Local Word Discovery for Interactive Transcription

William Lane and Steven Bird
Northern Institute
Charles Darwin University

Abstract

Human expertise and the participation of speech communities are essential factors in the success of technologies for low-resource languages. Accordingly, we propose a new computational task which is tuned to the available knowledge and interests in an Indigenous community, and which supports the construction of high quality texts and lexicons. The task is illustrated for Kunwinjku, a morphologically-complex Australian language. We combine a finite state implementation of a published grammar with a partial lexicon, and apply this to a noisy phone representation of the signal. We locate known lexemes in the signal and use the morphological transducer to build these out into hypothetical, morphologically-complex words for human validation. We show that applying a single iteration of this method results in a relative transcription density gain of 17%. Further, we find that 75% of breath groups in the test set receive at least one correct partial or full-word suggestion.

1 Introduction

In over a century of practice in descriptive linguistics, the pattern has been to prepare texts and a lexicon to support the construction of a grammar. The grammar includes a description of the phonology and morphosyntax, which inform the representation of the texts and lexicon, in a cyclic arrangement (Crowley, 2007, 139f). The three types of data are entwined in the so-called "Boasian trilogy" of texts, lexicon, and grammar.

More recently, another tradition of working with little-studied languages has grown up in the language technology community. It frames these as "low resource languages," lacking the text, speech and lexical resources that are needed for creating speech and language technologies (Krauwer, 2003). In many cases, these languages are not little-studied at all, it is just that the technological methods can only exploit texts and lexicons, not the grammar.

This brings us to the question: how can we leverage a grammar when working with a low resource language? In particular, how can we leverage a morphosyntactic description to accelerate the creation of texts and a lexicon for a morphologically complex language?

Our approach complements the practice of "learning to transcribe" (Bird, 2020), where non-speaker transcribers train themselves to recognize words in connected speech. We assume that transcribers are able to sparsely annotate spans of audio with any words they recognize. These words can be aligned with the output of an automatic phone recognizer, and the machine suggests new words conditioned on phones in the locus of known words (Fig. 1). We call this task *local word discovery*.

In the case of low-resource languages like Kunwinjku (ISO gup), we do not have enough text to train a language model to guide the suggestion of words in the locus of previously recognized words. However, as a morphologically-complex language with a published grammar, we do have information at the level of morphemes. Thus, we employ a morphological transducer to map previously recognized morphs with the surrounding noisy phone sequences to new morphologically-complex wordforms for manual verification. The constituent morphs of confirmed words are then added to the lexicon. Figure 2 shows the proposed local word discovery pipeline, which we deploy in a prototype interactive transcription system. We

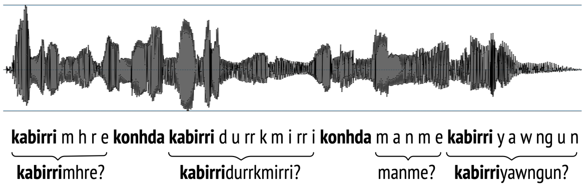
kabirri m h r e konhda kabirri d u rr k m i rr i konhda m a n m e kabirri y a w ng u n
kabirrimhre?
kabirridurrkmirri?
manme?
kabirriyawngun?

Figure 1: Local Word Discovery: A mix of morphs and phones have been recognized, and combined into hypothetical words

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2058–2067
November 7–11, 2021.

test the system with speakers of Kunwinjku.

The main contributions are: a new word discovery task which cultivates a morph lexicon; a new, low-friction, interactive speech transcription workflow for low-resource morphologically-complex languages which leverages local word discovery; and a prototype implementation that integrates a universal phone recognizer with a morphological transducer.[1].

We begin with a review of related work (Sec. 2), followed by an overview of the proposed task of local word discovery and our implementation of a model which performs this task (Sec. 3). We then explain how we set up an evaluative experiment of the model (Sec 4), and give results (Sec. 5), followed by conclusions (Sec. 6).

2 Previous Work

Early work on computer-assisted speech transcription grew out of the increasing effectiveness of automatic speech recognition (ASR) systems for resource-rich languages. For example, Nanjo et al. (2006) trained an ASR system on 228 hours of transcribed speech from the National Congress of Japan. Word recognition errors are manually corrected using various interfaces: multiple choice selection from confusion pairs, respeaking, and manual correction.

Subsequent work continues to build in human post-editing of increasingly accurate ASR output (Luz et al., 2008; Sanchez-Cortina et al., 2012). Thanks to their reliance on ASR, these systems depend on lexicons and large amounts of transcribed speech for training. The lack of performant ASR systems for low-resource languages makes this approach ill-suited to the linguistic documentation use case; we can only automate the first stage of the ASR pipeline, namely phone recognition.

2.1 Phone recognition

Phone recognizers have been able to produce impressive results in low-resource situations. For example, the Persephone system was trained on 50 minutes of phonetically-transcribed Chatino speech, and reached a 20% phone error rate. Trained on 224 minutes of phonetically transcribed Na speech, it reached a phone error rate of 11% (Adams et al., 2018). Their results suggest that as little as 30 minutes of phonetically transcribed speech are needed to achieve sub-30% phone error rate. Many others have been exploring this approach (Besacier et al., 2014; Adams, 2017; Dunbar et al., 2017; Littell et al., 2018; Jimerson and Prud'hommeaux, 2018; Adams et al., 2019).

Allosaurus provides a large pre-trained model tuned on speech from over 2,000 languages, allowing us to leverage learned parameters from a large amount of training data (Li et al., 2020). The technique of fine-tuning multilingual models to achieve better performance on lesser-resourced languages is well attested in areas such as universal machine translation and language modeling (Gu et al., 2018; Eisenschlos et al., 2019).

While most acoustic models handle multilingual data by taking the union of phoneme sets across languages, Allosaurus adds an allophone layer which maps narrow phone sets in one language to the phonemes of another. For example in English, all instances of [p] and [p^h] would map to *p*, while in Mandarin Chinese they would be kept distinct. As a result, there can be more consistent learning of similar sounds across languages. However, phone recognition falls far short of the word recognition required for transcription.

2.2 Word recognition

Phone sequences may be split into word-like units or "pseudowords" using unsupervised or semi-supervised methods (Johnson et al., 2006; Johnson and Goldwater, 2009; Sirts and Goldwater, 2013; Eskander et al., 2016) or with reference to a translation (Neubig et al., 2012; Adams et al., 2015; Godard et al., 2016, 2018). The hope is that manual conversion of pseudoword sequences to word sequences would be less onerous than entering a transcription from scratch.

Besacier et al. (2006) describe one such word discovery algorithm for Iraqi Arabic which leverages mutual information between consecutive phones along with word frequency counts to iteratively discover frequent pseudowords. They trained a language model and apply it on unsegmented data to infer the most likely segmentation.

They performed an extrinsic evaluation of the method in a speech-to-text system, where they found that simulating human supervision of the word discovery task by incorporating a lexicon of high-frequency known words led to better BLEU scores as well as a much smaller working lexicon—

[1] The finite state implementation of local word discovery and the interactive transcription demo which deploys it can be found at `https://cdu-tell.gitlab.io/tech-resources/`

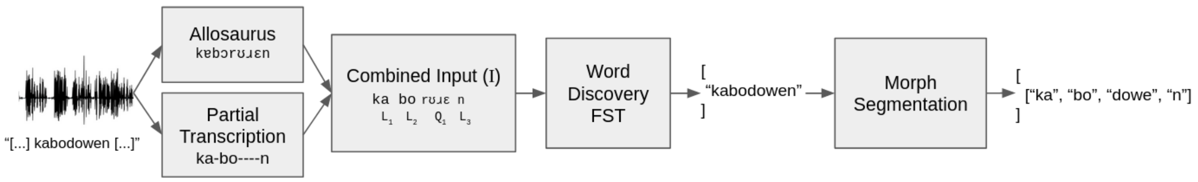

Figure 2: A local word discovery pipeline for morphologically complex words

2,200 words as opposed to 36,000 for the unsupervised phone-based approach—while maintaining the same translation coverage.

Zanon Boito et al. (2017) explored semi/supervised methods to discover words from unsegmented text in Mboshi using encoder decoder models. They obtained 27% of the target vocabulary, training on 5k sentences.

3 Local Word Discovery

This research takes place in the context of a series of engagements with the Bininj community of West Arnhem, in the far north of Australia. The community is centered in the town of Gunbalanya and a network of outstations, and predominantly speaks Kunwinjku. Schools, ranger programs and arts centres employ local people in cultural work where literacy in Kunwinjku is considered desirable, though not yet well established.

Kunwinjku has limited electronic texts and lexicons, but there is a comprehensive grammar (Evans, 2003). Transcription in this context is unavoidably collaborative, with a non-speaker transcriber working with a speaker and acquiring some of the language in the process (Rice, 2011; Hanke, 2017; Meakins et al., 2018). The non-speaker transcriber can transcribe familiar words in a first pass, and later prompt a speaker to produce any unrecognized words so they can be added to the lexicon and spotted automatically. Over time, they become part of the vocabulary of the non-speaker transcriber, who is able to confirm their appearance more readily in future.

Such transcription work is held up by the presence of unknown words, disfluencies, coarticulation, and noise. It is wise to skip difficult passages at first, and transcribe words that can be easily recognized, only later coming back to fill in the gaps once the priorities for careful, contiguous transcription have been established. This practice has been called *sparse transcription* (Bird, 2020).

Sparse transcriptions become contiguous through iterative, interactive processes such as collaborative work with speakers, or by leveraging word spotting techniques to detect other instances of identified lexemes across a larger corpus.

Sparse transcription serves a number of real-world use cases aside from contiguous transcription, e.g. spotted words serve as an index into the audio, facilitating keyword-based retrieval across large corpora; and lexical entries and associated metadata can be used in language learning.

3.1 Task definition

The starting point for local word discovery is an audio file, preprocessed using a phone recognizer (Li et al., 2020; Adams et al., 2018). We view the output as a noisy, low-dimensional representations of the signal (Figure 3, line Q).

We assume an early transcription scenario, where non-speaker transcribers are learning to transcribe the language. The audio is manually annotated with lexemes that non-speaker transcribers can confidently recognize. For example, in Figure 3 line L shows some morphs that were recognized by non-speaker transcribers (and identified as lexemes L_i), automatically aligned to the output of the phone transcriber. Recognized lexemes are combined with line Q to produce a sparsely-transcribed phone sequence which serves as the input to the local word discovery algorithm. The residue of unrecognised phone spans are labelled Q_i. Local word discovery accepts input I, and returns a list of legal, morphologically-complex words, anchored to the phone sequence (e.g. Figure 4).

3.2 Local word discovery in interactive transcription

In the sparse transcription model, partial transcriptions are stored as entries in a glossary along with pointers to all other instances of the entry across a wider corpus (Bird, 2020). This data structure is conducive to training a model for word spotting, which can identify other instances of a glossary

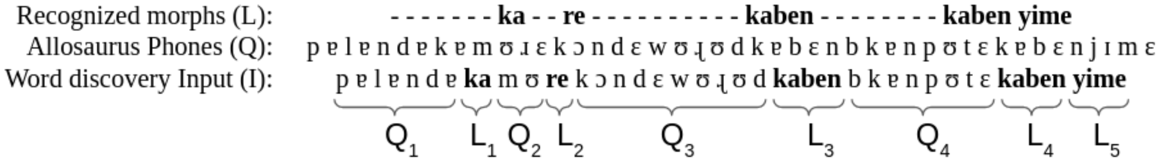
Recognized morphs (L): - - - - - - - ka - - re - - - - - - - - - - kaben - - - - - - - - kaben yime
Allosaurus Phones (Q): p ɐ l ɐ n d ɐ k ɐ m ʊ ɹ ɛ k ɔ n d ɛ w ʊ ɻ ʊ d k ɐ b ɛ n b k ɐ n p ʊ t ɛ k ɐ b ɛ n j ɪ m ɛ
Word discovery Input (I): p ɐ l ɐ n d ɐ ka m ʊ re k ɔ n d ɛ w ʊ ɻ ʊ d kaben b k ɐ n p ʊ t ɛ kaben yime
Q1
L1
Q2
L2
Q3
L3
Q4
L4
L5

Figure 3: Given an utterance, we assume a small lexicon of morphs which can be recognized by the non-speaker transcriber. Additionally, we assume an automatic phone transcription of the audio, e.g., from Allosaurus. Combining these two resources, we form the input to the proposed word discovery system.

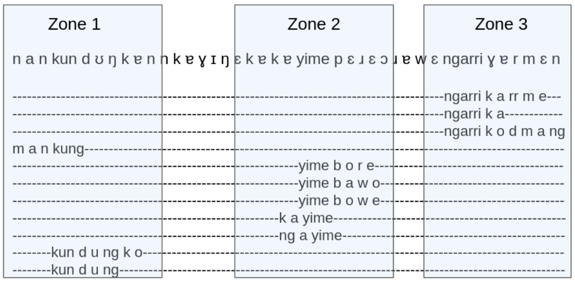

Figure 4: An automatic phone transcription, with the top results from the word discovery system. Each zone contains a set of predicted words that share an attested lexeme. In the lexical confirmation task, transcribers select the correct transcription from the list, if available.

entry across the whole corpus.[2]

"Local word discovery" offers a complementary mode of interactivity, as follows: human transcribers apply their own mental lexicon to transcribe sparsely, and local word discovery seeks to fill in gaps interactively as the transcriber works. Accepted word forms are added to the lexicon, and word spotting finds instances of lexemes across the whole corpus. Local word discovery is then applied in the loci of newly spotted lexemes, generating new words for the transcriber to confirm (Figure 5). The interactive feedback loop of local word discovery, global word spotting, and interactive confirmation comprise a novel transcription workflow, amplifying human effort in producing contiguous transcriptions.

3.3 Implementation

Given the task definition, we implement a baseline version of local word discovery using an FST to map attested morphs embedded in a noisy phone sequence to new, morphologically complex word forms. Our approach makes two assumptions, namely that the morphosyntactic description is sufficiently explicit and complete that it can be represented as an FST, and that a modest phone recognizer is available, e.g. by training a recognizer on a few hours of transcribed audio from related languages, or fine-tuning a larger pretrained model.

Speech representation. We adopt Allosaurus to provide a low-dimensional representation of speech which supports approximate matching against phone sequences predicted by the morphological transducer. Allosaurus provides a pretrained model which includes the ability to constrain the output vocabulary to a predefined set of phones (Li et al., 2020). The inventories of over 2,000 languages, including Kunwinjku, are supported in the default configuration. In practice, we found that the inventory for Kunwinjku was incomplete and we created our own, following (Evans, 2003).

Initial trials of Allosaurus on Kunwinjku produced unacceptably noisy representations, so we fine-tuned the model using 78 minutes of phonemically-transcribed spontaneous Kunwinjku speech (6 speakers). These are field recordings of speakers giving tours of their community, which include typical artifacts of natural speech including coarticulation, disfluency, and code switching.

We fine-tuned Allosaurus using k-fold cross-validation where k=6 (one fold per speaker, each time holding out one speaker's recordings for evaluation). After 50 epochs of fine-tuning we achieved the phone error rates shown in Figure 6. Across the 6 folds, we find that Allosaurus performs at an average phone error rate of 31.8%. This rate is acceptably good, given that we are not requiring high accuracy transcription, but an approximate representation to support the proposed local word discovery method.

Finite state word discovery. In order to perform word discovery on a stream of phones, we need a component capable of recognizing and performing morphological segmentation on full words. We

[2] See San et al. (2021); Le Ferrand et al. (2020); Chen et al. (2016); Yuan et al. (2017) for recent work on low resource word spotting models.

Step		Transcription	Lexicon
1.	**Task G.** Transcriber recognizes pronominal morph "ngarri" from audio and places it in phone stream	**ngarri** b ɪ m b u n m i ɲ j ɪ b ɪ m b u j i	{}
2.	**LWD.** Local word discovery completes the word, finding "ngarribimbun". The word is confirmed, and constituent morphs are added to the lexicon: {ngarri, bim, bu, n}	**ngarribimbun** m i ɲ j ɪ b ɪ m b u j i	{ngarri, bim, bu, n}
3.	**Task S.** Word spotting methods apply the updated lexicon to the audio and finds a potential second match for "bim"	**ngarribimbun** m i ɲ j ɪ **bim** b u j i	{ngarri, bim, bu, n}
4.	**LWD**. Local word discovery is applied again to find "yibimbuyi"; constituent morphs added to the lexicon	**ngarribimbun** m i ɲ **yibimbuyi**	{ngarri, bim, bu, n, yi_1, yi_2}

Figure 5: An interactive transcription workflow, which leverages local word discovery to increase the density of sparse transcriptions. We integrate the new task of local word discovery with existing Task S (word spotting) and Task G (growing the glossary) (Bird, 2020), to form a new interactive workflow.

Speaker	Time (hh:mm:ss)	PER
GN	00:12:21	.289
TG	00:09:21	.338
DY	00:23:28	.417
RN	00:15:35	.289
SG	00:10:27	.256
MM	00:07:44	.321
	Total: 01:18:56	AVG: .318

Figure 6: Allosaurus phone error rate (PER) for each speaker held out as validation, with the model fine-tuned on the remaining speakers.

have a detailed linguistic description (Evans, 2003), and an implementation of the morphology as an FST (Lane and Bird, 2019). This FST recognizes valid morphotactic sequences in Kunwinjku, and transduces to a morphological analysis. We can take the lower side of this transducer to obtain an FSA which recognizes the language of licensed surface forms of full words. From here, we build up the regular expression around the surface form to allow for the skipping of arbitrary characters on either side of the word. We have opted to output a padding character when we encounter characters which do not belong to the recognized word, for the purpose of retaining offset information. See Algorithm 1 for the implementation of the word discovery component in XFST format.

Accounting for phone noise. We used PanPhone (Mortensen et al., 2016) to acquire vector representations of each phone for both the Allosaurus representation and the orthographic-to-IPA mapping. We computed the cosine distance between each phone in both representations, forming a matrix of distance calculations. In the FST, we defined a transducer which maps from the orthographic character to a set of plausible phones. To define the set of phones per grapheme, we picked a cosine distance threshold of $K = .3$, and any phone below that threshold is deemed similar enough to be treated interchangeably with the canonical phone for that grapheme (Figure 7).

```
Define NoisyPhones  "n" (->) [ "n" | "m" | "ɲ" ].o.
                    "l" (->) [ "l" | "r" | "ɭ" ].o.
                    "r" (->) [ "r" | "l" | "ɭ" ].o.
                    "y" (->) [ "j" | "ɪ" | "ɛ" ].o.
                    "h" (->) [ "ʡ" ] ...
```

Figure 7: An excerpt from an implementation of the NoisyModel FST in XFST. The list of phones specified on the right is treated by the FST as being acceptably translated into the orthography on the lefthand side of the optional insertion operator.

4 Experiment Setup

We explore the concept of local word discovery on sparsely-transcribed audio by measuring the change in transcription densities before and after applying local word discovery implemented with the FST.

The first step is to define an initial lexicon which we use to sparsely transcribe a collection of audio. We used the collection of transcribed utterances from speaker SG as the test set, and the automatic phone recognition model which was fined tuned on all but SG's speech. From the transcriptions of SG's speech, we identify the 10 most frequent morphs and locate them in the speech. This is the input for word discovery (Figure 3, line *I*). This produced 126 annotated utterances: 126 breath groups represented by their phone stream, with individual tokens of the lexemes from the initial lexicon

Algorithm 1 Finite State Word Discovery from Sparsely Transcribed Input

```
Require: Grammar                    ▷ The FST which transforms a valid surface string into its morph analysis
  define WSpace [..] (->) “ ”;                              ▷ Optionally insert a single whitespace
  define LexA [Grammar .o. WSpace].l;         ▷ All surface forms optionally interspersed with a space
  define LexB [0:“ ”] LexA [0:“ ”];                      ▷ Suppress space on either side of lexeme
  define LexC [ “-”:? ]* LexB [ “-”:? ]*;    ▷ recognize lexemes, padding all non-member characters
  define LexD [LexC] .o. NoisyPhones;        ▷ Recognize LexC, flexibly allowing for phones in equivalence classes
```

aligned to the speech.

For each of the 126 lines of input, we calculated their baseline transcription density as the sum of character lengths of spotted lexemes divided by the sum of characters in the gold transcription. For example, if the utterance is “k a r ɪ **re**”, where “k”, “a”, “r”, and “ɪ” are phones and “re” is a lexeme, and its gold orthographic transcription is “karrire”, then the baseline transcription density of this utterance is $2/7$, or 28.6%.

We ran each of the 126 sparsely transcribed utterances through the local word discovery pipeline defined in Section 3.3. The output is a list of partial or full word completions, anchored in known lexemes (see Figure 4). For each utterance, we examined the suggestions and accepted those that were correct based on the gold transcription (simulating the manual confirmation of a speaker). We report the transcription density increase relative to the baseline density, since the model seeks only to increase density around the locus of existing annotations.

5 Experiment Results

Across the test set of 126 utterances, we found that 47.6% of them received correct, full word suggestions, and 75.4% received correct partial word suggestions.

Individual utterances varied widely in terms of baseline transcription densities, and how much local word discovery with an FST was able to contribute. In terms of characters transcribed solely by accepting full word suggestions anchored at the locus of known lexemes across all utterances in the corpus, we saw a transcription density growth of 17.34% relative to the baseline density. Summary statistics on the performance of local word discovery on the SG collection can be seen in Figure 8. Full word density is the number of characters transcribed before applying local word discovery plus the number of characters transcribed by accepting full word suggestions from the system output, over the the number of characters in the gold orthographic transcription. For individual utterances, we calculate the increase in transcription density, a subset of which can be seen in Figure 9.

SG Corpus	
Baseline Density (chars)	34.6%
Full Word Density (chars)	40.6%
Relative Increase	17.3%
% Breath Groups with Full Word Suggestions	47.6%
% Breath Groups with Partial Word Suggestions	75.4%

Figure 8: Summary statistics on the performance of local word discovery on the SG corpus of 126 utterances (breath groups).

Accepting only the full-word suggestions across the SG collection leads to the creation of 76 new *unique* entries for the lexicon. In the context of a full interactive transcription pipeline, this represents 76 new possible exemplars for lexeme spotting across the wider corpus. For reference, this experiment was seeded with just 10 unique glossary entries and produced more than 7 times that number of entries to seed a second round. As new instances of lexemes are confirmed, the local word discovery pipeline can be run again to discover more full-words around these new loci.

5.1 Relying on a human-in-the-loop

One of the weaknesses of the FST implementation of local word discovery is that allowing the transducer to treat any similar sounds as interchangeable opens up the space of possibly recognized words. The number of system suggestions could easily be detrimental to the transcription workflow, if the human-computer interaction is not properly

DocID	Input	Baseline Density	Full Word Density	Relative Increase	Correct Full Words	Correct Partial Words
doc73	ngarri k ʊ l u ng ng a ngarri ɟ u k m ɛ	59.3%	74.1%	25.0%	{ngarrikolung}	{ngarrikolu}
doc74	n u ng a n yimeng ng a r ɛ r ɛ t m ɛ k ɛ re b u kun w ʉ t a ng a ng a t u k	30.0%	46.0%	53.3%	{nungan, ngardduk}	{ngad}
doc75	w a l kun t u ng ka n ka n ka m u r ng ʊ a ɛ k u n b ɛ k f i t a m a k	32.0%	36.0%	12.5%	{kankan, kundung}	{kanka, dung, kan}
doc76	ngarri wok ng i m m ɛ n	80.0%	80.0%	0.0%	{}	{}
doc77	k ʊ kun w a t ɛ k ə n bu ngarri t r u k m i t i kabirrimarnbun ng a t b ɛ r ɛ	49.1%	61.4%	25.0%	{ngadberre}	{ngadberre, kunwardde, ngad}
doc78	b ɛ t ʃ r k a t i ɲ ɛ b a s t ʊ ng kayime t a n yiman	19.1%	19.1%	0.0%	{}	{}
doc79	k u r u ng kayime r a n m a r i ŋ ngarri w u m ɛ bonj	31.7%	43.9%	38.5%	{kurrung}	{kayimerran}
doc80	man m ɛ kabirri ŋ ʊ n bonj	75.0%	95.0%	26.7%	{manme, kabirringun}	{kabirringu, ngu, ngun}
doc81	w a ɲi ka b u l k b ɛ l ɛ t k m ɛ ngarri b ɛ n b ɛ ng k a n k ʊ b ʊ b a i	26.1%	26.1%	0.0%	{}	{bengka, kab, kabo, kan}
doc82	n a ʔ n ɛ kore b ɛ ng ʊ l ɛ yime k i ɛ k w ʊ ɟ b a l a ɟ ɛ n	32.50%	37.5%	15.4%	{ken}	{na}
doc83	ka r i yime kore b u a t ɛ r r birri k ʊ	42.9%	51.4%	20.0%	{karriyime}	{}

Figure 9: A sample of the results from the local word discovery experiment on the test set of 126 breath groups by speaker SG. Percent increase is calculated relative to the baseline transcription density.

handled. Accordingly, we implemented a transcription system which uses local word discovery to assist the transcriber, providing word suggestions per keystroke. In this implementation, it is often the case that the FST hallucinates an unmanageable number of suggestions conditioned on a fuzzy interpretation of the phone stream.

One solution is to rely on the human who is providing interactive feedback in real time. For example, suppose "kabirri" is a transcribed lexeme in a stream of phones. Local word discovery finds 10 possible continuations that are consistent with the following phone stream. As the user considers suggestions and continues to type, the system filters the suggestions to match. So, "kabirrib" yields 7 results, "kabirribu" yields just 3. In this example, the correct transcription "kabirribukkan" is present in all result sets.

5.2 Community experience

The automatic evaluation of local word discovery results in sets of hypothesized transcriptions for sub-spans of audio. The non-speaker transcriber can leverage interactivity with the model to give their best first pass transcription, and prioritize more difficult passages for confirmation with a speaker. The task for speakers of the language then is one of confirmation: presented with pre-prioritized and pre-scoped spans of audio, they confirm hypothesis derived through the human-computer interaction.

With this in mind, we visited the township of Gunbalanya and sat with a speaker, SB, for a transcription session. Our primary goal with this engagement was to assess whether the task of confirming prioritized work was a low-friction entry point to transcription work for speakers who have no experience with transcription. SB is a young adult, fluent in the language, yet not confident with reading and writing. He expressed uncertainty as to whether he was qualified to assist with transcription, and he suggested that a community elder might be more suitable. We assured him that the task only involved listening to recordings and talking about what we heard. After this he agreed to participate.

We used the same SG collection of utterances which we used for the automatic evaluation of local word discovery. The output of the pipeline organizes suggestions by zones, where each anchor lexeme and its associated suggestions form a distinct zone grouping (e.g., Fig. 4). As we progressed through the zones, we played the associated audio region and discussed the available options for transcription. SB was soon joined by GM, a community elder who wanted to observe the collaborative transcription process. GM volunteered his insights as well, and encouraged SB to pursue language work such as this. "This is like education you know," GM said to SB, pointing to the computer we were using together.

We worked with these two speakers, each with different levels of confidence in the written language, and both were capable of participating in the task effectively. This suggests that this is a low-friction entry point to language work. The task is simple and it grows the lexicon with well-formed words attested in the speech corpus. Lower barriers to participation democratize the work of transcription, enabling asynchronous collaboration with speakers.

6 Conclusion

The literature on low-resource languages has framed such languages as lacking the required *texts* and *lexicons* for developing the usual suite of speech and language technologies. Recent work in this vein has generally not explored the use of published linguistic *descriptions*, the third linguis-

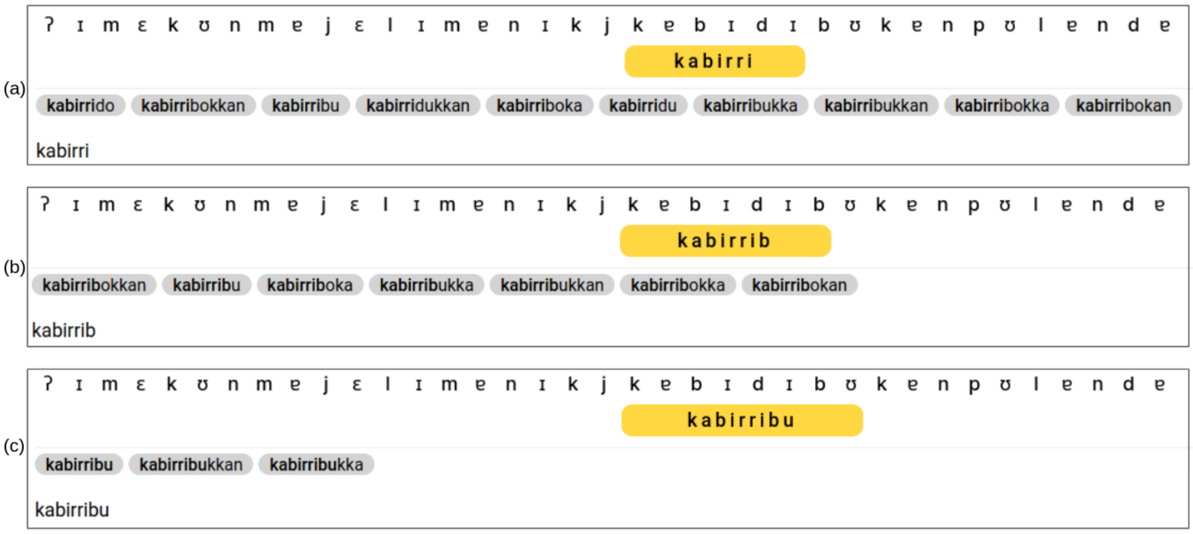

Figure 10: Screenshots from the interactive transcription system where we deploy local word discovery. As the user types, local word discovery is applied and suggestions appear below. The set of suggestions is refreshed per keystroke, so the user can let current suggestions guide their input choice and interactively receive updated suggestions based on continuing input choices. The target word, “kabirribukkan”, is present in all result sets.

tic data type in the Boasian trilogy, perhaps because descriptions are seen to require too much manual labour to convert into computational grammars, or because the resulting grammars are seen to be too brittle for working with natural speech.

Nevertheless, we believe such descriptions can play a role in supporting the creation of texts and lexicons, while reducing the dependence on language models. A description, suitably interpreted, can constrain the forms that can hypothesized in a given textual context, and this information can be used to inform (rather than limit) the choices made by human transcribers. In this paper we have explored this idea and applied it to a morphologically complex language.

We have proposed a new computational task of “local word discovery” to complement the practice of sparse transcription. We have discussed an approach to local word discovery that uses an existing morphological analyzer to process a sequence of known lexemes aligned to a noisy stream of phones. The method suggests possible completions of morphologically complex surface forms that are grounded at the locus of known lexemes and conditioned on the phonetic environment. On test data from Kunwinjku, local word discovery increases transcription density by 17.3% and contributes 76 new unique glossary entries. These new entries then serve as new loci for further iterations of local word discovery. These results show that local word discovery a promising means of generating transcription suggestions which grow the lexicon and produce more dense transcriptions.

This approach enables a novel transcription workflow where a non-speaker transcriber does a first pass, transcribing easily identifiable words, and a speaker comes along behind to work on the residue, while the system is performing word spotting and local word discovery in the background. We deployed the model in an interactive transcription system and tested it in the field and saw that local word discovery, together with the other stages of the new transcription workflow, enabled low friction interactions between transcribers and the system, speeding up the transcription process.

Acknowledgments

We are grateful to the Bininj people of Northern Australia for the opportunity to work in their community. Our thanks to several anonymous reviewers for helpful feedback on earlier versions of this paper. This research was covered by a research permit from the Northern Land Council, and was supported by the Australian government through a PhD scholarship, and grants from the Australian Research Council and the Indigenous Language and Arts Program.

References

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

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan. European Language Resources Association.

Oliver Adams, Graham Neubig, Trevor Cohn, and Steven Bird. 2015. Inducing bilingual lexicons from small quantities of sentence-aligned phonemic transcriptions. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 248–55, Da Nang, Vietnam.

Oliver Adams, Matthew Wiesner, Shinji Watanabe, and David Yarowsky. 2019. Massively Multilingual Adversarial Speech Recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 96–108, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Laurent Besacier, Bowen Zhou, and Yuqing Gao. 2006. Towards speech translation of non written languages. In *Spoken Language Technology Workshop*, pages 222–25. IEEE.

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

Hongjie Chen, Cheung-Chi Leung, Lei Xie, Bin Ma, and Haizhou Li. 2016. Unsupervised bottleneck features for low-resource query-by-example spoken term detection. In *INTERSPEECH*, pages 923–27.

Terry Crowley. 2007. *Field Linguistics: A Beginner's Guide*. Oxford University Press.

Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The Zero Resource Speech Challenge 2017. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pages 323–30. IEEE.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5702–07, Hong Kong, China. Association for Computational Linguistics.

Ramy Eskander, Owen Rambow, and Tianchun Yang. 2016. Extending the use of Adaptor Grammars for unsupervised morphological segmentation of unseen languages. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 900–10, Osaka, Japan. Association for Computational Linguistics.

Nicholas Evans. 2003. *Bininj Gun-wok: A Pan-dialectal Grammar of Mayali, Kunwinjku and Kune*. Pacific Linguistics. Australian National University.

Pierre Godard, Gilles Adda, Martine Adda-Decker, Alexandre Allauzen, Laurent Besacier, Helene Bonneau-Maynard, Guy-Noël Kouarata, Kevin Löser, Annie Rialland, and François Yvon. 2016. Preliminary experiments on unsupervised word discovery in Mboshi. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, pages 3539–43.

Pierre Godard, Marcely Zanon Boito, Lucas Ondel, Alexandre Berard, François Yvon, Aline Villavicencio, and Laurent Besacier. 2018. Unsupervised word segmentation from speech with attention. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, pages 2678–82.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 344–54, New Orleans, Louisiana. Association for Computational Linguistics.

Florian Hanke. 2017. *Computer-Supported Cooperative Language Documentation*. Ph.D. thesis, University of Melbourne.

Robbie Jimerson and Emily Prud'hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 4161–66, Miyazaki, Japan. European Language Resources Association.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparametric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 317–25, Boulder, Colorado. Association for Computational Linguistics.

Mark Johnson, Thomas Griffiths, and Sharon Goldwater. 2006. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. *Advances in Neural Information Processing Systems*, 19:641–48.

Steven Krauwer. 2003. The Basic Language Resource Kit (BLARK) as the first milestone for the Language

Resources Roadmap. *Proceedings of the International Workshop Speech and Computer*, pages 8–15.

William Lane and Steven Bird. 2019. Towards a robust morphological analyzer for kunwinjku. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 1–9, Sydney, Australia. Australasian Language Technology Association.

Eric Le Ferrand, Steven Bird, and Laurent Besacier. 2020. Enabling interactive transcription in an indigenous community. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3422–28, Barcelona, Spain. International Committee on Computational Linguistics.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan Black, and Metze Florian. 2020. Universal phone recognition with a multilingual allophone system. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 8249–53. IEEE.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–32.

Saturnino Luz, Masood Masoodian, Bill Rogers, and Chris Deering. 2008. Interface design strategies for computer-assisted speech transcription. In *Proceedings of the 20th Australasian Conference on Computer-Human Interaction: Designing for Habitus and Habitat*, pages 203–10.

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

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 3475–84. Association for Computational Linguistics.

Hiroaki Nanjo, Yuya Akita, and Tatsuya Kawahara. 2006. Computer assisted speech transcription system for efficient speech archive. In *Proceedings of the Western Pacific Acoustics Conference*, Seoul, Korea.

Graham Neubig, Taro Watanabe, Shinsuke Mori, and Tatsuya Kawahara. 2012. Machine translation without words through substring alignment. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 165–74. Association for Computational Linguistics.

Keren Rice. 2011. Documentary linguistics and community relations. *Language Documentation and Conservation*, 5:187–207.

Nay San, Martijn Bartelds, Mitchell Browne, Lily Clifford, Fiona Gibson, John Mansfield, David Nash, Jane Simpson, Myfany Turpin, Maria Vollmer, et al. 2021. Leveraging neural representations for facilitating access to untranscribed speech from endangered languages. *arXiv preprint arXiv:2103.14583*.

Isaias Sanchez-Cortina, Nicolas Serrano, Alberto Sanchis, and Alfons Juan. 2012. A prototype for interactive speech transcription balancing error and supervision effort. In *Proceedings of the International Conference on Intelligent User Interfaces*, pages 325–26.

Kairit Sirts and Sharon Goldwater. 2013. Minimally-supervised morphological segmentation using Adaptor Grammars. *Transactions of the Association for Computational Linguistics*, 1:255–66.

Yougen Yuan, Cheung-Chi Leung, Lei Xie, Hongjie Chen, Bin Ma, and Haizhou Li. 2017. Pairwise learning using multi-lingual bottleneck features for low-resource query-by-example spoken term detection. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5645–49. IEEE.

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