Workshop on Language Corpora in Australia

3 July 2023

Online live, via Zoom, 9am – 5.40pm

https://anu.zoom.us/j/82388177779?pwd=aHFEbFMwRVdqL3VzTmdDNm41QlA5UT09

Organisers: Catherine Travis & Li Nguyen

	Session 1	
9:00	Catherine Travis, Li Nguyen	Intro and Welcome
9:10	Catherine Travis	Sydney Speaks
9:35	Felicity Cox, Joshua Penney, Andy Gibson	Multicultural Australian English: The New Voice of Sydney
10:00	Steven Coats	The Corpus of Australian and New Zealand Spoken English
10:25	break	
	Session 2	
10:50	Louisa Willoughby, River Smith, Trevor Johnston	The Auslan Corpus and the Monash University Node of the Language Data Commons of Australia (LDaCA)
11:15	Gerry Docherty	An Overview of the "WestAuseE" Corpus
11:40	Elena Sheard	Sydney Speaks Lifespan Corpus
12:05	Erwanne Mas, Anne Przewozny	The PAC-Australia Corpus: A Small Spoken Corpus of Australian English for Sociophonetic and Dialectological Investigation
12:30	lunch	
	Session 3	
1:15	Celeste Rodriguez Louro, Glenys Collard	The Yarning Corpus: Aboriginal English in Southwest Western Australia
1:40	Alison Mount, Roy Barker, Jane Simpson	Muruwaringgu ngana yaan.gu - Creating a Corpus for Community from Recordings by Muruwari Man Jimmie Barker (1900-1972)
2:05	Sasha Wilmoth, Felicity Meakins	Small Language, Big Data: Building the Gurindji Kriol Corpus
2:30	break	
	Session 4	
3:00	Carmel O'Shannessy	Longitudinal corpus of language contact and change: Warlpiri and Light Warlpiri
3:25	Sally Dixon	The Ipmangker Corpus from Central Australia
3:50	Tünde Szalay, Kirrie Ballard, Felicity Cox, Beena Ahmed	AusKidTalk: Collecting a Corpus of 3- to 12-year-old Australian Child Speech
4:15	break	
	Session 5: Highlights	
4:30	Simon Gonzalez	A Text Database from Reddit: A Case for Australian English
4:35	Marissa Takahashi, Matthew Bettinson	Harnessing Online Public Discourse: Exploring Australian Twittersphere and NewsTalk Collections
4:40	Lucia Fraiese	Outta country: The Boarders' Corpus of Australian Aboriginal English
4:45	Cara Penry Williams	Interview 1: A Corpus of and about Young Adult Australian English Speakers
4:50	Sophie Richard	The University of Western Australia (UWA) Narrative Corpus
4:55	Question time for Highlights	
5:10	Catherine Travis, Li Nguyen	Discussion and next steps
5:40	Close	

9:00 – 9:10 Intro and Welcome
Catherine Travis and Li Nguyen
(Australian National University)

Over decades of work in Australia, significant collections of language data have been amassed, including of varieties of Australian English, Australian migrant language, Australian Indigenous languages, sign languages and others. These collections represent a trove of knowledge not only of language in Australia, but also of Australia's social and cultural history. And yet, not all are well known and many lack published descriptions. The purpose of this workshop is to provide an opportunity to share information about existing language corpora in Australia, with a view to producing a special issue of the Australian Journal of Linguistics that introduces a selection of these corpora, explores how they can contribute to our understanding of language, society and history in Australia, and considers avenues such corpora open up for future research.

This workshop is being run as part of the Language Data Commons of Australia (LDaCA), which is working to build national research infrastructure for the Humanities and Social Sciences, facilitating access to and use of digital language corpora for linguists, scholars across the Humanities and Social Sciences, and non-academics.

Session 1: 9.10-10.25

9:10 – 9:35 Sydney Speaks
Catherine Travis
(Australian National University)

Sydney Speaks is a sociolinguistic project exploring variation and change in Australian English. The Sydney Speaks corpus brings together spontaneous speech from 260 Sydneysiders, combining two legacy sub-corpora recorded in the 1970s and 1980s (Barbara Horvath's Sydney Social Dialect Survey and the NSW Bicentennial Oral History Project) with a contemporary corpus, recorded from 2015 to the present. Participant birth years span 100 years (from the 1890s to 1990s), and the sample is socially diverse, with participants coming from Anglo-Celtic, Italian, Greek and Chinese backgrounds, and representing a range of occupations, such as plumbers, hairdressers, university students, teachers, and doctors. The socio-historical information in the recordings provides rich contextualisation for interpretation of the linguistic patterns observed.

To date, 130 hours (over 1.25 million words) have been fully transcribed and time aligned at both the utterance and segment level, and a wide range of phonetic, morphosyntactic and discourse-pragmatic features have been examined. From these analyses, we have been able to observe ongoing changes for example in diphthongs such as FLEECE and FACE; a lengthening of word-final *-er* from the 1970s; the introduction of quotative *go* in the 1970s and its replacement by *be like* in the 2010s; and the introduction of the modal of obligation *need to* in the 2010s. In considering a role for ethnicity, we find that ethnic minorities may lead, lag or proceed alongside the majority community in these changes, and what appears to best account for that is their positioning within the socioeconomic hierarchy, evidence of a fundamental intersection between ethnicity and social class.

Sydney Speaks is an ongoing project that was funded by the ARC Centre of Excellence for the Dynamics of Language (2014-2022). It is housed in the ANU library (https://dx.doi.org/10.25911/m03c-yz22), and has been onboarded to LDaCA, where it can be accessed through an approval process.

9:35 – 10:00 Multicultural Australian English: The New Voice of Sydney

Felicity Cox, Joshua Penney, Andy Gibson
(Macquarie University)

Australia is one of the most ethnically diverse countries in the world with Sydney its most multicultural city (ABS 2017), yet our understanding of Australian English (AusE) speech patterns is largely Anglo-centric and fails to represent the community accurately. In 2016 (ABS 2016), 43% of people in Greater Sydney were found to be overseas-born, 67% had overseas-born parents, and 38% of households used a non-English language at home. These statistics, however, are opaque to the scope of diversity which varies from extremely diverse areas like Auburn where 84% of households use a non-English language, compared with only 12% in Pittwater. The *Multicultural Australian English* project aims to explore the role of ethnolinguistic diversity in the speech of AusE speaking young people.

We have recorded 184[1] AusE speaking adolescents from five areas of Sydney selected according to levels of diversity and dominant heritage languages spoken within the community: Northern Beaches (English), Cabramatta (Vietnamese), Bankstown (Arabic), Paramatta (Indic languages), Outer Inner West (Mandarin/Cantonese). Speakers were recorded in their schools by a local research assistant (RA). They completed a picture-naming task which included 185 single words (sampling targeted phonetic features) and 41 short phrases (capturing specific junctural contexts), and they engaged in an RA facilitated conversation with a peer discussing topics including local/Australian culture, school social structure, and their reflections on language. Speakers completed an adapted ethnic orientation survey (Hoffman & Walker 2010, Clothier 2014/in progress) and parents completed a comprehensive language background survey.

[1] Data collection is ongoing.

All data have been annotated orthographically and phonemic boundaries automatically time aligned. Word/phrasal tasks have been hand-corrected. The corpus will provide source material to help fill the considerable gap in our knowledge of variation in present-day AusE to enable a more representative picture of the variety. It will inform our understanding of relationships between language, culture, ethnicity and identity.

10:00 – 10:25 The Corpus of Australian and New Zealand Spoken English

Steven Coats
(University of Oulu, Finland)

The Corpus of Australian and New Zealand Spoken English (CoANZSE; Coats 2022) is a 190-million-word corpus of Automatic Speech Recognition (ASR) transcripts from YouTube channels of local councils and other governmental bodies in 472 locations in Australia and New Zealand, suitable for geospatial analyses of lexical, morpho-syntactic, pragmatic, and phonetic variation. In addition to linguistic variation, the corpus, which comprises transcripts of local government content, contains discourse on a wide range of political, social, and economic topics, making the it potentially interesting to researchers working in social science and humanities fields more generally.

While existing resources such as the AusTalk corpus (Estival et al. 2014; Cassidy et al. 2017) contain high-quality audio, making them suitable for phonetic analysis, CoANZSE may offer some advantages: First, it is significantly larger in size than existing corpora of Australian or New Zealand English, making it more likely to contain specific lexical items, grammatical constructions, and utterance sequences. Second, the transcripts are complete and searchable, and third, CoANZSE includes precise geolocation metadata. Fourth, audio and video content can be accessed via open-source scripts. As of May 2023, audio extraction and forced alignment is being undertaken for the entire corpus. The presentation provides a brief overview of the procedures used for creating the corpus and notes two preliminary analyses: First, an investigation of double modals, a rare syntactic feature; and second, a consideration of vowel quality in Australia based on millions of naturalistic vowel tokens. Cox and Palethorpe (2019) examined 5,722 vowel tokens produced by speakers in Sydney, Melbourne, Adelaide, and Perth in order to characterize regional differences in vowel quality. CoANZSE offers the possibility for analysis at a finer level of granularity: the force-aligned text grids for the 489 videos from a single channel, that of the City of Adelaide, for example, contain 3,136,405 vowel tokens. The scale of CoANZSE can thus allow detailed phonetic analyses to be conducted targeting particular lexical items or utterance types. In addition, manual (or automated) annotation of demographic categories can make analysis of variation according to traits such as apparent age, gender, or ethnic group possible.

Session 2: 10.50-12.30

10:50 – 11:15 The Auslan Corpus and the Monash University Node of the Language Data Commons of Australia (LDaCA)

Louisa Willoughby, River Smith, Trevor Johnston
(Monash University)

This project is building on and extending two Auslan datasets—a corpus and a dictionary. The new corpus component will include a corpus of deafblind tactile Auslan and potentially multimodal spoken language corpora. The new dictionary component will be able to accommodate individual signs from other sign languages (e.g., Australian indigenous sign languages) and potentially co-speech gesture data. In this presentation we focus on the Auslan corpus. The Auslan corpus contains video recordings of a representative sample of native or near-native deaf Auslan users, balanced for gender and age, drawn from five Australian cities. Recordings of language use were made of 20 signers (in pairs) in each city during 50 three-hour sessions, involving both natural conversation and elicitation materials. There are approximately 1000 clips edited by task representing 150 hours of recordings. The recordings were collected with the intention that it would form the basis of an annotated machine-readable linguistic corpus. Annotation of the corpus began in 2005 but, to date, less than half of the task clips have either individual sign glosses or translations.

Approximately 200 have additional detailed multi-tiered linguistic annotations. Studies of the lexicon and grammar of Auslan using the corpus include studies on the use of space, sign grammatical classes, sociolinguistic variation, phonological variation, grammaticalization, lexical frequency, and the relationship between depiction, indexing, gestures and enactment (see selected studies below). The Auslan corpus (recordings and annotation files) and dictionary (Auslan Signbank) are now archived at Monash. The LDaCA initiative is building infrastructure linking these two resources. This will make access to data in the corpus much easier for researchers as well as make further annotations of the corpus more accurate and less time consuming which, in turn, will further enhance research opportunities.

11:15 – 11:40 An Overview of the "WestAuseE" Corpus

Gerry Docherty
(Griffith University)

The "WestAusE" corpus arose from a large-scale study of the phonetic characteristics of speakers from Perth, WA (funded by ARC DP130104275). It comprises digital recordings of 150 young speakers (aged 18-22, 72m/78f) engaged in 30 min same-sex dyad conversations, plus the same speakers also reading a list of isolated words and a standard set of sentences. The speakers were entirely-schooled in Perth and were classified re: their residence in suburbs in the top or lower decile of the ABS SEIFA neighbourhood rankings,

or in the middle range of the same scale. A small sub-set were also recorded a second time in unscripted conversation with an older family member in order to allow comparisons across peer-to-peer and peer-to-parent speech styles. The recordings were made in the period 2013-2016, and the corpus was constructed within a LaBB-Cat environment (Fromont & Hay 2012). A subset of 40 speakers from the main corpus, balanced for speaker sex and neighbourhood, was identified for manual correction of automatic segmentation prior to acoustic analysis of a range of consonantal and vocalic features. Work on this corpus is continuing to develop some time after the end of the initial funding for the project, and further recordings are progressively being added to this manually-corrected sub-section of
the main corpus. This contribution will present details of the corpus, highlighting strengths and weaknesses, and focusing on areas that either have already been the focus of sociophonetic investigation or which are in the pipeline for investigation.

11:40 – 12:05 Sydney Speaks Lifespan Corpus

Elena Sheard
(University of Canterbury, New Zealand)

The combination of trend and panel data is 'crucial' in building more informed models of the role of individual speakers in language change (Sankoff and Blondeau 2007: 561). The Sydney Speaks Lifespan Corpus is not only a rare example of a panel study that is both ethnically diverse and representative of members of ethnic minority communities (cf. Sankoff 2018), it is directly complemented by Sydney Speaks trend data from the same ethnic communities and the Anglo majority (Travis, Grama, and Gonzalez In Progress). It is therefore of substantial value for understanding the role of ethnic minority communities and individuals in language change, particularly in multi-ethnic urban contexts. The Sydney Speaks Lifespan Corpus participants are five Greek-background and five Italian-background Sydneysiders who were born in Australia to parents who migrated from Greece or Italy. The ten participants were first recorded as teenagers in the late 1970s as part of the Sydney Social Dialect Survey (Horvath 1985), and again as middle-aged adults in 2019 as part of the Sydney Speaks project, following the project's data collection and storage protocols. The interviews are in Australian English, and audio recordings and corresponding transcripts are stored in a LaBB-CAT server (Fromont and Hay 2012). A key finding obtained through the integrated analysis of Sydney Speaks Lifespan Corpus panel data with Sydney Speaks trend data is that for changes in the realisations of FACE, FLEECE, GOAT, MOUTH, and PRICE, there is not a single 'direction of change' that panel participants are following so much as shared generational targets to which they are headed from their respective starting points as teenagers (which is affected by ethnic background). The distinction between retrograde change and lifespan change for these diphthongs is therefore more blurred than has been documented in the literature to date, in which these trajectories have functioned as distinct phenomena (cf. Sankoff 2018).

12:05 – 12:30 The PAC-Australia Corpus: A Small Spoken Corpus of Australian English for Sociophonetic and Dialectological Investigation

Erwanne Mas, Anne Przewozny
(Université de Toulouse, France)

The PAC corpus in Australia has been under development since 2003, building on the paradigms of corpus phonology (Durand, Gut and Kristoffersen, 2014; Durand and Przewozny, 2012; Przewozny, Viollain and Navarro, 2020) and variationist sociolinguistics (Labov 1994, 2001; Tagliamonte, 2006). The present paper aims at describing the three subcorpora of 2003, 2015 and 2018 as well as ongoing work. The sound files of our 37 informants from the Australian corpus of the PAC programme (The Phonology of Contemporary English: usages, varieties, structure) were recorded in 2003, 2015 and 2018 in a number of urban and regional settings across Australia: the capital cities of Sydney and Melbourne, Deniliquin (Riverina region) and the coastal town of Ulladulla, as well as White Cliffs (Central Darling Shire), NSW. Another field survey, to take place in June and July 2023, will be devoted to the historical city of Ipswich in Greater Brisbane, the rural town of Charleville (Shire of Murweh) and the suburb of Caloundra West (Sunshine Coast) in regional Queensland. The PAC-Australia corpus aims to document Australian English across sets of Australian speakers within urban and regional parts of Australia, as well as to build on previous works in the sociophonetics of Australia (e.g., Harrington & al., 1997; Cox, 1998, 2008; Billington, 2011; Butcher, 2012; Burridge, 2020). The PAC-Australia informants are all native speakers of Australian English (two of them being Aboriginal English speakers) and were selected on the basis of their territorial anchoring to their places of residence. The following protocol was applied: after signing a consent form, our informants were asked to proceed with two reading tasks (two wordlists and a 500-word text). They then had to participate in a conversational-style interaction with the interviewers, on the basis of specific sociolinguist guidelines (with items such as place of birth, place of residence, languages spoken at home, schooling, parents' and grandparents' background etc.), followed by the recording of informal interaction between the informant and a person from within their own network. Our corpus comprises 37 respondents as of May 2023. This makes up more than 50 hours of spoken data and anonymised sociolinguistic information which are stored on the MyCore CNRS server (the French Centre National de la Recherche Scientifique) and filed in the PAC programme archives.

We examine the relevance of 'small' spoken corpora such as PAC-Australia for the study of Australian English varieties from a sociophonetic point of view. Acoustic analyses on real-time data were carried out using Praat and then normalised with Lobanov's speaker normalisation method. The first findings show that a vocalic chain shift may be under way in our female informants from NSW compared to female informants from Victoria. Secondly, we expose some recent results on Standard Aboriginal English drawn from research on lexical stress patterns that confronts some PAC-Australia data with some AusTalk data and with dictionary-based data (Przewozny & Martin 2023). Finally, we discuss the ongoing PAC survey across Queensland which is meant to contribute to comparative research on vocalic change within Australia and in other varieties of English and to current research in dialectology across Australia (Docherty, Gonzalez & Mitchell, 2015).

Session 3: 1.15-2.30

1:15 – 1:40 The Yarning Corpus: Aboriginal English in Southwest Western Australia

Celeste Rodriguez Louro, Glenys Collard
(University of Western Australia)

Australian Aboriginal English (AE) is a post-contact variety used by approximately 80% of First Nations people in Australia (Rodríguez Louro & Collard, 2021a: 5). At least 400 traditional First Nations languages were spoken in Australia before invasion (Bowern, 2023: 56). Today, only 20-25 First Nations languages will be passed on to the next generation (Karidakis & Kelly, 2018: 106). AE thus stands as a powerful encoder of ethno-cultural identity.

Funded through an ARC DECRA (Rodríguez Louro, 2018-2022), this project seeks to uncover variation and change in AE as used in Nyungar country, Southwest WA. We designed a cross-cultural research model that allowed to document sociolinguistic aspects of AE *yarning* – a First Nations cultural form of storytelling and conversation – which we used to capture the voices of those rarely featured in sociolinguistic research.

Collected in 2019/2020, the *Yarning Corpus* consists of acrolectal AE data gathered in metropolitan Perth, WA. It features video-recorded interactions with 58 AE speakers born between 1931 and 2009. These materials, housed at UWA, amount to 517,600 words of unscripted, interactional speech data from 36 women and 22 men aged 10-88 who speak AE as their vernacular, the register that emerges 'automatically and unthinkingly' (Labov, 2013: 3). We focus on speakers who use AE as their main and only language of communication, and who do not speak traditional languages fluently.

The *Yarning Corpus* has allowed us to enrich sociolinguistic research methods (Rodríguez Louro & Collard, 2021b) by utilising an approach grounded in a relational ethic (Collard & Rodríguez Louro, Submitted). It has also allowed us to explore how the quotative system of AE is changing (Rodríguez Louro, Collard, Clews & Gardner, Forthcoming), and to examine how cultural frameworks such as a group orientation have shaped variation and change in AE (Rodríguez Louro & Collard, Under contract).

1:40 – 2:05 Muruwaringgu ngana yaan.gu - Creating a Corpus for Community from Recordings by Muruwari Man Jimmie Barker (1900-1972)

Alison Mount, Roy Barker, Jane Simpson
(Australian National University)

Jimmie Barker (1900–1972) was probably the first Indigenous Australian to use audio recording to document Aboriginal languages and culture, recording over 113 hours which include documentation and analysis of his own language Muruwari and surrounding Aboriginal languages, and reflections on his life, Australian and international history. These recordings are archived at the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS). Jimmie's grandson, Roy Barker, has auditioned much of the

collection to determine access and use restrictions. He has led a team of researchers to transcribe and annotate 31 recording hours using ELAN. Our corpus comprises 29 recording hours of unrestricted material, totaling 15.5 hours of speech and 13.5 hours of silence recorded between 1969 and 1972 in Brewarrina, Lightning Ridge and Goodooga. Most recordings constitute self-elicitation and monologues by Jimmie, although a few contain elicitation of other Muruwari speakers, and speakers of other Aboriginal languages. In some recordings, Jimmie also simulates a dialogue by responding to audio and written stimuli from a collaborator at AIAS, AIATSIS's predecessor.

A key purpose of the corpus is to improve the findability and interoperability of the material. We have expanded the metadata for each recording, and annotated transcripts with 'tags' that reflect the overarching themes of Jimmie's collection. The team is partnering with AIATSIS to make the recordings accessible through an online interactive cultural portal for the public. The design of the portal relies heavily on the ELAN transcription, metadata annotation, and the thematic 'tags'. The team is also using the corpus to support the development of digital language and cultural learning resources, for the Barker family, and Muruwari, Indigenous Australian, and non-Indigenous communities, in line with Jimmie's explicit wishes. The collection is already supporting the development of a community spelling system, a dictionary, cultural mapping, and audio-visual pronunciation guides.

2:05 – 2:30 Small Language, Big Data: Building the Gurindji Kriol Corpus

Sasha Wilmoth, Felicity Meakins
(University of Melbourne, University of Queensland)

At over 170 hours, or 850,000 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. 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 size and consistency of the corpus has enabled innovative research into questions of language variation, contact, emergence, and change, including Hua et al. (2021) , Meakins et al. (2019) , Sloan et al. (2022) and Meakins and Wilmoth (2020) . The corpus was collected by Felicity Meakins and Cassandra Algy in Kalkaringi and Daguragu (NT) with 157 child and adult speakers of Gurindji Kriol between 2004-2017. The data includes conversation, free narrative, picture-prompt narrative and other picture-prompt tasks. Each utterance is translated into English and annotated morpheme by morpheme for lexeme, language source, part of speech and features of POS e.g. transitivity for verbs, animacy and grammatical relations for nouns. Recordings were transcribed in the CHAT format using the CLAN software (MacWhinney, 2014) . Each file contains several lines of metadata, time-aligned transcriptions, as well as tiers for morphological information, translation, and notes. An excerpt from one file is shown in (1), with the metadata for the session and the first utterance.

(1) Excerpt from FM13_35_3a:

```
@UTF8
@Begin
@Languages: gu, en
@Participants: FBP NAME.
@Location: Lawi. Monster story. NAME is 12 years old. NAME
NAME, NAME and NAME are recording. 1:49min.
@Media: FM13_35_3a, audio
@Date: 29-OCT-2013.
*FBP: nyawa _ma karu _ma im pleigraun _ta pleibat warlaku _yawung .
5526_10797
%mor: dem|nyawa&amp;g=this suf:top|_ma&amp;g=TOP n:human|karu&amp;g=child
suf:top|_ma&amp;g=TOP pro|i&amp;k=3SG n:inanimate|pleigraun&amp;k=playground
case:loc|_ta&amp;g=LOC v:intran|plei&amp;k=play
n:animal|warlaku&amp;g=dog der:having|_jawung&amp;g=PROP .
%eng: This kid is in the playground playing with a dog .
```

Several precise conventions were followed in order for the format to be compatible with CLAN's automatic searching and morphological tagging. Cleaning of the data represented a major effort, and is the focus of this paper. This was accomplished using 20 custom-written Python scripts, complex regular expressions, as well as other tools such as OpenRefine. The result is detailed, accurate, and completely consistent transcription and annotation of the entire corpus. The corpus is available in the original CHAT format (used by CLAN), as well as ELAN and Toolbox formats, time-aligned to video and audio recordings. In this paper, we describe the steps taken to accelerate corpus development, including automatically generating the base CHAT files before transcription, standardising spelling, and reducing manual effort in tagging all subject NPs. All scripts are available on GitLab, 1 however more generally, it is our overall approach to data cleaning, drawn from industry best practice, which may provide a useful model for others working on the development of corpora of low-resource languages.

Session 4: 3.00-4.15

3:00 – 3:25 Longitudinal Corpus of Language Contact and Change: Warlpiri and Light Warlpiri

Carmel O'Shannessy
(Australian National University)

Australian languages show both maintenance and continuity, and contact-induced change. To address questions about processes involved in continuity and change, documentation of ways of speaking over time by similar cohorts is required. The longitudinal corpus of Warlpiri and Light Warlpiri was initiated in 2002 to address questions arising from these perspectives.

Warlpiri is a Pama-Nyungan language spoken in four Warlpiri communities in Australia's Northern Territory, being learned by all generations. Light Warlpiri is a recently-emerged mixed language spoken by younger adults and children in one Warlpiri community, Lajamanu. These speakers also speak Warlpiri.

The corpus is at The Language Archive, hosted by the Max Planck Institute for Psycholinguistics, The Netherlands. The speakers are adults and children speaking Warlpiri, from four Warlpiri communities, and Light Warlpiri, from one community; 121 adults and 249 children have been recorded. Recordings are on video and audio and almost all are transcribed in CHAT and/or ELAN formats. There are four main types of data:

1) longitudinal naturalistic interactions between adults and children, and children with each other, with toys and picture books as prompts (Egan 1986; O'Shannessy 2004; San Roque et al 2012).
2) elicited narratives - individual adults and children tell stories from the picture book prompts;
3) elicitation of a reflexives and reciprocals video description task (Evans et al, 2011); tasks for phonetic analysis (picture naming and sentence reading); animated response task; animated task about transitivity, and
4) discussions of grammar points that are not frequent in the other data.

The corpus provides the base of documentation work on the mixed language, Light Warlpiri, addressing typological and sociolinguistic questions about the emergence of the new language. It also provided data for two ANU Linguistics Honours studies. The Warlpiri component has been used in professional development workshops by Warlpiri educators in the four communities. The corpus is used to document language variation and change in each of Warlpiri and Light Warlpiri, and provide data to address a range of questions, including child development and multilingualism. Part of the corpus is reproduced in the international DoReCo and MultiCast corpora, and in the CoEDL language and text corpus.

3:25 – 3:50 The Ipmangker Corpus from Central Australia

Sally Dixon
(University of New England)

Many corpora are motivated by the analytical goal of describing some aspect of a language or languages. In this presentation, I will ask the complementary questions of how and why we go about documenting a language repertoire. To explore these questions, I will describe the Ipmangker Corpus, a collection of 50+ hours of video-recorded child language use collected in the small, remote Indigenous community of Ipmangker, on Alyawarr and Kaytetye country, Central Australia. Children in this community grow up in a very rich language ecology, with traditional Australian languages, contact languages, Standardised Australian English, global English varieties as well as regional Australian English all to be heard in the community. The main every day language of young people is Alyawarr English, a new English-lexified contact language.

The corpus focuses on naturalistic child-child and child-adult interactions from six target children, filmed at four, 6-monthly intervals between the ages of 5 and 8. Naturalistic recordings are set in range of home and school contexts. In addition, there are recordings of children narrating wordless picture books designed to elicit some grammatical structures that predominate in discussions of the emergence of Australian mixed-languages (O'Shannessy 2004). Approximately 50% of video recorded interactions have

time-aligned transcriptions in ELAN format. The corpus is currently stored at ANU, but will soon be moved to AIATSIS, with access conditions controlled by Ipmangker community members.

The corpus has been designed to capture the breadth of children's language repertoires, and as such includes use of structures readable as Standardised Australian English and structures readable as conforming to Alyawarr English. Comparing the use of variable structures across these contexts has allowed for valuable insights into the nature of contact language repertoires and their management (e.g. Dixon 2021; 2018).

3:50 – 4:15 AusKidTalk: Collecting a Corpus of 3- to 12-year-old Australian Child Speech

Tünde Szalay, Kirrie Ballard, Felicity Cox, Beena Ahmed
(University of Sydney, Macquarie University, University of New South Wales)

Child speech is difficult to collect and annotate, greatly limiting the development of speech recognition tools for children [1]. Less than 20 child speech corpora exist worldwide, with only three usable to develop speech processing systems, none from Australian children. AusKidTalk [2] remedies this by providing a corpus of Australian child speech accessible via a public research data repository. AusKidTalk consists of audio-visual recordings of Australian English typically developing and disordered speech from 600 children aged 3 to 12 years. Recordings are made of children participating in a variety of activities presented via a tablet application. The protocol includes 1) 117 single words of varying syllable length and lexical stress patterns that collectively sample all Australian English consonants and vowels at least once, 2) 36 sentences of varying word-lengths, intonational patterns and grammatical structures, 3) a narrative story telling task that uses a story picture sequence, 4) speech describing emotions elicited by two short video clips and 5) 40 nonsense words that provide developmental data on unseen word repetition. The protocol results in approximately 30 mins of continuous clean usable speech from each 3-5 year olds; 45 mins from 6-8 year olds; and 60 mins from 9-12 year olds. The data is collected across four locations in Sydney and stored in a central server at UNSW Sydney. Success in data collection was linked to the developed interactive, child-friendly recording setup and engaging activities used to elicit speech. A multi-step annotation workflow reduced manual word level orthographic annotation time by augmenting a developed, custom task-specific automatic speech transcription system that produced textgrids with manual correction. AusKidTalk will provide a large speech corpus to study the development of speech and language in Australian children plus enable the development of child-specific speech recognition-based applications, benefiting Australia's regional and remote communities.

Session 5: 4.30-5.05- HIGHLIGHTS

4:30 – 4:35 A Text Database from Reddit: A Case for Australian English

Simon Gonzalez
(Australian National University)

The advent of social media platforms has transformed the way people communicate and express themselves. These platforms generate an enormous amount of textual data, which can provide valuable insights into various aspects of human behaviour, language use, and societal trends. One compelling reason for creating such databases is the ability to study the change of language in real-time and capture the dynamic nature of communication within online communities. Despite the growing interest in text databases derived from social media platforms, there is a notable absence of such resources specifically focused on Australian English. This gap poses challenges for researchers seeking to analyse linguistic phenomena, sociocultural patterns, and regional variations within the Australian context. To address this limitation, we present a novel dataset from Reddit, in Australia, which fills the existing void and offers valuable insights into language use in this specific linguistic domain.

The dataset consists of over 200K posts by over 10K users spread across all regions in Australia. These posts represent a variety of topics, covering areas such as politics, entertainment, sports, and daily life. The dataset comprises a total of ~10M words, providing a concise yet comprehensive representation of the Australian English lexicon within the context of social media discourse. Among these words, there are 300K unique words, capturing the core vocabulary employed by Australian English speakers in online communication.

Reddit's anonymity, pseudonymity, and no limit of number of characters per post often encourage users to express themselves more freely, leading to the emergence of novel linguistic phenomena such as internet slang, memes, and other innovative language practices. These linguistic innovations can be studied to better understand the rapid evolution of language in online spaces and its relationship with identity, social networks, and cultural trends.

4:35 – 4:40 Harnessing Online Public Discourse: Exploring Australian Twittersphere and NewsTalk Collections

Marissa Takahashi, Matthew Bettinson
(Queensland University of Technology)

We introduce two expansive datasets capturing the dynamics of online public discourse in Australia: the Australian Twittersphere (AuTS) and NewsTalk. AuTS is a well-established, longitudinal dataset that collates a broad array of "Australian" tweets, while NewsTalk is a newer, multi-platform compilation of reader comments on Australian news stories sourced from news platforms including ABC,

Guardian, News Corp, Nine Entertainment, MSN, along with regional newspapers and independent news outlets.

The AuTS collection has supported a range of interdisciplinary research, providing valuable insights into the shifting landscape of public discourse. However, with the maturity of platforms like Twitter, challenges emerge, and discourse migrates to new platforms. NewsTalk, initiated in December 2022, addresses this evolving environment. It currently amasses over half a million comments, with an increasing rate of 140,000 additions monthly, reflecting contemporary platforms for discourse such as news websites and Reddit. Both collections offer a near real-time, large-scale, comprehensive capture of public discourse on topical issues in Australia, including politically charged subjects like the Indigenous voice to Parliament and the AUKUS Defense Agreement. Beyond mere data acquisition, we provide tools for efficient search, tidy data management, processing, and analysis across these collections. The Australian Twittersphere is a longitudinal collection of tweets from a periodically updated list of Twitter accounts that are identified as Australian (i.e., have a stated connection to Australia in the free text fields of the account profile). The current collection has collected an average of 42 million tweets per month. The average monthly active user in the current collection is 466,288. NewsTalk is a new data-as-a-service offering that comprises eighteen bespoke site comment harvesters, a Reddit harvester. Researchers can access NewsTalk via a website to explore and download data, or they can use a provided API and Python package for integrating NewsTalk data into analytical workflows.

4:40 – 4:45 Outta country: The Boarders' Corpus of Australian Aboriginal English
Lucia Fraiese (University of Western Australia)

The Boarders' Corpus of Australian Aboriginal English (BAE) is a growing dataset of spontaneous conversation among First Nations youth in Australia. The field site, renamed by the students as St Mary's Hills to safeguard the institution's anonymity, is a boarding school located in Whadjuk Nyungar country, Southwest Western Australia. It is the home away from home for First Nations teens who every term leave their homes to pursue a private education in the mainstream system. Boarders at St Mary's Hills hail from across WA and from the Northern Territory. The students speak traditional languages such as Walmajari and Miriwoong, and /or Kriol when conversing with relatives face-to-face and on the phone. They also speak Australian Aboriginal English, a post-invasion contact-based variety of English used by approximately 80% of First Nations people in Australia (Rodríguez Louro & Collard, 2021a: 2).

The BAE currently comprises 20+ hours of audio-recorded material, featuring 23 female speakers and four male speakers aged 12-17. Data collection, which began in September 2022, is scheduled to continue until December 2023 as part of a sociolinguistic ethnography akin to Eckert's (1989) canonical work with adolescents in a Detroit highschool. The recorded data consists of First Nations students engaging in conversation with peers, as well as with the author. In line with Rodríguez Louro and Collard (2021b: 7), no predetermined questions are employed. Instead, yarning, a First Nations cultural form of storytelling and conversation, is used to collect data in a culturally safe way. As the first sociolinguistic corpus of First

Nations boarders, this dataset will contribute to our understanding of the linguistic experiences of First Nations communities in mainstream institutions. It will also throw light into the role that language plays in the construction of boarders' identities. Additionally, the ethnographic nature of the project will help to uncover the social meaning of linguistic variation.

4:45 – 4:50 Interview 1: A Corpus of and about Young Adult Australian English Speakers
Cara Penry Williams (University of Derby, La Trobe University)

This presentation details a corpus created in Melbourne to study Australian English as used by its young adult speakers. I initially outline and explain the sampling and interview design. The invitation to participate in the research asked that a potential interviewee was (1) a speaker of Australian English, (2) 20– 30 years of age, (3) not a linguist or a linguistics student, (4) unknown to the researcher, and (5) grew up in Melbourne. Ethnicity was intentionally not controlled and participants were not required to be monolingual. The interview was designed specifically for the topic and borrowed from but did not replicate existing sociolinguistic methods. Short tasks were included present ideas about identity and encourage engaged discussion of the specifics of language variation whilst avoiding shaping this too much (Penry Williams 2019).

Within the resulting audio corpus there are 17 speakers, matching the criteria above, and one interviewer of similar background (me, in 2008). In my first analysis of these data, I ruled out two participants for comparability. The data were transcribed in Praat (Boersma & Weenink 2017) into Santa Barbara transcription DT2 (Du Bois et al. 1992, Du Bois 2006) with variants tagged. Using a Praat Script created for a US project (Kendall 2006/2009), these were exported and compiled in MS Excel, where around 20 language features were tagged in a column each, displayed in context and with timestamps. The corpus is stored privately.

Important findings included updating quite sparse literature on Australian English in terms of a number of points of contemporary use. Analysis also centred on the participants accounts of what was meaningful to discuss in relation to language variation. This data-driven approach shaped my research focus in a number of ways and challenged me to contend with complexities of metapragmatic discourse that were not described elsewhere.

4:50 – 4:55 The University of Western Australia (UWA) Narrative Corpus
Sophie Richard
(Université de Tours)

The UWA Narrative Corpus consists of 210 narratives produced by 58 Australians who have Mainstream Australian English (MAE) as their L1 and who, for the most part, were born and raised in Western Australia (55/58), specifically in the Perth metropolitan area (47/55), and still lived there when recorded. Data collection took place in Perth between 2013 and 2016.

The aim was to elicit Labovian (Labov & Waletzky 1967), performed (Wolfson 1978) and monologic narratives (Romano, Porto & Molina 2013: 73-74) from a socially diverse sample of MAE speakers to investigate tense/aspect variation in narratives and unveil the sociolinguistic factors conditioning variant choice. Using a set of specifically designed prompts, participants were engaged in storytelling with the interviewer. Those "storytelling sessions" – loosely adapted from the traditional sociolinguistic interview (Labov 1984: 32) – were audio recorded and a written questionnaire was used to gather socio-demographic information from participants.

The final corpus comprises 7.5 hours of recorded speech (87,103 words) – exclusively narrative discourse – produced by 58 speakers of MAE (29 males, 29 females), engaged in various occupations (e.g., plumber/gas fitter, primary teacher, psychologist), and aged 16 to 69 at the time of recording. Individual narratives have been extracted and transcribed from the original recordings following a set of conventions partly based on Levey (2006: 149). The corpus is stored privately and will be uploaded onto CoCoON (Digital Oral Corpus Collections), a platform used to catalogue, store and archive oral corpora.

The data enabled fine-grained analysis of narrative structure and tense/aspect switching in MAE. Since narratives of personal experience epitomise the vernacular and reflect community norms (Labov 1984: 32), the corpus represents an invaluable source of semi-naturalistic data to investigate lexical, phonological, morpho-syntactic or discourse variation, offering future time depth, and a regional focus on metropolitan Perth.

4:55 – 5:10 Question time for highlights

5:10– 5:40 Discussion and next steps