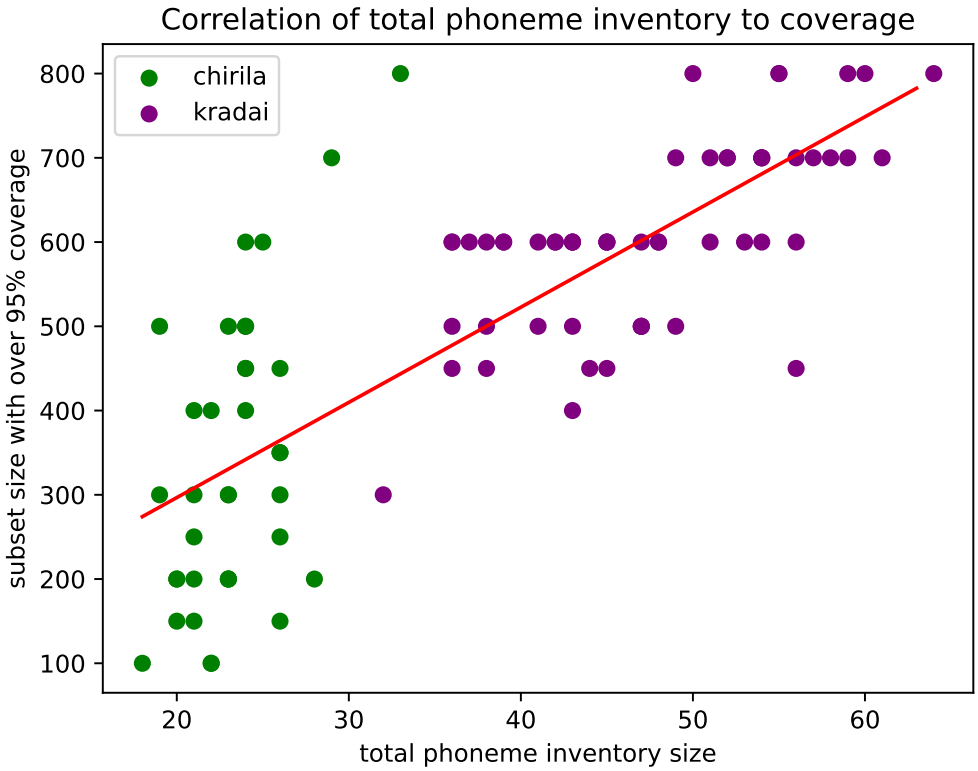

Figure 17. Correlation of total phoneme inventory to coverage at 95% confidence

language achieved full phoneme coverage for that language in at least 95% for the 1500 runs in our tests. We can see that the Kra-Dai language on average has much larger phoneme inventories compared to the languages from the Chirila dataset. To show the correlation, we have drawn a line of best fit using Ordinary Least Squares (OLS). OLS creates a regression line that minimizes the distance of all the points given to that line. The OLS formula produced this equation for the line of best fit:

$$y = 11.30x + 70.66 \tag{1}$$

$$r^2 = 0.624$$

We have a positive x coefficient which shows that as the x variable gets larger so does the y variable which matches our expectations of larger phoneme inventories needing larger wordlist sizes. The r^2 value of 0.624 shows that around 60% of the variance in the y variable (wordlist size) can be explained by the x variable (phoneme inventory size). This shows us that there is very strong evidence to suggest that these variables are correlated and that there is a clear trend. Because OLS provides an equation for the line of best fit, it is then possible to extrapolate these results to other languages. One problem with this is that if no fieldwork has been done yet, how do we know the phoneme inventory of a language? But if we did know, for example, the language family ahead of time, it should be possible to use this equation to give a more accurate

recommendation for the wordlist size that is desired. However, this equation was made with limited data from only two datasets so these results should be used with that in mind.

5. Recommendations and Conclusion. Drawing from these results, we can recommend new takeaways to help inform language documentation planning. Building upon the conclusions from Dockum & Bowern (2019) which also looked at coverage and distribution metrics is that between these two metrics, coverage is the limiting factor. The distribution metrics, MAE, MSE, and MaxAE, becomes more important when we are consistently getting high coverage. However because most wordlists already struggle to achieve full coverage, the distribution metrics become less important. This is because if a phoneme is missing from a wordlist, then it will also negatively affect the distribution metrics. So when using this work to help with fieldwork logistics, the coverage metric is the most important metric to consider when making those decisions. But this does create a little bit of a chicken and egg situation where to know how long a wordlist should be, you want to know how many phonemes exist in a language. If this is a language that has not been documented before, how do you know how many phonemes it has?

The best recommendation we can make based on the results of this study is to first reference Table 1 and pick a level of confidence that is appropriate for the intended use of the wordlist. Next, insofar as possible, estimate an expected phoneme inventory size in a language, given information already known about it (e.g. typological profile, language family, geographic region). Then use Equation (1) (or equivalent equation based on the level of confidence needed) to make a statistically well-motivated estimate of how large the minimal wordlist for each language under study will need to be.

In sum, the findings of this paper support the findings of Dockum & Bowern (2019) and Baird et al. (2022), and we also conclude that wordlists should be longer than has often been common practice for survey-style documentation work. The recommendation in Dockum & Bowern (2019) of 400 words for a wordlist may be an overly simplified rule of thumb, but is still a reasonable and actionable recommendation that is backed up by the findings from this paper and is still a good baseline if it is not possible to use any of the more nuanced methods suggested by this paper.

There will inevitably be limitations to the conclusions drawn here. This paper does not fully answer all of the questions posed at the start. However, the most important conclusion from this paper is still that wordlists should be longer than they have often traditionally been, if we want to accurately glean from them an accurate phonological profile of a language variety. The results from this study also strongly advocate for a data-driven, procedural approach to both logistical planning of survey-style fieldwork, and to research design for studies that plan to work with legacy field data. While additional research is needed to further advance this discussion of wordlist evaluation, nonetheless, this paper has introduced additional findings and measures to move this topic forward.

For full bibliographic detail of the wordlists used in the study, please take a look at the full thesis version of this paper for details. It can be accessed at https://works.swarthmore.edu/theses/957/

References

Baird, Louise, Nicholas Evans & Simon J. Greenhill. 2022. Blowing in the wind: Using 'North Wind and the Sun' texts to sample phoneme inventories. *Journal of the International Pho-*

netic Association 52(3). 453–494. https://doi.org/10.1017/S002510032000033X.

Bowern, C. 2015. *Linguistic Fieldwork: A Practical Guide*. London: Palgrave Macmillan UK 2nd edn. https://doi.org/10.1057/9781137340801.

Bowern, Claire. 2016. Chirila: Contemporary and Historical Resources for the Indigenous Languages of Australia. *Language Documentation & Conservation* 10. Publisher: University of Hawaii Press.

Brown, Cecil, Eric Holman, Søren Wichmann & Viveka Velupillai. 2008. Automated Classification of the World's Languages: A Description of the Method and Preliminary Results. *STUF - Language Typology and Universals, v.61, 285-308 (2008)* 61. https://doi.org/10.1524/stuf.2008.0026.

Dockum, Rikker & Claire Bowern. 2019. Swadesh lists are not long enough: Drawing phonological generalizations from limited data. *Language Documentation and Description* 16(0). https://doi.org/10.25894/ldd112. Number: 0 Publisher: Aperio Press.

Dryer, Matthew S. & Martin Haspelmath (eds.). 2013. *Wals online (v2020.3)*. Zenodo. https://doi.org/10.5281/zenodo.7385533.

Gasser, Emily & Claire Bowern. 2013. Revisiting phonotactic generalizations in australian languages. In *Proceedings of the annual meetings on phonology*, .

Good, Jeff & Calvin Hendryx-Parker. 2006. Modeling contested categorization in linguistic databases. In *Proceedings of the emeld 2006 workshop on digital language documentation: Tools and standards: The state of the art*, 20–22.

Holman, Eric, Søren Wichmann, Cecil Brown, Viveka Velupillai, André Müller & Dik Bakker. 2008. Explorations in automated language classification. *Folia Linguistica, v.42, 331-354 (2008)* 42. https://doi.org/10.1515/FLIN.2008.331.

Holman, Eric W., Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, Oleg Belyaev, Matthias Urban, Robert Mailhammer, Johann-Mattis List & Dmitry Egorov. 2011. Automated Dating of the World's Language Families Based on Lexical Similarity. *Current Anthropology* 52(6). 841–875. https://doi.org/10.1086/662127. Publisher: The University of Chicago Press.

Key, Mary Ritchie & Bernard Comrie (eds.). 2023. *IDS*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://ids.clld.org/.

List, Johann-Mattis, Robert Forkel, Simon J Greenhill, Christoph Rzymski, Johannes Englisch & Russell D Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data* 9(1). 1–16.

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.

Wichmann, Søren, Eric W. Holman & Cecil H. Brown. 2022. CLDF dataset derived from Wichmann et al.'s "ASJP Database" v20 from 2022. https://doi.org/10.5281/zenodo.7079637.