Learning Through Transcription

Mat Bettinson and Steven Bird
Northern Institute, Charles Darwin University, Darwin, Australia
matthew.bettinson@cdu.edu.au, steven.bird@cdu.edu.au

Abstract

Transcribing speech for primarily oral, local languages is often a joint effort involving speakers and outsiders. It is commonly motivated by externally-defined scientific goals, alongside local motivations such as language acquisition and access to heritage materials. We explore the task of 'learning through transcription' through the design of a system for collaborative speech annotation. We have developed a prototype to support local and remote learner-speaker interactions in remote Aboriginal communities in northern Australia. We show that situated systems design for inclusive non-expert practice is a promising new direction for working with speakers of local languages.

1 Introduction

Speech transcription is typically motivated by the desire for lasting accessible records of language. However transcription can also be a method for linguistic inquiry and language acquisition (Bowern, 2008; Meakins et al., 2018). In this case it is a form of note-taking as the transcriber strives to make sense of what they hear. This practice is well established in documentary and descriptive linguistics, and it leads to detailed transcriptions that include metalinguistic detail. Computational linguists rely upon these annotated datasets to train language models. Non-specialist outsiders may find their alphabetic decoding skills useful for learning the local languages that are spoken in the places where they live and work, as evidenced by the number of learning resources that depend on having a written representation for these primarily oral languages.

We propose computational support for an activity we call *learning through transcription*. The form we propose is that of a system that supports transcription as a series of learning interactions. The focus of the computation is shifted from automation to computer supported cooperative work. We describe a design and engineering effort to address this need in remote Aboriginal communities in northern Australia.

Here, local people speak one or more local languages, along with various degrees of proficiency in English. Some non-indigenous Australians seek competency in Aboriginal languages to carry out cultural projects with local people. Some locals need to develop literacy in one of the local languages to support knowledge work in art centres, ranger programs, health clinics, schools, tourism operations, and so on. Accordingly, speech transcription is a practice that supports language acquisition and literacy development.

We report on a system for iterative word-level transcription modelled on 'sparse transcription' (Bird, 2020b). Developed through a course of Research-through-Design (cf. RtD in Zimmerman et al., 2007), 'Sparzan' is a vehicle for investigating methods for amplifying human effort in transcribing speech in primarily oral, local languages. We cover data models for learning through transcription, technologies and user interfaces for interactive transcription, and systems engineering.

A topic foregrounded by the COVID-19 pandemic is the tyranny of distance when outsiders seek to conduct fieldwork with people in remote linguistic communities (Williams et al., 2021). We investigate remote collaboration through a novel video messaging appliance which serves as a vector for learner-speaker interactions embedded in transcription work.

This paper is organised as follows. In Section 2 we describe the role of speech transcription for oral languages, including learning through transcription, interactive transcription, and situated systems design. In Section 3 we describe the Sparzan system, including the transcription client, the Lingobox appliance, and an example application. In Section 4 we reflect on the approach and draw lessons for future work.

Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 83 - 92

2 Background

Many local languages, including endangered and Indigenous languages, are purely oral, and there is no naturally-occurring context for deploying text technologies (Bird, 2022). Oral languages are not just languages that lack a writing system; the existence of an oral culture unlimited by writing leads to an entirely different situation (Ong, 1982). Local matters of concern include caring for the country, transmitting ecological knowledge to the next generation, and managing intercultural workplaces. This may mean that there is a need to record and transcribe ancestral knowledge about key places and practices, to compile vocabularies of local flora, fauna, and material culture, and for two-way language learning between the local vernacular and a language of wider communication. Apparently simple tasks like accessing an archive of historical recordings become more complex when one considers that we lack a standardised orthography for a community-agreed reference dialect supported by robust speech recognition. Thus, there is both a need for transcription, and a need to innovate when it comes to making transcriptions, through the design of novel processes and interfaces.

2.1 Learning through transcription

When we speak of language learning, we take a different focus to the usual kind of language learning that depends on previously prepared materials and resources, and that uses methods that are well-described in the field of second language teaching and learning (e.g. Nunan, 1999; Cook, 2016). Australian Aboriginal languages are rarely taught in formal settings, and so non-indigenous people tend to acquire local languages in an independent, self-directed way. One case in point is linguists, whose field methods often incorporate learning.

In one conception of fieldwork on local languages, outsiders enter with their agenda to capture a language, bringing with them a strong focus on creating textual resources for use in linguistic analysis and for training computational models. We immediately run into the so-called 'transcription bottleneck', which is being tackled in various ways, mostly depending on universal phone recognition (Besacier et al., 2014; Hasegawa-Johnson et al., 2016; Adams, 2017; Zanon Boito et al., 2017; Marinelli et al., 2019). Phone recognition for Indigenous languages nevertheless depends on recruiting linguists and local people to create phone level transcriptions. This approach downplays the cultural significance of the content, focussing on idiosyncrasies of form to the point where variations in pronunciation, even speech disfluencies, should ideally be transcribed (Bird, 2020a).

Instead, we begin with the agency of local communities and inquire about what people are already doing. In many Indigenous communities, this includes collaboration with outsiders on culturally meaningful tasks connected to land management, ecological knowledge, and transmitting traditional practices to the next generation. We believe that language work can sit in this space, so long as it is possible to design natural workflows. It may be as simple as shifting the discussion from 'how do we transcribe this utterance using a phonetic alphabet?' to 'what is the cultural significance of this word?' (Bird, 2022).

When we do this, we arrive at a kind of speech transcription which is not based on the idea of exhaustive transcription, but which identifies the significant words and phrases that are useful for organising and accessing audio collections, i.e. sparse transcription (Bird, 2020b). Sparse transcription is particularly suited to situation where newcomers enter a community and begin to learn language in the course of working with local people. Instead of phone recognition, this approach relies on a different off-the-shelf language technology, namely keyword spotting (Garcia and Gish, 2006; Gales et al., 2014; Le Ferrand et al., 2021).

At any stage of this process of learning a language through transcription, we have a personal lexicon of known words and phrases. We can engage speakers in discussions of the meaning of these words, perhaps using the contact language. We can listen to recorded passages with speakers, pick out further key words, elicit their meaning, and add them to the lexicon. This is an approach to language work which is more grounded in local concerns, e.g., interest in transmitting the content, and interest in supporting the learning journey of a newcomer.

2.2 Interactive transcription

The amplification of human effort with machine assistance is a practical approach suited for local languages. Often the most effective form of machine assistance is facilitating collaboration. Systems or *groupware* for computer supported cooperative work (CSCW) are now common in the workplace

(Khoshafian and Buckiewicz, 1995). Language-based CSCW systems have been described and implemented for language documentation and linguistic fieldwork (Hanke, 2017; Cathcart et al., 2012).

Interactive transcription sees the transcriber draw upon machine resources in real-time. Sparzan transcriber (see Sec. 3.2) is a related development to an interactive transcription prototype with an FST language model-backed real-time phone alignment and word completion for a polysynthetic language (Lane et al., 2021). In the present work we deprioritise established language models and metalinguistic analysis, instead focusing on conventional word-level transcription with assistance from keyword spotting and human-to-human language interactions.

2.3 Designing for inclusion

Transcription tools occupy a vital place in the construction of annotated corpora. Transcribers wish for high-quality easy-to-use software, that interoperates with other tools, and that support collaborative workflows (Finlayson, 2016; Thieberger, 2016). In this context we note that production-grade software development is beyond our resources, but some features are relevant as we seek to design for a realistic and useful artefact (rather than a prototype), in accordance with the research-through-design methodology (Zimmerman et al., 2007).

In contrast to the majority of speech transcription tools, our design focus lies not with expert transcribers and the production of annotated corpora, but rather in selective transcription by people working on the ground. Servicing this audience requires supporting non-experts of various types, including Western newcomers to remote Aboriginal communities, along with local people. In recent years there has been growing attention towards updated methodologies and fresh takes on software design to meet these needs. SayMore offered a design aimed at community participants as transcribers, integrating support for audio based workflows over metalinguistic annotation (Hatton, 2013). Similarly, tools developed under the Aikuma umbrella introduce designs for mobile and web-based tools, also aimed at community participation (Bird et al., 2014; Bettinson and Bird, 2017).

It is established practice for field linguists to perform transcriptions with real-time assistance of speakers (Meakins et al., 2018; Sapién, 2018). Yet access to speakers is often limited. Bespoke technology design can help make the most of time spent in the community, chiefly as mediating tools to support face-to-face interactions (Bettinson and Bird, 2021a). There have been proposals for remote collaboration as a form of linguistic crowdsourcing (Hatton, 2013; Bettinson, 2015). The Aikuma-Link prototype explored a mobile-based design to distribute consulting tasks to speaker's phones (Bettinson, 2020, p.87).

The global pandemic has challenged us all to innovate in remote working practice, including interactions with Indigenous language speakers (Williams et al., 2021). The design context of remote Indigenous communities in North Australia is quite different from mainstream culture, urban corporations and the tools that have evolved to serve them. There is rising understanding of the need for technology not to substitute for interaction but rather to support relationships Taylor et al. (2019). A common point of agreement is that video communication supports work practice and relationship maintenance. As a parallel investigation into the general problem of working consultations in remote communities, we developed an appliance-based video messaging service called Lingobox (Bettinson and Bird, 2021b). The design challenge we take up here is to integrate Lingobox as a means to support learner-interactions within a system for collaborative transcription practice.

3 Project Sparzan

In this section we describe a system for collaborative computationally-assisted speech transcription. It is a synthesis of prior work in speech transcription methodology, interactive transcription workflows and remote interaction through tangible technology. The system's working title Sparzan derives from **spars**e t**ran**scription, the fundamental transcription model we adopt here. Two use cases are supported: collaborative lexicon building and individual language acquisition. In both cases, speech is transcribed by non-native speakers as a way to expand their individual and collective understanding, which is why we call this learning through transcription.

The Sparzan architecture is given in Figure 1, comprising a web application including the transcription activity, backed by server 'stack' responsible for the data model logic, data storage and computational agents dispatched as asynchronous

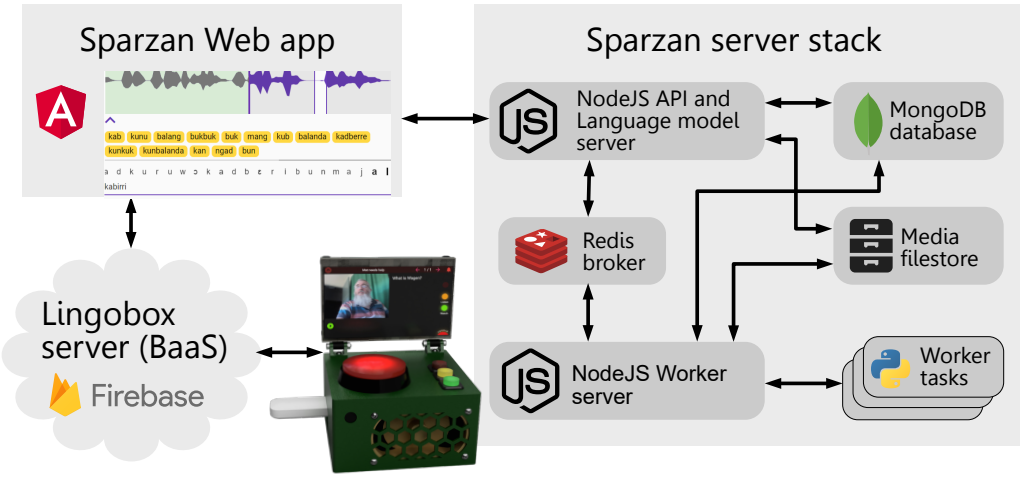

Figure 1: The Sparzan system architecture implementation is similar to common web services. The web-based transcription app is backed by centralised computational resources.

worker processes. Remote language consultation is achieved by integrating the external Lingobox service, hosted on a commercial backend-as-a-service (BaaS). In the following sections we elaborate on key components of the Sparzan system.

3.1 Sparse transcription for learning

Sparse transcription relies on tasks that are well suited to the competencies of the audience in our use case, i.e. as a series of 'interpretive, iterative and interactive processes that are amenable to wider participation' (Bird, 2020b, p713). However, the original proposal does not take asynchronous collaboration into account. For example the tasks of growing and refactoring a glossary (Tasks G and R) presume the existence of a single glossary that is in a perpetual state of motion towards a fully validated and authoritative state.

However, designs to support learning through transcription must recognise the existence of individual understandings of language, or more completely, individual dynamic language systems (De Bot et al., 2007). Systems for computer assisted language learning (CALL, Levy, 1997) often model the knowledge held by individual learners in order to craft personalised learning opportunities (cf. the Input Hypothesis, Krashen, 1992). While the individual strives to acquire the sum of generalised knowledge, the individual *transcriber* is also an actor in incremental processes to extend this knowledge. This observation is not limited to lexical knowledge either, but holds for any type of language knowledge that is being investigated, such as morphosyntax, or sociolinguistic variation. Thus, we need a way to differentiate curated knowledge and individual knowledge.

Accordingly, we extend the sparse transcription model to include two lexical data structures, the glossary (individual) and the lexicon (general). We anticipate one lexicon per language variety but multiple glossaries (one per transcriber). Now we refine the refactoring task to be non-destructive to the glossary, instead using content from the glossary when creating or updating lexical entries during consultation with a language authority.

The lexicon is useful for computationally assisted learning. We use phone-based keyword spotting to identify plausible instances of words in a speech segment and offer a list of suggestions to the learner without prejudice. Should the learner know the word, they may accept the suggestion, thereby establishing a link from the learner's glossary to the lexical entry. Otherwise, if the word be unknown, the suggestion is treated as a learning prompt that supports an exploration of meaning (lexical definition) and usage (concordance views). Crucially, these learning prompts need not be correct; phone-based word spotting errors are typically plausible learner errors in their own right, and thus they are useful as practice to discriminate similar-sounding words (as noted in, Bettinson and Bird, 2021a).

In sum, 'transcription for learning' is a form of computationally-assisted self-study. Learner-speaker interactions are an essential complement to self-study but we must also recognise the reality that time with speakers is limited. Anchoring learner-speaker interaction in context supports the systematic capture of speaker knowledge, reducing

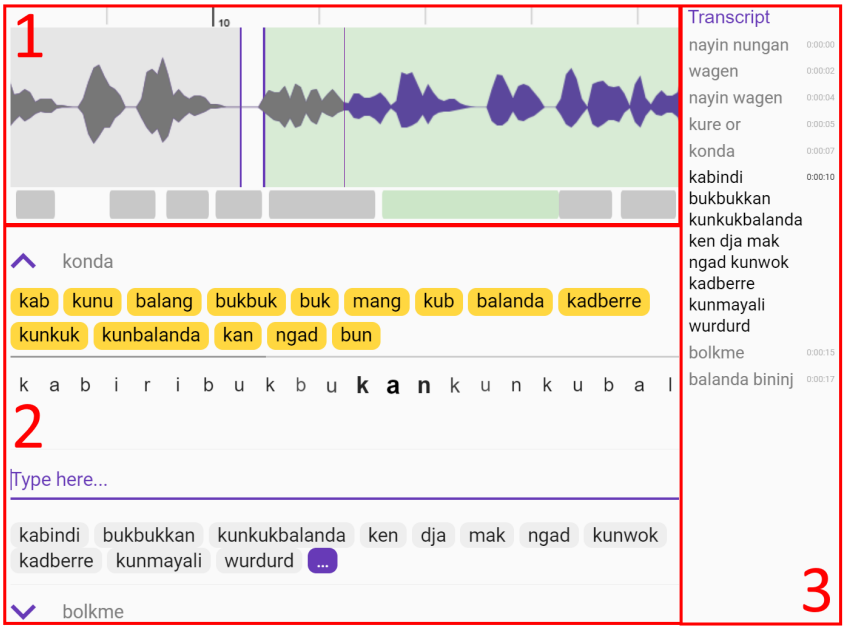

Figure 2: The Sparzan transcriber activity. User interface zones are annotated in red: 1. the signal zone, 2. the segment zone and 3. the transcript zone

repetition and improving learning efficiency and the acquisition of valuable language data. In the next section we illustrate how this is achieved in a transcription activity.

3.2 Sparzan transcriber

In this section we describe the transcription activity of the web app. The design is inspired by prior work in simplified transcription interfaces, particularly those with a focus on oral workflow support such as SayMore and Aikuma-NG. Simplicity is at the heart of the design goals for Sparzan Transcriber for two reasons: as general tactic to increase usability (Nielsen, 1994); and to support transcription as a common resource where consultants and transcribers work together. A natural consequence of co-located tool use is that observers learn to become operators. That is an important design consideration to support the self-sufficiency and digital agency rights of Aboriginal Australians (Carew et al., 2015).

Sparzan Transcriber is split into three zones (Fig. 2): signal, segment and transcript zones. Each zone displays a mapping against time and affords a method of navigation in the media file. The signal zone maps time horizontally, comprising a scrolling waveform (10 second window) and a fixed voice activity (VAD) bar underneath (entire duration). The segment zone maps time vertically, displaying a single speech segment at a time derived from an initial automatic voice activity segmentation. The transcript zone is a vertical map of timestamped transcriptions to offer an uncluttered context of transcription content.

Transcription is achieved by consulting the assistance offered in the segment zone, and typing into a temporary text input box. The current prototype offers phonemically word spotted candidates (yellow chips in Fig. 2) and an automatic phonemic transcription (the text line underneath). When the segment changes, audio playback begins automatically and is reflected in the scrolling signal view and in an animated phone display that highlights phones associated with the current point of playback (visible as the bold ‘kan’ in Fig. 2).

In contrast to many other transcription tools, transcriptions are not free text. They are a sequence of word tokens rendered as ‘chips’ so as to support interaction such as querying lexical entries and viewing a concordance of projects associated with the lexicon. A transcription is a sequence of word token chips with an temporary chip mapped to current text input. The transcriber interacts via the temporary text input box and left/right cursors to achieve insert, edit and delete operations.

The operator may request help from a native speaker via the Lingobox service (Sec. 3.4). To do so the operator creates a new Lingobox request by recording a webcam video and customising the Lingobox prompt (Fig. 3). Customisations include

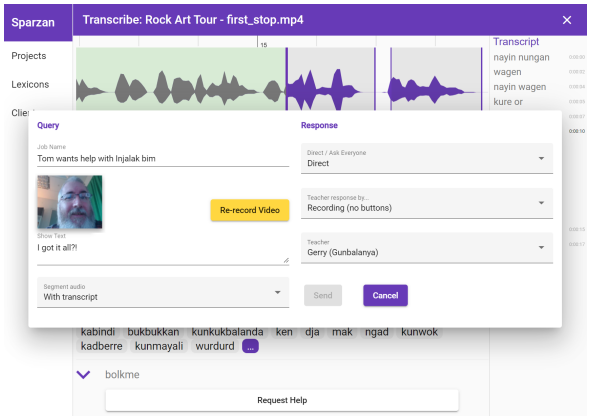

Figure 3: Requesting speaker assistance within a Sparzan transcriber session

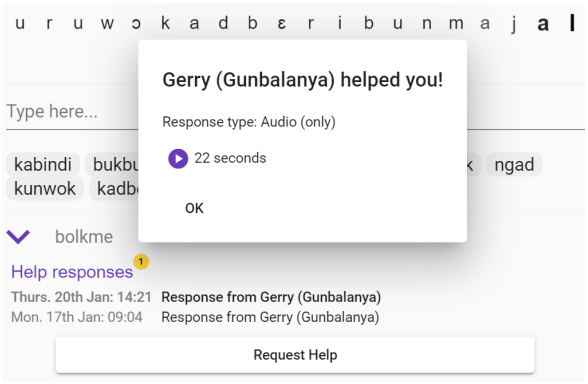

Figure 4: Speaker responses are anchored to transcription spans in Sparzan transcriber

on-screen text, including the segment audio, the recipient of the request (if there are multiple language authorities) and what the type of desired response. Typically one would ask a question about the audio of a current segment, such as confirming a transcription choice. When the speaker has provided a response at a later time, the Sparzan web app indicates the response against a given transcription and clicking on the notification takes the operator directly to the speech transcription segment to access the response (Fig. 4).

3.3 Sparzan server

The server stack comprises two Node.js server apps: a main business logic server, and a 'worker' app. The main server app implements data model transactions and handles client interaction through HTTP and WebSocket APIs. The worker application brokers computational workload via jobs dispatched as asynchronous worker threads, injecting the results back into the database and notifying the main server app on their completion. Structured data is stored in a MongoDB instance while job persistence and inter-app communications are achieved via an in-memory database (Redis).

When the client uploads new media, a media processing job is created which batches up a number of vital tasks: extracting audio peak information, breath group segmentation, automatic phonemic transcription via Allosaurus, customised for Kunwinjku (Li et al., 2020; Le Ferrand et al., 2021) and finally phonemic word spotting of existing lexical entities. These operations must complete before a transcription session may begin, however a phone-based word spotting task is also created, with results to appear in Sparzan transcriber as that operation completes. During normal operations, additional word spotting jobs are executed periodically so that newly added items appear in the candidate suggestions.

User experience of server-backed web applications is highly dependent on network performance. Sparzan's backend supports real-time transcription with a server-based data model with a low-latency WebSocket API transport. This is sufficient to support client transcription in regional centres, but is impractical for supporting transcription sessions in remote communities. On a provisional basis, we support remote community participation through the Lingobox appliance described in the next section. This service utilises a data synchronising strategy designed to function adequately on 'bush internet'.

3.4 Lingobox

Lingobox is an appliance designed to support consulting interactions framed as personal video requests (see Fig. 5) and intended to be deployed in community workplaces, such as language and arts centres, where it behaves much like an answering machine. It was developed to explore an effective replacement for paid consulting interactions that would usually take place in the course of face-to-face fieldwork (Bettinson and Bird, 2021b). We opted to design a custom appliance to solve a number of intractable limitations of mobile devices that have emerged from several years of experience developing stand-alone and server-connected mobile apps. Limitations include poor audio on mobile devices, the need to manage and secure devices, and the low effectiveness of attention strategies such as notifications, and the general lack of prominence and association with a place of work.

Figure 5: Lingobox, an appliance to support remote interaction in language work

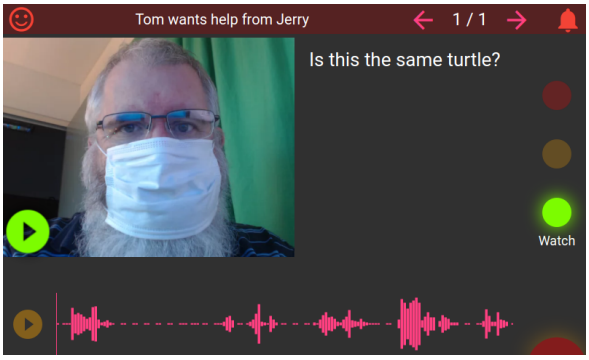

Figure 6: Lingobox on-screen display for a typical transcription-anchored help request

The hardware features illuminated buttons, a tilt-adjustable LCD screen, a integrated cellular modem, and high quality audio recording and playback. Distinct from depersonalised crowdsourcing techniques, the intent here is to support learner-teacher relationships and to place the burden of effort on the person asking for help. New requests are added to a stack of requests with an audible ping, and the large red recording button periodically flashes when there are unanswered requests. Requests are created within the context of ongoing transcription work, and they typically (but optionally) include the segment of audio that is currently being transcribed. Consultants act on requests through a staged process such as: playing the video request, playing the media (e.g. the segment of audio being transcribed), eliciting a spoken response, and conveying it back (Fig. 6). Sparzan provides a rudimentary form of workspace awareness (Gutwin and Greenberg, 1996) to draw attention towards consultant responses. This is achieved through notifications that draw the user directly to the relevant transcription segment.

4 Discussion

Many people have noted the pressing need to bootstrap data collection for primarily oral, local languages. This is largely a human effort, but it can be scaled up with the support of efficient workflows and assistive technologies. For example, having a searchable lexicon facilitates updating individual entries. However, piecemeal solutions, where we improve the efficiency of individual tasks, only deliver incremental improvements. Fully integrated solutions allow us to explore broader questions and to exploit fortuitous opportunities.

In the usual pattern of language teaching, learning resources are compiled by 'experts'. Learners with their repetitive mistakes and frequent errors have no role to play in crafting lessons. However for local languages there may be few learning resources of the type expected by western learners, and minimal capacity for creating such resources. Nevertheless, there are still learning resources to be found. In particular, word recognition errors – obstacles for transcription – are plausible learner errors. These are not useless mistakes to be discarded, but potential prompts for learners to consider and correct. *A word that one person has learned and systematically corrected in the course of transcription may become a prompt for another learner.*

The system remains a prototype and we have not yet not been able to test it on the ground because the communities remain closed. This work has some other limitations. Foremost is that the design has been conducted in a university lab, rather than in a co-design process with our partners in the community. A second shortcoming is that the learning potential of collaborative transcription has yet to be explored. Using the resulting data to create learning content for dedicated learning apps is promising direction for future work.

A third shortcoming is the architectural reliance on low-latency networks. Solutions of this type require server infrastructure, but 'bush internet' network conditions rule out the simple convenience of the 'cloud'. One solution is to deploy compact, low-

cost server architecture in the field, as we have previously explored with BushPi, a ruggedised battery-powered server to support local use of collaborative language apps (Bettinson and Bird, 2021a). Thus, there is an ongoing need for on-country design and engineering to devise practical, community-based solutions, building on previous attempts in this space (Cathcart et al., 2012; Hanke, 2017).

Despite these shortcomings, we believe this work amounts to a novel and effective design pattern for remote asynchronous collaboration on meticulous language work, serving a variety of documentary and pedagogical goals. The potential for computer supported cooperative language work remains relatively unexplored, and faces an egregious challenge: how do we go from minimally-viable research prototypes to robust, supported, and sustainable solutions (cf. Finlayson, 2016, p27)? We believe there remains a clear need for foundational research on technology for working with primarily oral, local languages, supporting a broad range of stakeholders, for the benefit of community goals in sustaining linguistic diversity.

5 Conclusion

We have described a system for cooperative language work, including speech transcription and language learning. It was developed during a period where Aboriginal community interactions were severely limited due to the COVID-19 pandemic, but where cooperative work and the underlying relationships needed to be sustained. We developed assistive technology to support (and even encourage) language acquisition in the course of transcription and the associated learner-speaker interactions. We set aside expert-defined practice, and instead designed for inclusive participation of learners and speakers, regardless of their technical competencies. In the process, we have demonstrated that the effort of systems engineering for specific sociolinguistic contexts has direct relevance for language data collection and for local language technologies in general.

Acknowledgements

We are grateful to the Bininj people of Northern Australia for the opportunity to work with them on the Kunwinjku language (ISO gup). This research has been supported by a grant from the Australian Research Council entitled *Learning English and Aboriginal Languages for Work*, and the Indigenous Languages and Arts Program entitled *Mobile Software for Oral Language Learning in Arnhem Land*. Our work with Bininj is covered by a research permit from the Northern Land Council and approvals from the board of Warddeken Land Management and the CDU Human Research Ethics Committee.

References

Oliver Adams. 2017. *Automatic Understanding of Unwritten Languages*. Ph.D. thesis, University of Melbourne.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Mat Bettinson. 2015. Towards Language Documentation 2.0: Imagining a Crowdsourcing Revolution. Presentation at the International Conference on Language Documentation and Conservation, http://hdl.handle.net/10125/25302.

Mat Bettinson. 2020. *Enabling Large-Scale Collaboration in Language Documentation*. PhD thesis, University of Melbourne, Melbourne, Australia.

Mat Bettinson and Steven Bird. 2017. Developing a suite of mobile applications for collaborative language documentation. In *Second Workshop on Computational Methods for Endangered Languages*, pages 156–164.

Mat Bettinson and Steven Bird. 2021a. Collaborative fieldwork with custom mobile apps. *Language Documentation and Conservation*, 15:411–432.

Mat Bettinson and Steven Bird. 2021b. Designing to support remote working relationships with indigenous communities. In *Proceedings of OzChi '21: 33rd Australian Conference on Human-Computer Interaction*.

Steven Bird. 2020a. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, page 3504–19, Barcelona, Spain.

Steven Bird. 2020b. Sparse transcription. *Computational Linguistics*, 46:713–44.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Steven Bird, Florian R Hanke, Oliver Adams, and Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–5. Association for Computational Linguistics.

Claire Bowern. 2008. *Linguistic Fieldwork: A Practical Guide*. Palgrave Macmillan.

Margaret Carew, Jennifer Green, Inge Kral, Rachel Nordlinger, and Ruth Singer. 2015. Getting in touch: Language and digital inclusion in Australian indigenous communities. *Language Documentation and Conservation*, 9:307–23.

MaryEllen Cathcart, Gina Cook, Theresa Deering, Yuliya Manyakina, Gretchen McCulloch, and Hisako Noguchi. 2012. LingSync: A free tool for creating and maintaining a shared database for communities, linguists and language learners. In *Proceedings of FAMLi II: Workshop on Corpus Approaches to Mayan Linguistics*, pages 247–250.

Vivian Cook. 2016. *Second Language Learning and Language Teaching*. Routledge.

Kees De Bot, Wander Lowie, and Marjolijn Verspoor. 2007. A dynamic systems theory approach to second language acquisition. *Bilingualism: Language and cognition*, 10:7–21.

Mark Alan Finlayson. 2016. Report on the 2015 NSF Workshop on Unified Annotation Tooling. Technical report, Computer Science and Artificial Intelligence Laboratory (MIT). http://hdl.handle.net/1721.1/105270, accessed February 2022.

Mark Gales, Kate Knill, Anton Ragni, and Shakti Rath. 2014. Speech recognition and keyword spotting for low-resource languages: BABEL project research at CUED. In *Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 16–23. ISCA.

Alvin Garcia and Herbert Gish. 2006. Keyword spotting of arbitrary words using minimal speech resources. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing*. IEEE.

Carl Gutwin and Saul Greenberg. 1996. Workspace awareness for groupware. In *Conference Companion on Human Factors in Computing Systems*, pages 208–209.

Florian R Hanke. 2017. *Computer Supported Collaborative Language Documentation*. Ph.D. thesis, University of Melbourne. Http://hdl.handle.net/11343/192578.

Mark A Hasegawa-Johnson, Preethi Jyothi, Daniel McCloy, Majid Mirbagheri, Giovanni M di Liberto, Amit Das, Bradley Ekin, Chunxi Liu, Vimal Manohar, Hao Tang, et al. 2016. ASR for under-resourced languages from probabilistic transcription. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 25:50–63.

John Hatton. 2013. SayMore: Language documentation productivity. Talk at 3rd International Conference on Language Documentation and Conservation (ICLDC3). http://hdl.handle.net/10125/26153.

Setrag Khoshafian and Marek Buckiewicz. 1995. *Introduction to groupware, workflow, and workgroup computing*. John Wiley & Sons, Inc.

Stephen Krashen. 1992. The input hypothesis: An update. *Linguistics and Language Pedagogy: The State of the Art*, pages 409–431.

William Lane, Mat Bettinson, and Steven Bird. 2021. A computational model for interactive transcription. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pages 105–111.

Éric Le Ferrand, Steven Bird, and Laurent Besacier. 2021. Phone based keyword spotting for transcribing very low resource languages. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, pages 79–86.

Michael Levy. 1997. *Computer-assisted language learning: Context and conceptualization*. Oxford University Press.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, et al. 2020. Universal phone recognition with a multilingual allophone system. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 8249–8253. IEEE.

Federico Marinelli, Alessandra Cervone, Giuliano Tortoreto, Evgeny A Stepanov, Giuseppe Di Fabbrizio, and Giuseppe Riccardi. 2019. Active annotation: Bootstrapping annotation lexicon and guidelines for supervised NLU learning. In *Proceedings of Interspeech 2019*, pages 574–578.

Felicity Meakins, Jenny Green, and Myfany Turpin. 2018. *Understanding Linguistic Fieldwork*. Routledge.

Jakob Nielsen. 1994. *Usability Engineering*. Elsevier. This is the bible for any discussion around producing easy-to-use products. It is vital to cite this is if one is talking about usability.

David Nunan. 1999. *Second Language Teaching & Learning*. Heinle & Heinle Publishers.

Walter Ong. 1982. *Orality and Literacy: The Technologizing of the Word*. Routledge.

Racquel-María Sapién. 2018. Design and implementation of collaborative language documentation projects. In *Oxford Handbook of Endangered Languages*, pages 203–24. Oxford University Press.

Jennyfer Lawrence Taylor, Wujal Wujal Aboriginal Shire Council, Alessandro Soro, Paul Roe, and Margot Brereton. 2019. A relational approach to designing social technologies that foster use of the kuku yalanji language. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*, OZCHI'19, pages 161–172. Association for Computing Machinery.

Nicholas Thieberger. 2016. Language Documentation Tools and Methods Summit Report. Technical report, Centre of Excellence for the Dynamics of Language (CoEDL). http://bit.ly/LDTAMSReport, accessed February 2022.

Nicholas Williams, W. D. L. Silva, Laura McPherson, and Jeff Good. 2021. Covid-19 and documentary linguistics: Some ways forward. *Language Documentation and Description*, 20:359–377.

Marcely Zanon Boito, Alexandre Bérard, Aline Villavicencio, and Laurent Besacier. 2017. Unwritten languages demand attention too! Word discovery with encoder-decoder models. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 458–65.

John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research Through Design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 493–502. ACM.