Language Documentation & Conservation

Volume
19 (2025)

Pages
348 - 367

Received
22 April 2025

Handle
https://hdl.handle.net/10125/74839

Online
http://nflrc.hawaii.edu/ldc

Citation
Wilmoth, Sasha, Felicity Meakins, Cassandra Algy. 2025. Small language, big data: Building the Gurindji Kriol corpus to model the emergence of a mixed language. *Language Documentation & Conservation* 19: 348-367.

Small language, big data: Building the Gurindji Kriol corpus to model the emergence of a mixed language

Sasha Wilmoth
University of Melbourne

Felicity Meakins
University of Queensland

Cassandra Algy
Karungkarni Art

At 178 hours and 853,348 words, the Gurindji Kriol corpus (Meakins & Algy 2004) is currently the largest annotated corpus of an Australian Indigenous language, and is a significant record of the community's language use in a complex multilingual environment. Together with the Gurindji corpus, four generations of language use and change in the Gurindji community are represented, including the rare emergence of a mixed language. In this paper, we present details on the development of this corpus, in particular the complex processes of corralling this data into a consistent format that enables quantitative and computational work. The scale, breadth and consistency of the corpus has enabled innovative research into questions of language variation, contact, emergence and change; and has helped the Gurindji community to better understand linguistic changes and continuities across generations. Data-cleaning and annotation are often overlooked in discussions of data management within the field of language documentation. However, they are important steps in any quantitative research, and the amount of work required can be significantly reduced with thoughtful automation. Our approach, drawn from industry best practice, may provide a useful model for others working on the development of corpora of low-resource languages.

E-ISSN 1934-5275

1. Introduction[1]

In this paper, we report on the development of the Gurindji Kriol corpus, collected by the second and third author in Kalkaringi and Daguragu (Northern Territory, Australia) between 2004–2017 (archived in the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC); Meakins & Algy 2004). Gurindji Kriol is a recently emerged mixed language which combines the grammar and lexicon of the traditional language Gurindji with the English-based Creole called Kriol. The Gurindji Kriol corpus is currently the largest annotated corpus of any Australian Indigenous language with 178 hours of video- and audio-recording and 853,348 words transcribed. While it is a significant record of a community's language use over time and between generations, its design and scale also make it a rich dataset for addressing theoretical questions surrounding variation and change in a language contact situation. However, to explore such issues using quantitative methods, the transcription and annotation of the primary language recordings must be consistent and accurate, that is, it must be sufficiently 'clean' such that token extraction can be automated using computational methods, in a form that is appropriate for analysis. This is a greater challenge for languages such as Gurindji Kriol with no existing natural language processing (NLP) resources. Nonetheless the process of cleaning up and annotating the Gurindji Kriol corpus has reaped dividends for quantitative research on this language and for progressing our understanding of the emergence of new languages in situations of language contact (e.g., Meakins et al. 2019; Meakins & Wilmoth 2020; Sloan et al. 2022). Processing and developing the corpus required a significant investment of time and effort, and is the main focus of this paper.

Data cleaning is seldom discussed despite the fact that it is a crucial step in any corpus-based linguistic research and in commercial language technologies (although see Chua et al. 2018; Zupon et al. 2021, inter alia). This lack of detail in the literature is noted by Babinski et al (2022: 5), in a review of the usability of archival collections: "(m)any linguistic publications mention something along the lines of 'data were extensively cleaned' with no further explanations". They go on to list examples of this, before continuing:

> As these publications mention how crucial data cleaning and preprocessing were to their methods and results, it is important that these cleaning methods are transparently communicated. Not only does this aid in reproducibility and replicability of important work, good documentation of data processing is good practice. Stanley (2021) notes that the order of operations of cleaning and processing a dataset can affect the results; thus such information should be recorded and reported. Finally, if researchers are extensively cleaning datasets before use, those doing language documentation might be able to adjust methods to make the use of such collections more straightforward. (Babinski et al: 5)

This paper aims to address this lack of discussion of data processing, as part of the broader effort to make data and methods in language documentation more transparent, accountable, and reproducible (Thieberger & Berez-Kroeker 2011; Mosel 2012; Gawne et al. 2017; Berez-Kroeker et al. 2018; Mosel to appear). In the various language documentation handbooks and textbooks, and in venues such as this journal, one finds much methodological discussion around archiving, metadata, recording techniques, citing primary data, and so on; even though these issues may vary from project to project, there is broad agreement on best practice. On the other hand, there is very little discussion of data cleaning and processing methods. This may be because data cleaning is considered 'hacky and atheoretical' by some (as a colleague recently put it to us informally), as it can be more idiosyncratic to a specific project, file format, or language. We do not believe that this is inevitable. While some of the details of our workflow are specific to Gurindji Kriol, our research questions, or the CHAT format (discussed in §3.1), we hope that our overall approach may be useful to others working with corpora of under-documented languages, as language documentation and linguistic typology become increasingly data-driven and corpus-based (see e.g., Levshina 2022; Schnell & Schiborr 2022), particularly in their interfaces with computational linguistics and sociolinguistics (e.g., Drager 2018). Particularly, we wish to emphasise three main takeaways which are applicable to any project, and which are discussed in §5:

[1] We are grateful for the support of many individuals and organisations who contributed to the development of the Gurindji Kriol corpus. We acknowledge all of the Gurindji Kriol speakers who have contributed to the corpus. Transcription and coding was contributed by Jackie van den Bos, Kate Charlwood, Hannah Choi, Elizabeth Hall, Jessica Solla, Vivien Dunn and Bodean Sloan. Technical contributions were made by Lucy Andresen, Simon Hammond, Romi Hill, Tom Honeyman, Nay San, Jonathan Taufatofua and Ola Olsson. This work has been funded and supported by the Australian Research Council (DP0343189, DP0985024, DE140100854, CE140100041), the Hans Rausing Endangered Languages Project (IPF0134), DOBES, the University of Queensland and Appen. We also thank two anonymous reviewers for their helpful comments.

1. Take a bird's eye view of the data.
2. Document everything.
3. Undertake iterative validation.

We begin this paper with an overview of the structure of Gurindji Kriol and the sociohistorical circumstances of its emergence. We also provide an overview of the contents and availability of the corpus. In §3, we describe our data-cleaning and annotation workflow, including metadata, transcription, morphological annotation, tagging of subject NPs, conversion between formats, and preparation for archiving. §4 discusses some of the research that the corpus has enabled.

2. Corpus description

2.1. History and sociolinguistic setting of Gurindji Kriol

Gurindji Kriol is a mixed language spoken in northern Australia in the twin communities of Kalkaringi and Daguragu. It is also spoken further north in Nitjpurru (Pigeon Hole) and Yarralin, which are home to Bilinarra and Ngarinyman people, respectively (see Figure 1). Kalkaringi and Daguragu are only eight kilometres apart with shared kin relations and administration, so they essentially function as a single community (Meakins 2008: 300). The name 'Kalkaringi' is used in this paper to refer to both communities.

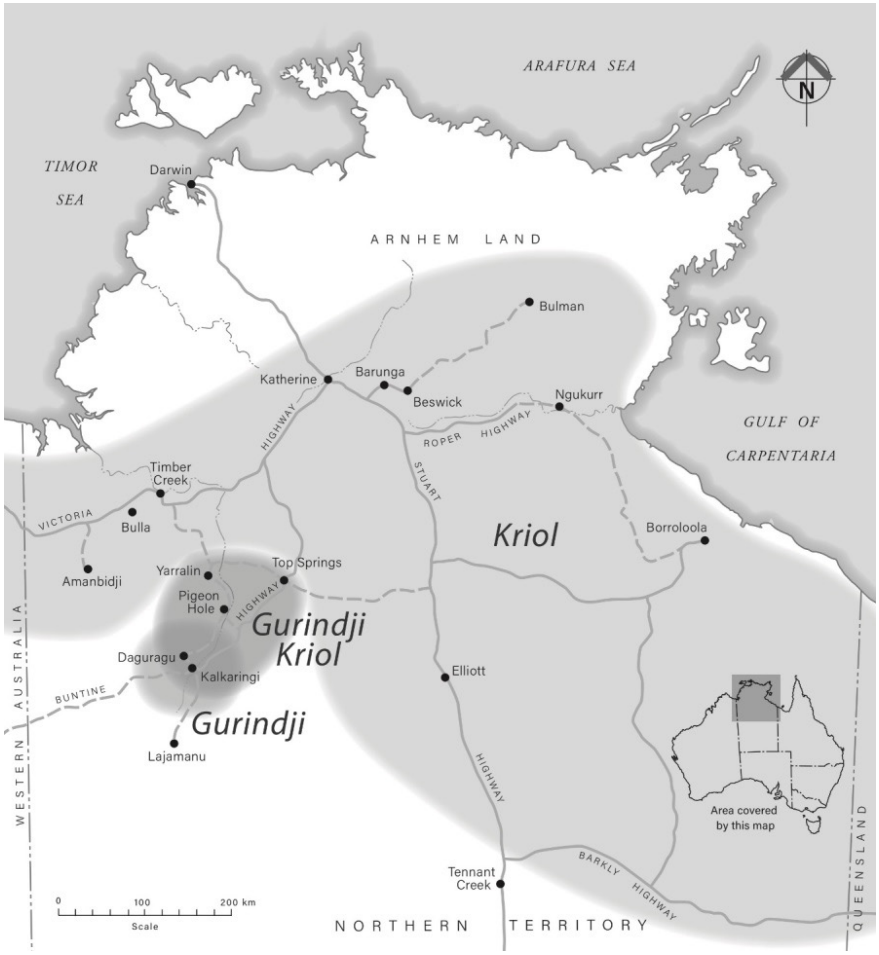

Figure 1. Areas in northern Australia where Gurindji, Kriol and Gurindji Kriol are spoken (Meakins et al. 2019)

Gurindji Kriol is derived from Gurindji (Ngumpin-Yapa, Pama-Nyungan), the traditional Australian language of the region, which is also closely related to Bilinarra and Ngarinyman; and Kriol, the English-lexifier Creole language spoken across much of northern Australia. Gurindji Kriol combines the lexicon and structure of these two languages. The noun phrase structure including case markers and derivational morphology largely derives from Gurindji with the verb phrase structure, including the TAM auxiliaries, such as *bin* 'past', originating in Kriol. An example of Gurindji Kriol is illustrated in (1) (with bolded elements of Kriol origin).

(1)	***Dat***	*warlaku-ngku*	***bin***	***bait-im***	***dat***	*marluka*	***fut****-ta*
	the	dog-NOM	PST	bite-TR	the	old.man	foot-LOC
	'The dog bit the old man on the foot.'						(Stewart & Meakins 2021: 74)

Gurindji Kriol derives from code-switching practices. In the 1970s, code-switching between Gurindji and Kriol was observed to be the dominant language practice in Kalkaringi (McConvell 1988). Over time, this pervasive code-switching eventually stabilised into an autonomous mixed language variety (McConvell & Meakins 2005; Meakins 2011a). The recent history of Gurindji people also provides a sociopolitical context for the emergence of the mixed language. The post-invasion period of Gurindji history is one of the better-documented periods of Australian history (Hardy 1968; Charola & Meakins 2016; Ward 2016). This is likely due to the Gurindji people's impact on Australian politics. The Gurindji people gained significant attention for their nine-year workers strike between 1966 and 1975, which eventually culminated in the transfer of the land at Wave Hill Station back to the Gurindji people in 1986. This was the first successful land claim for an Aboriginal group in Australia, and it generated a great deal of momentum for the broader Australian Indigenous land rights movement. The Gurindji people became famous for their determination during this time, and their eventual triumph further solidified the strength of the Gurindji identity. This strength of identity has probably driven the maintenance of many elements of the Gurindji language, despite Kriol becoming increasingly dominant in the area, superseding many traditional languages, and coming to represent a pan-north Australian Aboriginal identity (Meakins 2008: 86). These sociopolitical and linguistic factors provided the motivation for the emergence of Gurindji Kriol.

2.2. Corpus collection and design

The collection of Gurindji Kriol audio-visual recordings took place between 2004–2017. The recordings were made during the the Aboriginal Child Language (ACLA) project from 2004 to 2007,[2] the Jaminjungan and Eastern Ngumpin DoBeS project from 2007 to 2008,[3] a Hans Rausing Endangered Languages Project from 2008 to 2010,[4] and two Australian Research Council fellowships from 2009 to 2017.[5] These projects focussed on adult and children's Gurindji Kriol. The third author (Algy) was instrumental in the collection of these recordings. 122 speakers were recorded during the full 14-year period including the same children at different ages. In all, the Gurindji Kriol corpus consists of 178 hours of transcribed and annotated recordings from 122 speakers. Altogether, 146,933 utterances were transcribed and annotated (159,758 clauses, 853,348 words, 1,054,767 morphs). Transcripts and recordings from all the aforementioned projects are housed in PARADISEC (Meakins & Algy 2004).[6] Access is via Meakins and in line with instructions given by Gurindji participants in the ethics clearance process. The video recordings remain inaccessible to all but participants, per community preference.

[2] ARC Discovery Project DP0343189 2003–2007.

[3] https://dobes.mpi.nl/ projects/jaminjung/

[4] IPF0134; available in the ELAR archive—https://elar.soas.ac.uk/deposit/0273

[5] ARC APD fellowship 2009 to 2012 (DP0985024) and ARC DECRA fellowship from 2014 to 2017 (DE140100854).

[6] https://catalog.paradisec.org.au/collections/GK1, DOI: 10.26278/CD9K-3Y70

The recordings consist of conversation; picture-prompt narratives (e.g., Frog stories (Mayer 1994) and the Monster stories series (O'Shannessy 2004)); and picture-prompt sentences which come from a series of director-matcher tasks designed to target particular linguistic features (Meakins 2011a). Around 50 hours of the corpus consists of conversation, and the remaining consists of picture-prompt tasks. The picture-prompt tasks require speakers to express the characters, objects, and actions depicted by the pictures. All characters, objects, and actions occur at least three times (in most cases many more) in the pictures, guaranteeing the multiple use of a Kriol, Gurindji and/or innovative variant(s) by a speaker to express a particular linguistic feature. This is important for variationist research.

The director-matcher tasks were conducted as a series of games (picture-match, bingo, or 'go fish') (see Meakins 2009; Meakins & O'Shannessy 2010; Meakins 2011a: 50–54; 2011b; 2016; Meakins & Algy 2016; Meakins & Wilmoth 2020 for a detailed description of the tasks). In total, 148 pictures were used throughout the different tasks. The director and matcher sat near each other with matching sets of pictures in front of them. The director described a picture, which differed only minimally from another picture, in such a way that the matcher could identify the correct picture. The director and matcher's pictures were screened off from each other to ensure that the director was required to produce detailed descriptions. Director-matcher tasks have been used extensively in semantic typology research, such as in studies of spatial relations and reciprocals (Levinson & Wilkins 2006; Evans 2011), to produce comparable data across languages.

More generally, we also adopted the peer elicitation approach (cf. Meakins et al. 2018) to ensure that the data captured were as close to everyday speech as possible. Peer elicitation involves replacing the usual linguist-speaker dyad with peer-peer dyads or groups. In the case of elicited narrative descriptions using picture books, the person giving the instructions came from the same generation. In the case of director-matcher tasks, both the director and matcher came from the same generation. This ensured that the data was not skewed by adjustment of language use to accommodate a Gurindji person of another generation (or, more importantly, an English-dominant researcher). For example, a Gurindji speaker might accommodate for the presence of children or an English-dominant researcher by using more Kriol.

The corpus was transcribed in the CHAT format, although it is archived in CHAT, ELAN, and Toolbox formats (see §3.5.3). The CHAT format is essentially a plaintext format with particular formatting conventions to make it readable by the CLAN software.[7] This is described in more detail in §3.1. Meakins transcribed the recordings made between 2004–2010. Later, University of Queensland Summer Scholars transcribed the remaining corpora in three Summer Research programs: Jackie van den Bos, Kate Charlwood, Hannah Choi, Elizabeth Hall, Jessica Solla, Vivien Dunn, Bodean Sloan, and Sasha Wilmoth. Wilmoth continued to work on the collection through an Appen/CoEDL[8] internship (2016) and then a CoEDL-funded fractional research assistant position (2017). It was during this time that the data cleaning protocols were finessed by Wilmoth with Meakins. Further work by Romi Hill and Lucy Andresen supervised by Wilmoth was supported by Appen through their CoEDL partnership (2020). This was necessary as another batch of data was incorporated at a later stage in order to ensure the collection archived in PARADISEC represented the full collection; this data was recorded during the aforementioned projects but had not previously been processed to the same degree.

3. Data cleaning and annotation

In this section we describe the processing of the Gurindji Kriol corpus data to the point where quantitative and computational work becomes feasible. Processing here refers to ensuring the consistency and accuracy of the metadata, transcription and all annotations such as interlinear glossing, part-of-speech tagging and, in our case, identification of all subject NPs. While the corpus includes other types of annotation such as time-alignment and translation, these were

[7] CLAN (Computerized Language ANalysis; MacWhinney 2014) is a program which has been widely used to transcribe and analyse child language data, particularly in conjunction with the CHILDES (Child Language Data Exchange System) corpora. ELAN (2024, version 6.8) has since become the popular choice, and thus we chose to archive the corpus in both formats, with Toolbox format (Buseman 2019) being an intermediate step in the conversion process (see §3.5.3).

[8] Appen is a company, headquartered in Sydney, that develops speech and language data for artificial intelligence/machine-learning technologies such as automatic speech recognition. Wilmoth was employed as a linguist at Appen between 2015–2018. CoEDL refers to the Australian Research Council Centre of Excellence for the Dynamics of Language (2014–2022; CE140100041), of which Appen was an industry partner. While Appen supported the work described in this paper, they hold no rights to the data or other intellectual property produced as part of this project.

not integral to the research questions being asked of the corpus and were not systematically checked. In more general terms, data cleaning includes the following steps (after Maletic & Marcus 2005):

1. Define and determine error types.
2. Search and identify error instances.
3. Correct the errors.
4. Document error instances and error types.
5. Modify data entry procedures to reduce future errors.

Data cleaning and annotation at scale can be a formidable challenge, particularly for languages such as Gurindji Kriol which are under-resourced (in terms of available NLP resources). There were also additional challenges arising from the fact that the corpus was transcribed over many years by many different people, as well as some challenges specific to the CHAT format.

Much of the work involved in cleaning the data was undertaken while Wilmoth was employed at Appen; the approach taken for the Gurindji Kriol corpus was directly inspired by industry practice. However, as Wilmoth is a linguist by training and not a computer scientist or programmer, the computational skills that were employed are not highly sophisticated and do not require any specific training beyond a general familiarity with the Unix shell (the command line interface accessed through a Terminal), regular expressions, and a text editor such as Vim; Python was learned throughout the process of cleaning up the corpus, using free online courses, and some advice from colleagues. Python scripts written specifically for this project are available online.[9]

The CHAT files were mostly handled in the Vim text-editor or via the command line (that is, typing commands into a Terminal, rather than using a graphical user interface) using basic text processing tools such as grep and sed.[10] Custom Python scripts were written as needed when other basic command line tools did not suffice. The use of these tools confers a high degree of flexibility and efficiency, and has the benefit of working over multiple files at once instead of opening them up individually. In the final stage of converting between CHAT and other transcript formats, the open-source data-wrangling application OpenRefine became very useful (see §3.5.3).[11]

Where files needed to be worked on directly, this was typically done by collating all CHAT files in the corpus into a single text file using grep:[12]

(2) `grep '' */*.cha > all_files.txt`

The output has the original file path at the beginning of each line, and can then be worked on in Vim, such as running complex find-and-replace commands (at times globally, at other times checking each instance) or other error correction. This saves an enormous amount of time and ensures that mistakes can be reversed. A Python script is then used to split this file back up into individual sessions in their original locations (with an option to replace the original files or not).

[9] https://gitlab.com/swilmoth/Gurindji-Kriol-Corpus-Scripts

[10] Grep and sed are standard text-processing utilities in Unix/Linux and Unix-based operating systems such as MacOS. Grep is used for searching text for matches based on regular expressions; sed is used for editing text, including using regular expressions for find-and-replace operations. There are several free courses and guides online that cover command line basics, bash scripting (for automation), and text-processing using Unix (including regular expressions, text editors such as Vim/emacs, and text-processing utilities such as grep, awk and sed).

[11] https://openrefine.org/

[12] This command runs a search with grep, using an empty regular expression enclosed in apostrophes which matches every line in text files. This search runs over `*/*.cha`; that is, all files with the .cha extension located in all subdirectories of the current directory. The output of this search (comprising the full corpus excluding audiovisual data) is redirected to a single text file (`> all_files.txt`).

3.1. The CHAT format and the lexicon files

Before discussing specific steps in our data cleaning process, this section describes the technical aspects of the corpus. While the collection is archived with transcripts in several formats (CHAT, ELAN, and Toolbox), it was transcribed in the CLAN software, which uses the CHAT format. Alternative formats were only possible as a result of this data cleaning process, and the conversion is described in §3.5.3. The CHAT format, now a less common choice, has many affordances as well as drawbacks. Apart from linking in with the CHILDES database (MacWhinney 2014) and CLAN's built-in search functions, the main benefit of CHAT is that it is a plain text format and therefore easily manipulated via basic scripts and command line tools. This is in contrast to XML-based formats such as ELAN files, which are less easily manipulable with simple Python and command line tools.

The CHAT format as we use it (see MacWhinney 2019 for comprehensive documentation) consists of a metadata section and the transcript itself, which consists of transcription tiers (beginning with '*' and speaker initials), morphological annotation tiers (beginning with '%mor'), translation tiers (in our case beginning with '%eng') and comment tiers ('%com').

The metadata tiers in the Gurindji Kriol corpus were initially generated from a metadata spreadsheet by a Python script (generate_headers.py by Jonathan Taufatofua), however, errors are easily introduced as the file is further worked on. An example of the metadata tiers is shown below, with personal names redacted.

(3)
```
@UTF8
@Color words:     *FSE: 1 0 32767 0
@Begin
@Languages:       gu
@Participants:    FSE S____ Speaker, FCA Cassandra Investigator, FFM
                  Felicity Investigator
@ID:              gk|meakins|FSE|25;00.00|female|||Speaker|||
@ID:              gk|meakins|FCA||female|||Investigator|||
@ID:              gk|meakins|FFM||female|||Investigator|||
@Media:           FM16_78_1b, audio
@Location:        U_____'s house (Daguragu). S_____telling Horse and Cow
                  story to her youngest daughter A______. Other kids are
                  also present. 1:09min
@Date:            25-MAY-2016
```

An example utterance is shown in (4).

(4)
```
*FJA:   im _in wirr _im _an nyanuny shirt _nganyjuk . •169606_171435•
%mor:   pro|i&k=3SG\P suf:tam|_in&k=PST v:tran|werrim&k=wear suf|_
        im&k=TR suf:tel|_an&k=on pro:poss|nyanuny&g=3SG/DAT n:inani-
        mate|shirt&k=shirt
```

Note that the utterance is pre-tokenised, with clitics/affixes separated from stems by a space and an underscore. That is, transcribers marked morph boundaries as part of the transcription process, typing 'bolou _im _bat' instead of 'bo-

louimbat' (follow-TR-CONT). Punctuation marks are also separate tokens. Transcribers used a dual orthography where Gurindji-source elements were written using the Gurindji orthography and Kriol-source elements written using a variation of the Northern Territory Kriol orthography. The transcription tier also includes timestamps. These consist of the start and end time of the utterance, in milliseconds, separated by an underscore, and bracketed by Unicode control characters (U+0015, displayed as '•' in the examples in this paper). The mor-tier includes one mor-code corresponding to each token on the transcription tier. The mor-codes contain the following information, with optional elements in parentheses:

(5) PartOfSpeech(:SubType)|CitationForm(&Reduplication)&Language=Gloss

These codes are generated automatically from lexicon files; we used separate lexicon files for words of Gurindji versus Kriol origin (for incidental reasons due to the history of various documentation projects on these languages). For example, *wirr* on the transcription tier corresponds to `v:tran|werrim&k=wear` on the mor-tier, which was generated from the following entry in the lexicon file:

(6) `wirr {[scat v:tran]} "werrim&k" =wear=`

The mor-code in this instance tells us that this is a transitive verb of Kriol origin, meaning 'wear'. These lexicon files are stored with project scripts and resources but are not currently archived with the corpus.

3.2. Correcting metadata

As seen above, the metadata section contains a considerable amount of detail in a specific format. As the files were worked on by multiple people over several years, formatting errors were easily introduced. These errors were identified by using grep to extract metadata tiers in various ways. For example, the following command produces a list of tier headings which can be used to identify errors (such as whitespace before the '@' symbol, singular 'Language' instead of 'Languages', semicolons instead of colons, mixing tabs and spaces, and so on):[13]

(7) `grep '^\s*@' | cut -f1 | sort -u > metadata-headings.txt`

Errors can then be corrected across the corpus with regular expressions in Vim or with sed. Similar processes were applied to the basic formatting of all tier types before further correcting the content of transcripts.

3.3. Spelling and transcription

It is a much more complex task to identify and correct errors in the body of the CHAT file. In this section we describe error identification and correction in the transcription tier, while the next section describes error identification and

[13] The pipe ('|') is one of the main affordances of working via the command line; it chains together several commands such that the output of one is 'piped' to the input of the next. In this example, we first run a regular expression search with grep. The output of this is then 'cut' to include the first column (of a tab-delimited file), and then sorted into an alphabetised list of all unique metadata tier headers. This output is redirected to a new text file.

correction in the morphological annotation tier, as well as correction of the two tiers in tandem.

One advantage of the design of the corpus in the CHAT format is that morphological segmentation was done at the transcription stage. This obviates any need to segment morph boundaries automatically. The already segmented transcripts could then be easily checked against the lexicon files, which includes lexical stems as well as clitics and affixes (including all allomorphy). The goal of this process was to ensure that each token in the transcription tier was correctly spelt according to Gurindji and Kriol orthography, that it matched an entry in the lexicon file, and that some CHAT-specific conventions were followed (e.g., separating punctuation marks from surrounding tokens with space, ensuring that reported speech is introduced with a +" symbol). At this stage, we were also able to correct known misanalyses in word/morph boundaries.

This process consisted of an iteration between two scripts, findUnknownTokens.py and correctCHATSpelling.py. The first script compares a list of all transcribed tokens to the lexicon file(s) and outputs a file with two identical columns ready for correction, containing all tokens not found in the lexicon. The following example (8) is an abbreviated version of the full output. Note that at the bottom of the output are some words that had previously been corrected. The script takes a 'checked words' file as an optional extra argument. This means that when running the script a second, third or fourth time, you do not have to correct the same errors multiple times.

(8)
```
_rayinyj      _rayinyj
ppuntanup     ppuntanup
-aran         -aran
xx            xx
playing       playing
footy         footy
Kajirrii      kajirrii
lajamanu      lajamanu
writing       writing
_waija        _waija
ayglas        ayglas
_bat_karra    _bat_karra
wood          wood
marlaku       marlaku
making        making

=== The following tokens are already in the 'checked words' file.===
===They will be corrected if you run correctCHATSpelling.py. ===
kangkaru      kengkaru
shoulda       sholda
```

The next step would then be to go through this file and for each token either correct the second column, correct the token in context and remove it from the list, or add it to the lexicon and remove it from this list. This is illustrated below.

(9)

_rayinyj	_ranyj
ppuntanup	puntanup
-aran	_aran
xx	xxx
playing	plei _ing
~~footy~~	~~footy~~
Kajirrii	kajirri
lajamanu	Lajamanu
writing	rait _ing
_waija	_walija
ayglas	eyeglas
_bat_karra	_bat _karra
wood	wud
~~marlaku~~	~~marlaku~~
~~making~~	~~making~~

Add to the lexicon

Check in context: either *marluka* 'man' or *warlaku* 'dog'

Check in context: either *makin* 'to sleep' or *meik _ing* 'making'.

```
=== The following tokens are already in the 'checked words' file.===
===They will be corrected if you run correctCHATSpelling.py. ===
kangkaru      kengkaru
shoulda       sholda
```

We would then run correctCHATSpelling.py to correct the corpus automatically. Only the transcription tier is affected by this script. This is an iterative process; before moving on to morphological annotation, we would run this pair of scripts several times until the output generated by findUnknownTokens.py was empty. A crucial part of this process was that the cumulative 'checked words' file became a valuable dataset that is kept with all other scripts and project resources. This saved time when new data was later processed.

Several errors had also been introduced into the timestamps and had to be corrected; this was done using grep and regular expressions to fix formatting errors, and Python to identify start times that were after the end times.

3.4. Morphological annotation

Once transcription of the files was complete, a tier with morphological annotations (the mor-tier) was generated by CLAN, based on a separate lexicon file or files. An example of this is repeated below from (4).

(10)
```
*FJA:   im _in wirr _im _an nyanuny shirt _nganyjuk . •169606_171435•
%mor:   pro|i&k=3SG\P suf:tam|_in&k=PST v:tran|werrim&k=wear
        suf|_im&k=TR suf:tel|_an&k=on pro:poss|nyanuny&g=3SG/DAT n:in-
        animate|shirt&k=shirt
```

While the mor-tier was automatically generated, homophonous morphs required disambiguation; in §3.4.1 we discuss how this process was partially automated. As well as disambiguation, much of this mor-coding required correction during the data cleaning phase of the project, as described in §3.4.2. This newer data was subject to the same correction processes as the older data. Finally, for a study of variable subject marking (Meakins & Wilmoth 2020), all subject noun phrases were tagged; the process of automating this as much as possible is described in §3.4.3.

3.4.1. Disambiguation

As mentioned above, disambiguation of mor-codes was required for all homophones. For example, *langa* can be a locative preposition (from Kriol) or a noun 'ear' (from Gurindji). In such cases, CLAN generates codes for all possible interpretations, separated by a ^ character: `prep|la&k=LOC^n:bp|langa&g=ear`. A very high proportion of morphs required disambiguation. In one file chosen at random (FM12_a180), 76 out of 341 morphs (22%) required disambiguation. There is an option to correct these within the CLAN software, and this was done for much of the corpus during the University of Queensland Summer Research Programs. However, in the absence of a hand-coded CLAN grammar for the language, this is a very repetitive and slow process. For the newer data coded during the data-cleaning phase, disambiguation was done using regular expressions in Vim. Most regular expressions were applied globally, while others required manual checking of each example; some examples are detailed in Table 1. This reduced the amount of manual checking by 40–50%. Even for those morphs/contexts which required manual checking, it was significantly more efficient to address each morph/context across the whole corpus, instead of going through every morph/context in each file in chronological order. The regular expressions, and the order in which they were applied, were thoroughly documented.

Table 1. Examples of common ambiguities in morphological tagging

Context	Global
A verb followed by the transitive suffix *-im* is transitive (for verbs such as *barn* 'burn' which can be either transitive or transitive).	Yes
Ting 'whats-it-called' preceded by the demonstrative *dat*, any preposition, or a possessive pronoun is a noun (and not a verb 'do whats-it-called').	Yes
Garra followed by a noun is a preposition 'PROP', and not an auxiliary ('will') or transitive main verb ('have').	Yes
Any stem followed by a telic or continuative suffix is a verb, not a noun.	Yes
Si is usually a transitive verb 'see' in the data, and not a noun 'sea'.	No
Plei 'play' followed by an inanimate noun is usually a transitive verb. (In some cases it may be intransitive, if the noun is followed by the proprietive *-jawung*.)	No

3.4.2. Correction

There were several types of errors in the mor-coding that needed to be addressed:

1. Introduced typographical errors in the mor-codes (a hazard of the plain text format), or older mor-codes that no longer reflected updates to the lexicon files.
2. Tokens which were not in the lexicon at the time of adding the mor-codes in CLAN. These were automatically given codes such as `?|ngakpran` (in this example, the transcriber has mis-spelled ngakparn 'frog'). While the spelling in the transcription tier was resolved in the previous stage (§3.3), the mor-coding was not.
3. Mismatches between transcription and mor-tier. At some point in the lifecycle of the corpus, errors arose where the mor-codes may have matched some entry in the lexicon, but do not match the transcription.
4. Insertion of placeholder tokens on the mor-tier. CLAN does not add a token to the mor-tier for every token on the transcription tier. For example, 'xxx' (used to transcribe inaudible speech), English insertions (which are indicated in the transcription with an ampersand, e.g., '&bought'), and some punctuation marks did not have any corresponding token added in the mor-tier. Insertion of placeholder tokens was required both for correcting errors of type 3, as well as for converting to other formats (see §3.5.3). This is a drawback of the unstructured, plaintext format, as there is no explicit link between tokens on different tiers.
5. Misanalyses and mistakes in disambiguation.

Error types 1 and 2 were corrected in a similar iterative manner to transcription errors, as described above in §3.3, using validateMorCodes.py and correctMorCodes.py. Once all mor-codes were validated against the lexicon (i.e., a mor-code `suf:num|_nganyju&g=GROUP` has a correctly corresponding lexicon entry `_nganyjuk {[scat suf:num]} "_nganyju&g" =GROUP=`), we addressed errors of type 3 and 4. The first step in this process was to check that the number of tokens on each tier matches (taking into account tokens expected to be missing such as 'xxx' etc., as described above). We did so by running countTokens.py, and corrected errors manually. At this point, we could be sure that all spelling and mor-coding was valid according to the lexicon, and that there was the same number of tokens on each tier. However, there were still cases where a particular transcribed token does not have the appropriate corresponding mor-code, even though both tiers may seem valid on their own. These errors were uncovered and corrected with two Python scripts (checkMorcodeMatches.py and correctMismatches.py), in a similar iterative fashion to other steps. Subsequently, insertPlaceholders.py adds tokens to the mor-tier corresponding to 'xxx', punctuation marks and so on.

Type 5 errors were the most difficult to uncover, as some mor-codes matched both the transcription and lexicon file perfectly while still being incorrect. To identify as many of these as possible, morChecker.py (by Simon Hammond) was used. This script identifies sequences of mor-codes of a given length (n-grams) which have the same sequence of lemmas but have a difference in some other morphological information (sentence boundaries are also counted as tokens by this script). We began with longer n-grams because, though the output was smaller, this was more likely to represent errors rather than genuine morphosyntactic differences. For example, running the script with the length set to 9, the output was minimal, identifying only two occurrences of '*START dat tri i bin baldan la dat kajirri*' ('the tree fell down on the woman'). In one instance *baldan* 'fall down' had been correctly coded as an intransitive verb, while in the other it was incorrectly coded as a transitive verb. The mor-codes and transcriptions for these sequences were identical other than this coding error. For shorter sequences, such as 3- or 4- grams, the output is too verbose to check through entirely. However, likely errors can be identified according to the relative frequency of each morphological analysis. For example, the script identified 65 instances of the trigram *en dat boi* 'and that boy' where *boi* was coded as a noun, and one anomalous instance where *boi* was coded as an adjective. This process allowed us to identify and then manually correct likely errors such as these, significantly improving the quality of morphological annotation.

3.4.3. Coding subject noun phrases

Once the morphological annotation had been cleaned and checked, the next step in our workflow was to code all subject noun phrases. This is not a feature of CLAN but was a necessary step in coding the dependent variable in a study of variable subject case marking (Meakins & Wilmoth 2020). While Gurindji has an obligatory ergative marker on transitive subjects, the alignment system of Gurindji Kriol has shifted to a nominative-accusative system (Meakins 2009; 2011a; 2015; Meakins & O'Shannessy 2010), with case marking on subjects now optional. An example of this is shown in (11), with italics indicating Gurindji origin elements, and the subject noun phrases in bold.

(11) a. **Dat** ***mukmuk*** bin *kuli* la=im.
the owl PST attack OBL=3SG.O
'The owl attacked (the boy).'

b. **Dem** **bi-*ngku*** *kuli* la=im dat *warlaku=warla.*
those bee-NOM attack OBL=3SG.O the dog=CONTR
'(And) the bees really went for the dog instead.' (Meakins 2009: 82)

This study examined the factors conditioning this variation across 3,575 subject tokens. These were identified as subjects from a total of 15,618 candidate nouns and 10,272 candidate pronouns (a large amount of new data was added to the corpus at a later date, also requiring a similar amount of coding and cleaning). This volume of data is too great to inspect each noun phrase manually. Thankfully, it was possible to automate a large proportion of the subject coding. This was done with regular expressions in Vim, first adding a subject candidate tag to each noun phrase on the mor-tier, before successively removing these tags, or changing them to a confirmed subject tag. In certain cases, noun phrases were assumed to be subjects and were tagged automatically (such as when they have the erstwhile ergative suffix *-ngku* or allomorphs thereof). In other contexts, they were assumed to never be subjects (such as noun phrases following prepositions) and the subject candidate tag was removed. A series of increasingly specific contexts were automatically tagged or untagged, leaving approximately 3,200 noun phrases to be checked manually. While still requiring significant time and effort, this is approximately an 87% reduction in the amount of manual checking required. We return to the issue of subject-tagging in §4, as the coding done according to this process was significantly more accurate and this led to different findings compared to previous research on the topic by Meakins. This shows that a more methodical, semi-automated approach to data annotation has benefits for linguistic analyses, not only for reducing manual effort in preparing data for analysis.

3.5. Preparation for archiving

With the corpus fully cleaned up, all that remained was to prepare the data for archiving. There are two versions of the corpus. The community-facing version includes video, audio, and transcripts/metadata with personal names. The version used for research is audio only, with all names redacted from transcripts and metadata (but not audio). This latter version is archived in PARADISEC with access managed by Meakins.

3.5.1. Anonymisation of transcripts

In order to anonymise the transcripts, all words in the CHAT files beginning with a capital letter were extracted and sorted through manually. A sed script was run over the corpus to replace these across all tiers including transcription, translation and metadata. This sed file was simply a list of regular expressions similar to the following made up example.

(12)
```
s/\<Anna\>/NAME/g
s/\<Anne\>/NAME/g
s/\<Andrew\>/NAME/g
s/\<Anton\>/NAME/g
```

3.5.2. Aligning audio and video

It is common practice to record audio and video on separate devices; this often results in them being out of sync. While professional video editing software often includes an automatic audio syncing function, this would be very time-consuming to run over many hundreds of sessions. An additional problem arose when the video recording was switched on earlier than the audio recording. The transcription was aligned to the audio recording, so if the audio and video are synced, they will no longer match the timestamps of utterances in the transcript. To overcome these problems, a Python script (av_align.py by Ola Olsson) was run across the corpus. This script (based on ffmpeg, a software package used for handling video and audio files; Tomar 2006) calculates the most likely offset between the audio and video files, and adds either silence or blank frames to the beginning of the record as required. Crucially, it also adjusts the timestamps of a CHAT (or ELAN) file in the same directory. Unfortunately, this script has been increasingly difficult to run in recent years, due to the change to Python 3.0 (the script was written in Python 2.7), and various package updates and incompatibilities. A Docker container[14] was set up by Nay San at the time; this has not been tested recently but may still work.

3.5.3. Conversion between formats

CHAT was chosen as the format for the Gurindji Kriol corpus as it originated in the ACLA project which was a child language project, and CLAN has been widely used in child language research. However, as ELAN has now become the more popular choice in the field of language documentation, we decided to archive both formats for the sake of future reuse and interoperability. Unlike the CHAT format, the ELAN format is a highly structured XML document, with explicit links between annotations on different tiers. Figure 2 shows a screenshot of the CHAT file in Vim, while Figure 3 shows the same utterance in ELAN.

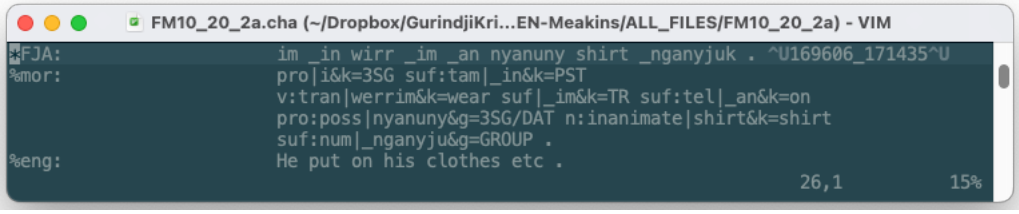

Figure 2. Example utterance in CHAT format (displayed in MacVim)

[14] https://coedl-knowledgebase.readthedocs.io/en/latest/posts/20170509_av_align-intro-tutorial/index.html

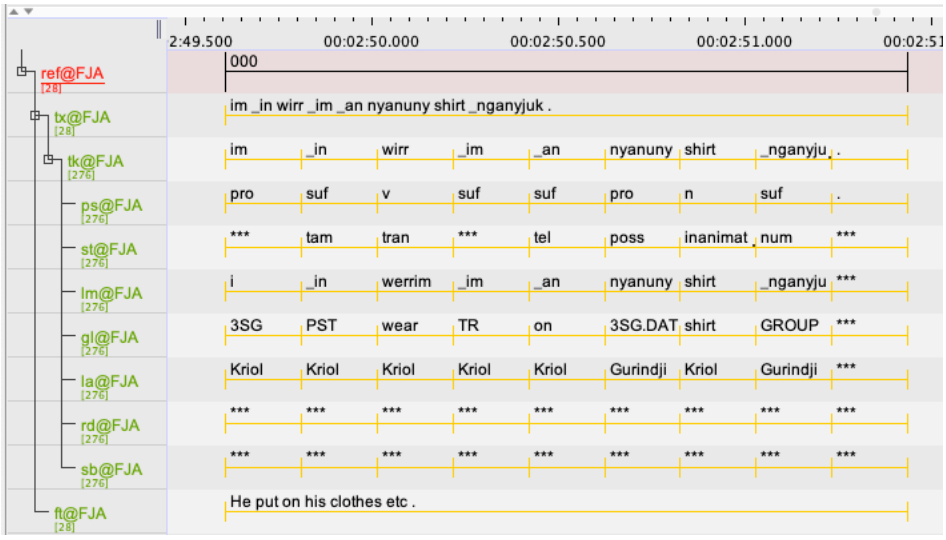

Figure 3. Example utterance in ELAN format

While ELAN does have an option to import CHAT files, this simply copies the mor-code verbatim onto a single tier and does not split up the different types of information across multiple tiers (which is one of the main affordances of ELAN). To overcome this limitation, we converted the CHAT files to Toolbox format first as an intermediate step. Converting to Toolbox first required disentangling the multiple types of information in the mor-codes. This was done using OpenRefine, a free, open-source software which is extremely useful for creating reproducible data-processing workflows. Tom Honeyman then assisted with devising a method to automatically export the corpus from the OpenRefine database into the Toolbox format. Figure 4 shows the same utterance in this format.

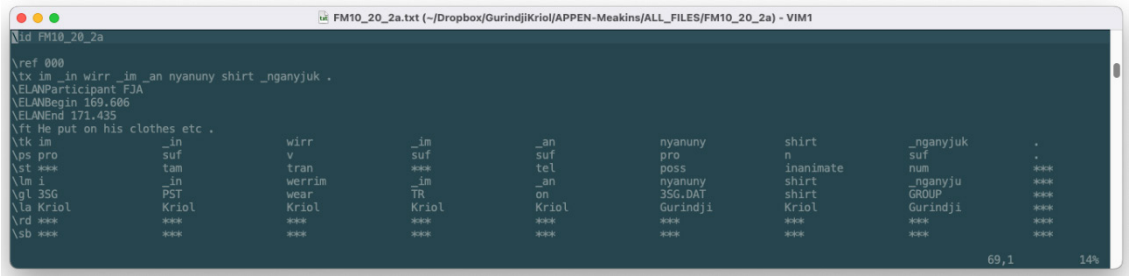

Figure 4. Example utterance in Toolbox format (displayed in MacVim)

Finally, the corpus was imported en masse into ELAN from this Toolbox format. None of this would have been possible without exceptionally consistent and clean data in the transcription and mor-tiers.

4. Enabling new research

The process described above represents a significant investment of time and effort (and therefore money). Although having squeaky clean data may be satisfying in and of itself, this investment is only truly worth it if it enables research that would not otherwise be possible. In this case, the investment in data processing has opened up several new lines of inquiry, and led to the development of new methodologies to address research questions that were previously intractable. In this section, we describe some of the research made possible by the scale and consistency of the corpus.

A series of studies by Meakins (linguist), Algy (field assistant), Lindell Bromham (evolutionary biologist), and Xia Hua (mathematician) are the first investigation of contact-induced change within a single speaker population that uses multiple variants (Meakins et al. 2019; Bromham et al. 2020; Hua et al. 2021; Hua 2022). This would not have been possible without a consistently coded Gurindji Kriol corpus. Hua (2022) developed the R Package BayesVarbrul which enables innovative multivariate analysis of language change. Specifically, it combines Wright–Fisher models[15] and the logic behind Varbrul[16] analyses under a Bayesian framework in order to investigate temporal change in linguistic data. The papers track changes in lexicon and grammar over three generations of Gurindji people, using data from 78 speakers across 120–185 variables, depending on the paper.

Another avenue of research has been determining whether changes across the first and second generations of Gurindji Kriol speakers are the result of a generational shift or an extended acquisition process. For example, Sloan et al. (2022) analysed the production of Gurindji-derived and Kriol-derived nouns and verbs in the speech of five different age groups of Gurindji Kriol speakers: children aged 5–9, 10–11, 12–14, and 15–17 years; and adults aged 21–35 years. This data is drawn from a subset of the Gurindji Kriol corpus consisting of 176 picture-prompt narratives from 70 different speakers, recorded between 2006 and 2017. The production of nouns and verbs is analysed both synchronically (in apparent time) and diachronically (in real time) in order to gain a deeper understanding of the processes responsible for intergenerational changes.

Perhaps the best demonstration of the importance of a consistently annotated corpus comes from work on optional ergativity, that is, the use or non-use of the ergative marker where the grammatical relation of the subject remains unaffected by non-use. Optional ergativity is driven by a number of factors. The original work by Meakins (2009; 2011a) identified animacy as a clear driver of ergative use, that is, inanimates were significantly more likely to be marked ergative than animates. However, in the study of intergenerational change in the use of the same case-marker, Meakins and Wilmoth (2020) found this variable not to be significant. This difference is the result of improved data annotation and extraction procedures. In the first two studies, transitive subjects including unmarked subjects were (excruciatingly!) extracted manually from an unannotated corpus, whereas Meakins and Wilmoth (2020) used a Python script across the corpus now coded for features such as 'subject', 'animacy' etc. The processing of coding subjects was discussed in §3.4.3. It is highly likely that many unmarked transitive subjects were simply missed in the first studies due to the manual nature of data extraction. This would have artificially inflated the frequency of subject marker use and skewed the statistical analysis. The difference in studies is a case in point about the importance of annotating corpora and annotating them well. Zero occurrences of a feature are impossible to find using automatic searches with unannotated data and very hard to do accurately just by eye-balling transcripts.

5. Conclusion

This paper has described the process of creating, annotating, and particularly of cleaning, the Gurindji Kriol corpus. At 853,348 words, and with no existing NLP resources for the language, the process of preparing this data for quantitative/computational research would not have been possible without a significant amount of consideration, planning, and careful automation. The result of this work is the largest annotated corpus of any Australian Indigenous language, which is amenable to innovative quantitative research. Our aim in this paper has been to describe our overall approach, which, though specific to our own needs, may provide a model for others developing large corpora of under-documented languages or working with legacy data that needs significant further processing. This approach has been drawn directly from Wilmoth's experience of managing language data annotation projects while working in industry. The following three principles encapsulate our approach and, in our experience, are often not applied in language documentation, where researchers usually do not approach data cleaning or annotation systematically.

[15] Wright-Fisher models are generally used in population genetics for describing a serial sampling process of genetic variants from one timestep to the next.

[16] Varbrul is a multivariate statistical model which uses observed token counts of variables and their variants to determine what factors influence the probabilities of choice of either variant. This is done by means of stepwise logistic regression, using a maximum likelihood algorithm. Varbrul can only calculate the use of individual variables whereas BayesVarbrul can account for multiple variables in a single model.

1. Take a bird's eye view of the data.

As a general principle, one should work on individual files as little as possible, and only when absolutely necessary. For example, instead of inspecting individual files, we used basic Unix tools and simple scripts to find and correct formatting errors. Similarly, to correct spelling and morphological tagging, we ran a script that takes every morph in the corpus, compares it to the lexicon file, and outputs anything that doesn't match. A corresponding script then corrects errors, only a tiny fraction of which need to be checked in context. Various scripts were used throughout this process that were not mentioned here, such as one which lists all Unicode codepoints found in the corpus and their frequencies, or one which simply lists all transcribed morphs in alphabetical order. This sort of basic, abstracted information about the corpus allowed us to identify many random errors incurred over the years.

2. Document everything.

Taking this bird's eye approach to the data makes documentation an easy part of the process. We not only document scripts and the order in which they're applied, we also document errors and how they ought to be corrected. This is part of the spelling correction process as described above, for example. Were spelling to be corrected in individual files, this effort would not be generalisable to similar errors across the corpus, and, importantly, would not be useful when any data is added in the future. This detailed documentation was especially helpful when, several years after the initial phase of the project, a significant amount of new data was added. While Wilmoth was not available to process all this data herself, she was able to supervise colleagues who could follow earlier documentation, and use the same scripts, regular expressions, and lists of errors and their corrections, speeding up the process significantly.

3. Undertake iterative validation.

This type of data-cleaning process is designed to be iterative. The spelling correction stage is one such loop, where the pair of scripts is run over all 853,348 words several times until the output is empty. Similarly, each type of error in morphological tagging is another iterative loop. Following each major stage of the data-cleaning process, we run through earlier steps yet again to ensure errors have not been introduced. With our refined and well-documented processes, this is quick and easy and allows us to catch errors before they cascade into other errors.

While data-cleaning always involves some amount of tedious work, it can be significantly sped up through thoughtful problem-solving and automation, taking into account these three principles. In recent years, there has been a push in language documentation to make data and methods more transparent and therefore replicable (Thieberger & Berez-Kroeker 2011; Mosel 2012; Gawne et al. 2017; Berez-Kroeker et al. 2018; Mosel to appear). As a result, the field has seen substantial improvements in archiving, data citation, and transparent discussion of fieldwork methodologies. However, in our experience, data processing, cleaning, and annotation is rarely discussed. This is even the case in NLP and Artificial Intelligence (AI) research, despite the oft-quoted mantra of 'garbage in, garbage out' in computing. For example, Sambasivan et al. (2021) describe pervasive data cascades in AI, where data quality issues lead to negative effects downstream. This is caused by a widespread undervaluing of data work, as opposed to the more 'glamourous' modelling work. As field linguists collect and annotate ever more data, producing rich documentary records, we will need to find ways to efficiently process and manage this data, aided by judicious use of automation (without sacrificing the insight that one gains from spending a significant amount of time immersed in one's data). As Thieberger and Berez-Kroker (2011: 93) write,

> If we take the time to produce well-formed, reusable data, our linguistic analyses will be more convincing, and our production of language materials will become more efficient. The initial outlay of time and effort [...] may seem high, but the investment will pay off in the end.

Well-structured and documented data can be used for many different ends over the long term; we also wish to emphasise that the *quality* of our data has downstream effects, particularly for quantitative and corpus-based research. The process described in this paper, though requiring a significant investment of time and effort, has provided a unique opportunity to research language variation and change in a complex contact situation. While the scale of the Gurindji Kriol corpus may be an outlier within the field at the time of writing, we hope to have provided a useful model for others developing larger documentary corpora.

References

Babinski, Sarah, Jeremiah Jewell, Kassandra Haakman, Juhyae Kim, Amelia Lake, Irene Yi, & Claire Bowern. 2022. How usable are digital collections for endangered languages? A review. *Proceedings of the Linguistic Society of America* 7(1). 5219. doi:10.3765/plsa.v7i1.5219

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer et al. 2018. 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

Bromham, L., X. Hua, C. Algy, & F. Meakins. 2020. Language endangerment: a multidimensional analysis of risk factors. *Journal of Language Evolution* 5(1). 75–91. doi:10.1093/jole/lzaa002

Buseman, Alan. 2019. Toolbox (version 1.6.4) [Computer software] SIL. (https://software.sil.org/toolbox/) (Accessed 2024-12-16.)

Charola, Erika & Felicity Meakins (eds.). 2016. *Yijarni: True stories from Gurindji country*. Canberra: Aboriginal Studies Press.

Chua, Mason, Daan van Esch, Noah Coccaro, Eunjoon Cho, Sujeet Bhandari, & Libin Jia. 2018. Text normalization infrastructure that scales to hundreds of language varieties. *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*. (https://aclweb.org/anthology/L18-1216) (Accessed 2023-08-23.)

Drager, Katie K. 2018. *Experimental research methods in sociolinguistics* (Research Methods in Linguistics). London: Bloomsbury academic publishing Plc.

ELAN (Version 6.8) [Computer software]. 2024. Nijmegen: The Language Archive, Max Planck Institute for Psycholinguistics. (https://archive.mpi.nl/tla/elan) (Accessed 2024-12-16.)

Evans, Nicholas (ed.). 2011. *Reciprocals and semantic typology* (Typological Studies in Language vol. 98). Amsterdam: John Benjamins.

Gawne, Lauren, Barbara F. Kelly, Andrea L. Berez-Kroeker, & Tyler Heston. 2017. Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation & Conservation* 11. 157–189. (https://hdl.handle.net/10125/24731)

Hardy, Frank. 1968. *The unlucky Australians*. Melbourne: Nelson.

Hua, Xia. 2022. BayesVarbrul: A unified multidimensional analysis of language change in a speaker community. *Journal of Language Evolution* 7(1). 40–52. doi:10.1093/jole/lzac004

Hua, Xia, Felicity Meakins, Cassandra Algy, & Lindell Bromham. 2021. Language change in multidimensional space: New methods for modelling linguistic coherence. *Language Dynamics and Change* 12(1). 78–123. doi:10.1163/22105832-bja10015

Levinson, Stephen C. & David P. Wilkins. (eds.). 2006. *Grammars of space: Explorations in cognitive diversity*, 1st edn. Cambridge University Press. doi:10.1017/CBO9780511486753

Levshina, Natalia. 2022. Corpus-based typology: Applications, challenges and some solutions. *Linguistic Typology* 26(1). 129–160. doi:10.1515/lingty-2020-0118

MacWhinney, Brian. 2014. *The Childes project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press.

MacWhinney, Brian. 2019. *CHAT manual*. TalkBank. doi:10.21415/3MHN-0Z89

Maletic, Jonathan I. & Andrian Marcus. 2005. Data cleansing. In Maimon, Oded & Lior Rokach (eds.), *Data mining and knowledge discovery handbook*, 21–36. Boston, MA: Springer US. doi:10.1007/0-387-25465-X_2

Mayer, Mercer. 1994. *Frog, where are you?* China: Puffin.

McConvell, Patrick. 1988. Mix-im-up: Aboriginal codeswitching old and new. In Heller, Monica (ed.), *Codeswitching: Anthropological and sociolinguistic perspectives*, 97–124. Berlin: Mouton de Gruyter.

McConvell, Patrick & Felicity Meakins. 2005. Gurindji Kriol: A mixed language emerges from code-switching. *Australian Journal of Linguistics* 25(1). 9–30. doi:10.1080/07268600500110456

Meakins, Felicity. 2008. Land, language and identity: The socio-political origins of Gurindji Kriol. In Meyerhoff, Miriam & Naomi Nagy (eds.), *Social lives in language*, 69–94. Amsterdam: John Benjamins. doi:10.1075/impact.24.08mea

Meakins, Felicity. 2009. The case of the shifty ergative marker: A pragmatic shift in the ergative marker in one australian mixed language. In Barðdal, Jóhanna. & Shobhana L. Chelliah (eds.), *The role of semantics and pragmatics in the development of case*, 59–91. Amsterdam: John Benjamins. doi:10.1075/slcs.108.06mea

Meakins, Felicity. 2011a. *Case-marking in contact: The development and function of case morphology in Gurindji Kriol* (Creole Language Library vol. 39). Amsterdam: John Benjamins.

Meakins, Felicity. 2011b. Spaced out: Intergenerational changes in the expression of spatial relations by Gurindji people. *Australian Journal of Linguistics* 31(1). 43–77. doi:10.1080/07268602.2011.532857

Meakins, Felicity. 2015. From absolutely optional to only nominally ergative: The life cycle of the Gurindji ergative suffix. In Gardani, Francesco, Peter Arkadiev, & Nino Amiridze (eds.), *Borrowed morphology* (Language Contact and Bilingualism vol. 8), 189–218. Berlin: De Gruyter Mouton.

Meakins, Felicity. 2016. No fixed address: The grammaticalisation of the Gurindji locative as a progressive suffix. In Meakins, Felicity & Carmel O'Shannessy (eds.), *Loss and renewal: Australian Languages since colonisation*, vol. 13, 367–395. Walter de Gruyter.

Meakins, Felicity & Cassandra Algy. 2004. *Gurindji Kriol corpus*. doi:10.26278/CD9K-3Y70

Meakins, Felicity & Cassandra Algy. 2016. Deadly reckoning: Changes in Gurindji children's knowledge of cardinals. *Australian Journal of Linguistics* 36(4). 479–501. doi:10.1080/07268602.2016.1169973

Meakins, Felicity, Jennifer Green, & Myfany Turpin. 2018. *Understanding linguistic fieldwork* (Understanding Language Series vol. 1). London: Routledge.

Meakins, Felicity, Xia Hua, Cassandra Algy, & Lindell Bromham. 2019. Birth of a contact language did not favor simplification. *Language* 95(2). 294–332. doi:10.1353/lan.2019.0032

Meakins, Felicity & Carmel O'Shannessy. 2010. Ordering arguments about: Word order and discourse motivations in the development and use of the ergative marker in two Australian mixed languages. *Lingua* 120. 1693–1713. doi:10.1016/j.lingua.2009.05.013

Meakins, Felicity & Sasha Wilmoth. 2020. Overabundance resulting from language contact: Complex cell-mates in Gurindji Kriol. In Arkadiev, Peter & Francesco Gardani (eds.), *The complexities of morphology*, 81–104. Oxford: Oxford University Press.

Mosel, Ulrike. 2012. Advances in the accountability of grammatical analysis and description by using regular expressions. In Nordhoff, Sebastian (ed.), *Electronic grammaticography*, 235–250. Honolulu: University of Hawai'i Press. (https://hdl.handle.net/10125/4537)

Mosel, Ulrike. to appear. Corpus building for under-researched languages - a practical guide. In Ahoua, Firmin, Dafydd Gibbon, & Stavros Skopeteas (eds.), *Linguistic fieldwork and language documentation: A course book on foundational skills*.

O'Shannessy, Carmel. 2004. The Monster Stories: Picture stimuli for elicited production (Unpublished Series). Nijmegen: Max Planck Institute for Psycholinguistics. (https://carmeloshannessy.net/research.html) (Accessed 2024-05-16.)

Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, & Lora M. Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*, 1–15. New York: Association for Computing Machinery. doi:10.1145/3411764.3445518

Schnell, Stefan & Nils Norman Schiborr. 2022. Crosslinguistic corpus studies in linguistic typology. *Annual Review of Linguistics* 8(1). 171–191. doi:10.1146/annurev-linguistics-031120-104629

Sloan, Bodean & Meakins, Felicity & Algy, Cassandra. 2022. Intergenerational changes in Gurindji Kriol: Comparing apparent-time and real-time data. Asia-Pacific Language Variation 8(1). 1–31. (doi:10.1075/aplv.21001.slo)

Stanley, Joey. 2021. Order of operations in sociophonetic data processing.

Stewart, Jesse & Felicity Meakins. 2021. Advances in mixed language phonology: An overview of three case studies. In Mazzoli, Maria & Eeva Sippola (eds.), *New Perspectives on mixed languages: From core to fringe.* (Language Contact and Bilingualism). Berlin: De Gruyter Mouton.

Thieberger, Nicholas & Andrea L. Berez-Kroeker. 2011. Linguistic data management. In Thieberger, Nicholas (ed.), *The Oxford handbook of linguistic fieldwork*. Oxford University Press. doi:10.1093/oxfordhb/9780199571888.013.0005

Tomar, Suramya. 2006. Converting video formats with FFmpeg. *Linux Journal* (146). 10.

Ward, Charlie. 2016. *A handful of sand*. Melbourne: Monash University Press.

Zupon, Andrew, Evan Elizabeth Crew, & Sandy Ritchie. 2021. *Text normalization for low-resource languages of Africa*. AfricaNLP.

Sasha Wilmoth
orcid.org/0000-0002-6626-9104

Felicity Meakins
orcid.org/0000-0003-4487-4351