Beyond a means to an end: A case study in building phonotactic corpora for Central Australian languages

Saliha Muradoğlu🐟 **James Gray**🐞 **Jane Simpson**🐦 **Michael Proctor**🕷 **Mark Harvey**🐟
🐦The Australian National University (ANU) 🐟University of Newcastle
🕷Macquarie University 🐞Independent Scholar
`Firstname.Lastname@ {`🐦`anu.edu.au,` 🐟`newcastle.edu.au,`
🕷`mq.edu.au,` 🐞`@alumni.anu.edu.au}`

Abstract

Linguistic datasets are essential across fields: computational linguists use them for NLP development, theoretical linguists for statistical arguments supporting hypotheses about language, and documentary linguists for preserving examples and aiding grammatical descriptions. Transforming raw data (e.g., recordings or dictionaries) into structured forms (e.g., tables) requires non-trivial decisions within processing pipelines. This paper highlights the importance of these processes in understanding linguistic systems. Our contributions include: (1) an interactive dashboard for four central Australian languages with custom filters, and (2) demonstrating how data processing decisions influence measured outcomes.

1 Introduction

With ubiquitous use of advanced NLP systems for language technology and linguistics (often by proxy), linguistic corpora and the processing it entails are often treated as a means to an end.

In this paper, we show that the process is vital in enhancing our understanding of linguistic systems. Each step in the processing pipeline embodies a linguistic decision that can be non-trivial. For example, when building a phonotactic corpus, we want each entry to be a root. But how do we judge what constitutes a root? Should the decision be structural or semantic? The definition of how to classify a root has been a subject of numerous literature (Harley, 2014; Embick, 2021; Gouskova, 2023). To help guide this decision making we present an interactive web interface, to highlight the flow-on effects of analysis decisions.

This system was designed with the following questions in mind: (1) Are vowels distributed evenly across syllable positions? (2) Does the vowel distribution by syllable position change across different parts of speech (POS)? (3) Do some vowels occur more frequently in the root final position? (4) Does the characteristic of the following consonant affect the distribution of vowels? (5) If the initial vowel is /a/, then what is the distribution of vowels in syllable 2; if the initial vowel is /i/, then what is the distribution of vowels in syllable 2; if the initial vowel is /u/, then what is the distribution of vowels in syllable 2?

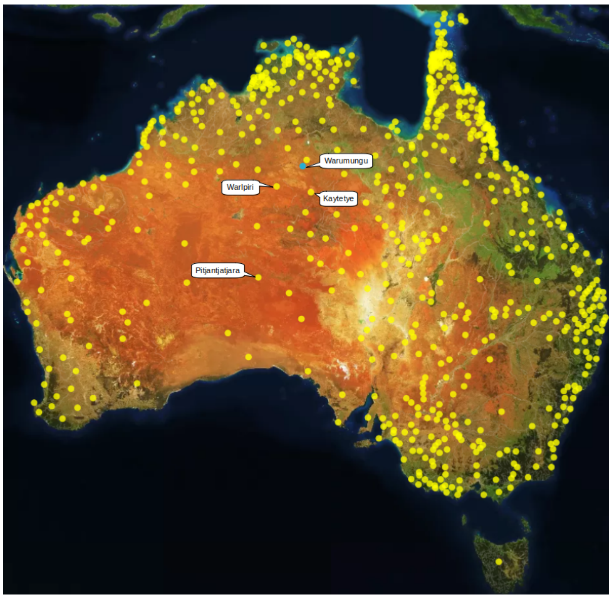
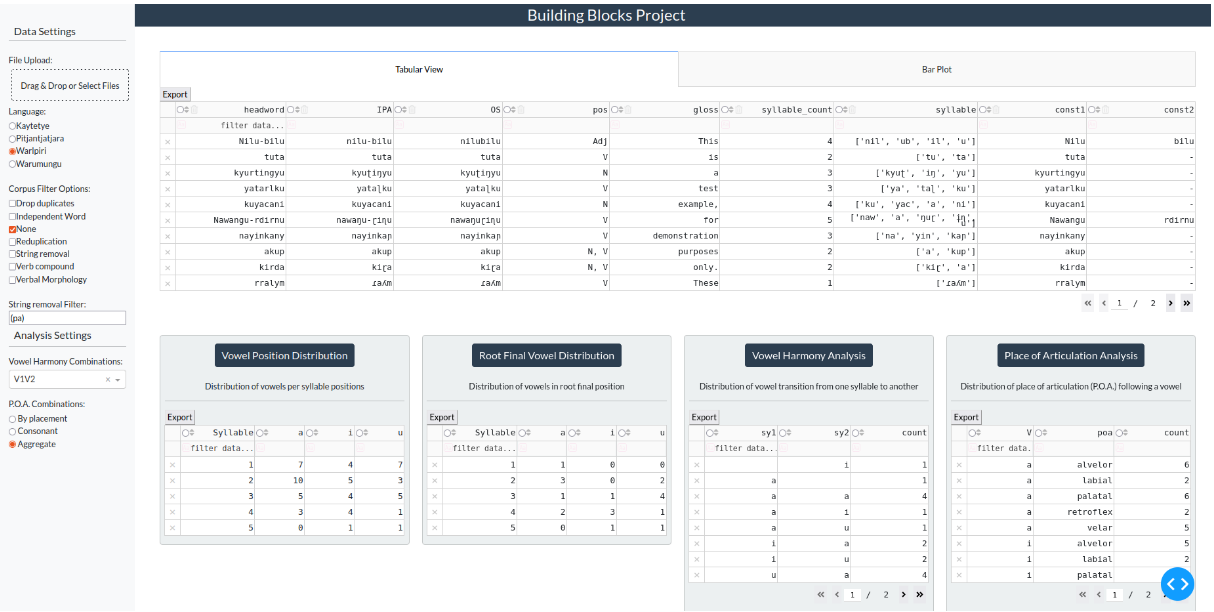


Figure 1: First Languages map of Australia with indicative locations for speakers of Kaytetye, Pitjantjatjara, Warlpiri and Warumungu. Image adapted from Gambay.

Our contributions are two-fold; first, we present an interactive dashboard for four central Australian languages with custom filter functions; second, we show that the processing of raw data into a desired format is embedded with decisions that alter the measured outcomes.

Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025), pages 32–37
March 2, 2025 ©2025 Association for Computational Linguistics



Building Blocks Project

Data Settings

File Upload:

Drag & Drop or Select Files

Language:

○Kaytetye
○Pitjantjatjara
●Warlpiri
○Warumungu

Corpus Filter Options:

□Drop duplicates
□Independent Word
☑None
□Reduplication
□String removal
□Verb compound
□Verbal Morphology

String removal Filter: (pa)

Analysis Settings

Vowel Harmony Combinations: V1V2

P.O.A. Combinations:

○By placement
○Consonant
●Aggregate

Tabular View | Bar Plot

Export

headword	IPA	OS	pos	gloss	syllable_count	syllable	const1	const2
filter data...								
Nilu-bilu	nilu-bilu	nilubilu	Adj	This	4	['nil', 'ub', 'il', 'u']	Nilu	bilu
tuta	tuta	tuta	V	is	2	['tu', 'ta']	tuta	-
kyurtingyu	kyuʈiŋyu	kyuʈiŋyu	N	a	3	['kyuʈ', 'iŋ', 'yu']	kyurtingyu	-
yatarlku	yataɭku	yataɭku	V	test	3	['ya', 'taɭ', 'ku']	yatarlku	-
kuyacani	kuyacani	kuyacani	N	example,	4	['ku', 'yac', 'a', 'ni']	kuyacani	-
Nawangu-rdirnu	nawaŋu-ɽiɳu	nawaŋuɽiɳu	V	for	5	['naw', 'a', 'ŋuɽ', 'iɳ', 'u']	Nawangu	rdirnu
nayinkany	nayinkaɲ	nayinkaɲ	V	demonstration	3	['na', 'yin', 'kaɲ']	nayinkany	-
akup	akup	akup	N, V	purposes	2	['a', 'kup']	akup	-
kirda	kiɽa	kiɽa	N, V	only.	2	['kiɽ', 'a']	kirda	-
rralym	ɹaʎm	ɹaʎm	V	These	1	['ɹaʎm']	rralym	-

« ‹ 1 / 2 › »

Vowel Position Distribution

Distribution of vowels per syllable positions

Export

Syllable	a	i	u
filter data...			
1	7	4	7
2	10	5	3
3	5	4	5
4	3	4	1
5	0	1	1

Root Final Vowel Distribution

Distribution of vowels in root final position

Export

Syllable	a	i	u
filter data...			
1	1	0	0
2	3	0	2
3	1	1	4
4	2	3	1
5	0	1	1

Vowel Harmony Analysis

Distribution of vowel transition from one syllable to another

Export

sy1	sy2	count
filter data...		
	i	1
a		1
a	a	4
a	i	1
a	u	1
i	a	2
i	u	2
u	a	4

« ‹ 1 / 2 › »

Place of Articulation Analysis

Distribution of place of articulation (P.O.A.) following a vowel

Export

V	poa	count
filter data.		
a	alvelor	6
a	labial	2
a	palatal	6
a	retroflex	2
a	velar	5
i	alvelor	5
i	labial	2
i	palatal	

« ‹ 1 / 2

Figure 2: *Interface and system design.* The left-aligned side bar entails a settings control panel, starting with upload options, followed by options for language and filtering function to be applied. In the bottom half of the side bar, two analysis settings are presented: the vowel positions for vowel harmony, and the level of detail required for place of articulation distribution. The center console entails two tabs: 'Tabular View' shows an interactive table of the data uploaded post-filtering, 'Bar Plot' shows the distribution of words with respect to word length (see Figure 3). Depicted beneath the console are four distributions calculated from the dataset: vowel distribution per syllable, root final vowel distribution, vowel harmony, and place of articulation distributions. (Note: examples shown are made-up, for demonstration purposes only).

2 Related Work

Anthony (2022) outlines the differences between online, offline and DIY corpus tools. Online tools are hosted on a cloud and accessible via the internet (such as *english-corpora.org, sketchengine.eu*. Offline are tools such as AntConc, WordSmith Tool or LancsBox (Brezina and Platt, 2023) which run on a local device. Finally, Do-It-Yourself (DIY) describes scripts developed by researchers. The major drawback of DIY tools is the programming skills needed, but are otherwise largely successful in providing tailored, innovative solutions for niche, language-specific concerns.

The majority of corpus tools are built to examine word-level statistics, such as frequency or concordance. While it is possible to adapt these to analyse intra-word components, it can be intricate. Addressing these concerns are often not possible with standard tools (Anthony, 2012). Biber (1988) advocates for DIY tools, given their adaptability and efficiency to phenomena and corpus size. Further, DIY circumnavigate propriety software. In our design, we propose a local web-based interface to ensure data privacy and longevity.

Previous studies have presented online calculators for phonotactic distributions: English (Vitevitch and Luce, 2004; Storkel and Hoover, 2010), Modern Standard Arabic (Aljasser and Vitevich, 2018) and Czech (Čechová et al., 2023). Most of these resources appear to be hosted on external servers and some are no longer available.

Phonotactic structures in Australian languages have been studied through the lenses of historical linguistics and typology (Macklin-Cordes and Round, 2020, 2022), and it is advantageous for researchers to be able to bring insights from historical, areal and typological phenomena to inform analyses at different stages of the workflow.

3 Languages

Australian languages are characterised by small vowel inventories – often three distinctions in place and quality (Fletcher, 2014; Baker, 2014). Over three quarters of Australian languages have a vowel inventory between three and six vowels (Round, 2023a). Consonant inventories exhibit elaborate place contrasts, but comparatively few manners of articulation (Fletcher, 2014; Round, 2023b). Vowel harmony is observed more than 50% of the time across adjacent syllables in several Australian languages (Round, 2023b).

The phonetic inventories of the languages considered here are listed in Table 1.

Kaytetye Kaytetye (ISO 639-3: gbb) is part of the Arandic branch of the Pama-Nyungan family. It is primarily spoken in Kaytetye country, which is approximately 300km north of Alice Springs (see figure 1 for approximate geographic region) (Turpin, 2000). Vowel inventory in Kaytetye has been a subject of discussion, with accounts varying from two (/a/ and /ə/) and four ([i], [a], [ə], [u]) (Harvey et al., 2023). In this paper we follow the four vowel analysis as we utilised the root corpus developed by (Panther, 2021)[1].

Pitjantjatjara Pitjantjatjara (ISO639-3: pjt) is a dialect of the Western Desert Language (Douglas et al., 1964) and is a part of the Pama-Nyungan family. In 2016, over 3,000 speakers were recorded (Wilmoth, 2022). It is closely related to the Yankunytjatjara dialect (Goddard, 1983, 2001). It follows the norm for Australian languages, with a three vowel system and a consonant inventory that spans many places of articulation but fewer manners of articulation (Tabain et al., 2014; Tabain and Butcher, 2014).

Warlpiri Warlpiri (ISO 639-3: wbp) is spoken in the northwest of Alice Springs by a few thousand people. It is a Pama-Nyungan language. It has one of the largest speaker populations of the Australian languages (Nash, 1980). It aligns with the typical inventory of Australian languages, featuring a three-vowel system and a consonant inventory with diverse articulation points but few articulation manners (Loakes et al., 2008). Warlpiri has been a subject of extensive study, particularly in the domain of syntax, given its free word order (Nash, 1980; Simpson, 1983, 2012).

Warumungu Warumungu (ISO 639-3: wrm) is spoken by a few hundred people in the central part of the Northern Territory of Australia around Tennant Creek. It is a member of the Desert Nyungic branch of the Pama-Nyungan family. It is closely related to Warlpiri (Simpson, 2017). The Warumungu sound system is typical of Australian languages. A three-way vowel system, five places of articulation and eight different possible manners of articulation. Warumungu differs by having a second stop series.

Language	Consonant Inventory	Vowel Inventory
Pitjantjatjara	{c, j, k, l, m, n, p, r, t, w, ŋ, ɭ, ɲ, ɳ, ɻ, ʈ, ʎ}	{a, i, u}
Warlpiri	{c, k, l, m, n, p, r, t, w, y, ŋ, ɭ, ɭʈ, ɲ, ɲc, ɳ, ɳʈ, ɽ, ɾ, ʈ, ʎ, ʎc}	{a, i, u}
Warumungu	{c, k, l, m, n, p, r, t, w, y, ŋ, ɭ, ɲ, ɳ, ɾ, ʈ, ʎ}	{a, i, u}
Kaytetye	{c, cɲ, k, l, l̪, m, n, n̪, p, r, t, t̪, t̪n̪, w, y, ɰ, ŋ, ɭ, ɳ, ɲ, ɾ, ʈ, ʈɳ, ʎ}	{a, i, u, ə}

Table 1: Vowel and consonant inventories of the four languages included in the analysis.

4 System

A key consideration for this project is flexibility in working with the various forms of available data and different approaches to encoding similar phenomena. For example, one linguist might choose to encode a gloss field with additional notes, while another does not. Corollary to this, custom filters and calculations can be added to the system.

An additional consideration is privacy, given the non-public nature of some of these databases. For this reason, the system is designed to be run locally via a Jupyter notebook on the operators computer.

We use Plotly dash module (Albini et al., 2022; Schroeder et al., 2022) to generate an interactive dashboard[2] .

4.1 Pre-processing

The system we present consists of two sections. The first, a preprocessing step that involves transforming hierarchical dictionary data into a tabular form. While the transformation can be extended to extract additional fields, for the purposes of building a root database this step extracts the **headword**, **POS** and **gloss**. Limiting to these three fields also allows for flexibility across various legacy sources and documentation styles.

An additional step is needed for the Warumungu data, since the pos and gloss fields in the dictionary file contained additional notes. It is language- and linguist-specific, but can be taken as an example for other such considerations.

[1] For an overview of Kaytetye phonetics and phonology see Harvey et al. (2015); Turpin and Ross (2012); Panther (2021).

[2] All code is available at `https://github.com/smuradoglu/phc`

4.2 Dashboard design

4.2.1 Tabular View

Once the tabular data consisting of the **headword**, **POS** and **gloss** triplet is uploaded into the system, six additional columns are added.

The headword is mapped to IPA based on language-specific vowel and consonant inventories. To allow for traceability, we have kept the headword entry as it is found in the original file (dictionary). The 'OS' column reflects the operational string that is used for consequent calculations. This field becomes more relevant as the filter options are added. Syllable count is calculated by counting the vowels in each word. This is meant as an independent operation from the adjacent syllable column, to validate the predictions.

The syllable column reflects predictions of syllable structure based on the NLTK legality principle module (Bird, 2006). This module is implemented using the Legality Principle, which states that syllable onsets and codas are only legal if they are found as word onsets or codas in the language. Since onsets are most likely maximised, the longest legal onset is prioritized.

The last two columns show the constituents of the headword entry separated by hyphens ('-'). This column is later used for filtering reduplications and verb compounds.

4.2.2 Bar Plot

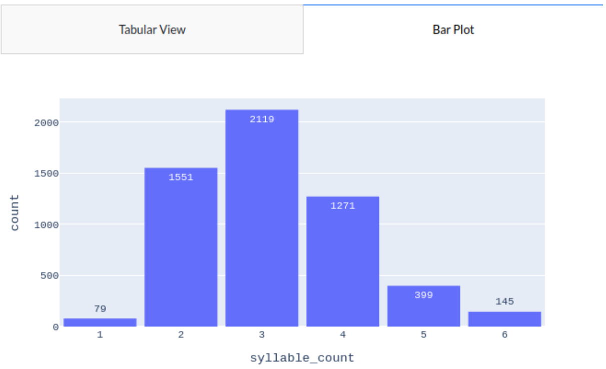


Figure 3: Bar plot showing distribution of words with respect to word length (syllables).

Using the syllable count from the table, a bar plot is produced (shown in Figure 3). This is a quick way to examine the distribution of word length (in syllables) with respect to number of words.

4.2.3 Filter options

String removal This is a straightforward function that filters the string sequence inputted by the user. It is motivated by the occurrence of '(pa)' in Warlpiri and Warumungu dictionary entries. '(pa)' is a semantically meaningless element which is sometimes added word finally to avoid illicit phonotactic consonant final words. As such, the default value is set to '(pa)'. However, it can be used to be a filter for any other string.

Reduplication This filter utilises the '-' marking out different sections of the word (separated out as **const 1** and **const 2** as shown in 2). It compares these two columns. If they match it only considers the first column for the subsequent calculations of syllable count and syllable structure. We remove the second occurrence to avoid a bias in the data towards those sound combinations.

Verb Compound In a similar manner to the reduplication filter, this option utilises **const 1** and **const 2** . It checks whether the second constituent is an entry in it's own right. If it is, it is not considered for the following calculations.

Verbal Morphology This filter is language-specific and as the name suggests, deals with the verbal morphology. In effect it strips verbs of their inflection. In the languages considered here, only suffixes are applicable.

Independent Word When checked, this option removed dependent words like clitics. These are typically marked as beginning with '-' in dictionaries. As such this function simply filters out words beginning with '-'.

Drop duplicates This option removes duplicates based on the proposed syllable structure. It is mainly useful after other filters have been applied (although it can be used to deal with duplicates in the uploaded file as well).

Drop English Loans This is only applicable to Pitjantjatjara for the languages we consider. It filters entries which can readily be identified as English loan words by the presence of "From English" in the gloss field.

Light Verbs This option is similar to the 'Verb Compound' option but because some of these constructions are not separated by space or hyphen, we list out the available constructions in Pitjantjatjara to filter them out.

Verb Analysis This option only pertains to Pitjantjatjara. The reason for this is that the constituents are not marked like Warlpiri and Warumungu, and striping verbs of the suffixes yields some questionable analyses for the root. Given that it requires further input by linguists, we have instead introduced this option to provide a hypothesis that can be verified by a linguist/language expert.

4.3 Analysis

Vowel Distribution The modelled syllables are taken as the input for this function. The syllable length is calculated[3]. The syllables are sorted according to position. Vowels are counted for each syllable position (i.e., for Pitjantjatjara, Warlpiri and Warumungu {a,i,u} is enumerated, for Kaytetye {a,i,u,ə}).

This function is aimed to address the question of how vowels are distributed across different syllable positions.

Root Final Vowel Distribution This is similar to the Vowel Distribution function, except instead of sorting based on syllable position, we sort based on on word length. Here the input is both the modelled syllables and their respective lengths.

Vowel Harmony The list of predicted syllables is taken as an input for this operation. A 'syllable matrix' is constructed where each word is considered in a new row and each column represents a syllable. For example, the sound sequence *kitji*[4] would be two columns [ki] and [tji]. This extends to the maximum syllable length observed in the corpus. For shorter words, the remaining columns are left empty. Vowels are counted across each column.

For this analysis, our interface provides the option of choosing the transition between vowel one and two (V1V2), vowel two and three (V2V3) and so on.

Place of Articulation For this calculation, the language selected (to determine the possible consonants) and the 'OS' column is taken as input. Each vowel and consonant is converted to a 'V' or 'C' to construct a word template. From the word template, all VC structures are pooled together and sorted based on placement (i.e., coda or onset). Once we collect all VCs and their syllable position, we labelled the consonant according to the place of articulation. We consider five places of articulation (labial, alvelor, retroflex, palatal and velar).

Here the dashboard provides several options: to provide an aggregate count across vowel and place of articulation, a more detailed view by accounting for placement. Lastly, a frequency table of vowel and consonant combinations in all extracted VCs.

5 Conclusion

We introduce a local web-based interactive dashboard designed for targeted analysis of phonotactic patterns, and illustrate its application to four Central Australian languages. This is a customizable tool which can be adapted for a variety of search and data conditioning tasks in a wide range of linguistic data, supporting interactive analyses of morpho-phonological phenomena. The toolkit works on the principle that an iterative interactive approach is required for robust linguistically-informed processing and analysis of complex and potentially inconsistent lexical datasets, especially in corpus composition decisions.

References

Gabriele Albini, Shane Mattner, Marcel Zeuch, Arne Petter, and Joel Ostblom. 2022. The Dash Open-Source Curriculum. `https://open-resources.github.io/dash_curriculum/preface/about.html`. [Accessed 07-12-2024].

Faisal Aljasser and Michael S Vitevitch. 2018. A web-based interface to calculate phonotactic probability for words and nonwords in modern standard arabic. *Behavior research methods*, 50:313–322.

Laurence Anthony. 2012. Of software tools for corpus studies: The case for collaboration. *Contemporary Corpus Linguistics*, pages 87–104.

Laurence Anthony. 2022. What can corpus software do? In *The Routledge Handbook of Corpus Linguistics*, 2 edition, pages 103–125. Routledge.

Brett Baker. 2014. 4. word structure in australian languages. In *The Languages and Linguistics of Australia*, pages 139–214. DE GRUYTER, Berlin, Boston.

Douglas Biber. 1988. *Variation across Speech and Writing*. Cambridge University Press.

Steven Bird. 2006. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL*

[3] This can be cross-checked with the syllable counts provided by counting the number of vowels.

[4] Part of the Pitjantjatjara word for tickle: *kitji-kitjini*.

2006 Interactive Presentation Sessions, pages 69–72, Sydney, Australia. Association for Computational Linguistics.

Vaclav Brezina and William Platt. 2023. Lancsbox [software]. *online: http://corpora. lancs. ac. uk/lancsbox*.

Petra Čechová, Luca Cilibrasi, Jan Henyš, Jaroslav Čecho, et al. 2023. Introducing a phonotactic probability calculator for Czech. *Naše řeč (Our Speech)*, 1:72–83.

Wilfrid Henry Douglas, Arthur Capell, and Stephen Adolphe Wurm. 1964. *An introduction to the Western Desert language*. University of Sydney.

David Embick. 2021. The motivation for roots in distributed morphology. *Annual Review of Linguistics*, 7(Volume 7, 2021):69–88.

Janet Fletcher. 2014. 3. sound patterns of australian languages. In *The Languages and Linguistics of Australia*, pages 91–138. DE GRUYTER, Berlin, Boston.

Cliff Goddard. 1983. *A semantically-oriented grammar of the Yankunytjatjara dialect of the Western Desert language*. The Australian National University (Australia).

Cliff Goddard. 2001. *Pitjantjatjara/Yankunytjatjara to English Dictionary*, 2 edition. IAD Press, Alice Springs, NT, Australia.

Maria Gouskova. 2023. Phonological asymmetries between roots and affixes. *The Willey Blackwell Companion to Morphology*.

Heidi Harley. 2014. On the identity of roots. *Theoretical Linguistics*, 40(3-4):225–276.

Mark Harvey, Susan Lin, Myfany Turpin, Ben Davies, and Katherine Demuth. 2015. Contrastive and non-contrastive pre-stopping in Kaytetye. *Australian Journal of Linguistics*, 35(3):232–250.

Mark Harvey, Nay San, Michael Proctor, Forrest Panther, and Myfany Turpin. 2023. The Kaytetye segmental inventory. *Australian Journal of Linguistics*, 43(1):33–68.

Deborah Loakes, Andrew Butcher, Janet Fletcher, and Hywel Stoakes. 2008. Phonetically prestopped laterals in Australian languages: a preliminary investigation of Warlpiri. In *INTERSPEECH*, pages 90–93.

Jayden L Macklin-Cordes and Erich R Round. 2020. Re-evaluating phoneme frequencies. *Frontiers in psychology*, 11:570895.

Jayden L Macklin-Cordes and Erich R Round. 2022. Challenges of sampling and how phylogenetic comparative methods help: with a case study of the pama-nyungan laminal contrast. *Linguistic Typology*, 26(3):533–572.

David George Nash. 1980. *Topics in Walrpiri grammar.* Ph.D. thesis, Massachusetts Institute of Technology.

Forrest Panther. 2021. *Topics in Kaytetye Phonology and Morpho-Syntax*. Ph.D. thesis, Doctoral dissertation. University of Newcastle, NSW, Australia.

Erich R Round. 2023a. 10. segment inventries. In Claire Bowern, editor, *The Oxford Guide to Australian languages*. Oxford University PressOxford.

Erich R Round. 2023b. 11. phonotactics. In Claire Bowern, editor, *The Oxford Guide to Australian languages*. Oxford University PressOxford.

Adam Schroeder, Christian Mayer, and Ann Marie Ward. 2022. *The Book of Dash: Build Dashboards with Python and Plotly*. No Starch Press.

Jane Simpson. 2012. *Warlpiri morpho-syntax: A lexicalist approach*, volume 23. Springer Science & Business Media.

Jane Simpson. 2017. *Warumungu (Australian – Pama-Nyungan)*, chapter 32. John Wiley & Sons, Ltd.

Jane Helen Simpson. 1983. *Aspects of Warlpiri morphology and syntax*. Ph.D. thesis, Massachusetts Institute of Technology.

Holly L Storkel and Jill R Hoover. 2010. An online calculator to compute phonotactic probability and neighborhood density on the basis of child corpora of spoken american english. *Behavior research methods*, 42(2):497–506.

Marija Tabain and Andrew Butcher. 2014. Pitjantjatjara. *Journal of the International Phonetic Association*, 44(2):189–200.

Marija Tabain, Janet Fletcher, and Andrew Butcher. 2014. Lexical stress in Pitjantjatjara. *Journal of Phonetics*, 42:52–66.

Myfany Turpin. 2000. *A learner's guide to Kaytetye*, volume 2. Iad Press.

Myfany Turpin and Alison Ross. 2012. *Kaytetye to English dictionary*. IAD Press.

Michael S Vitevitch and Paul A Luce. 2004. A web-based interface to calculate phonotactic probability for words and nonwords in english. *Behavior Research Methods, Instruments, & Computers*, 36(3):481–487.

S L Wilmoth. 2022. *The Dynamics of Contemporary Pitjantjatjara: An Intergenerational Study*. Ph.D. thesis, The University of Melbourne.