Sparse Transcription

Steven Bird
Northern Institute
Charles Darwin University
steven.bird@cdu.edu.au

The transcription bottleneck is often cited as a major obstacle for efforts to document the world's endangered languages and supply them with language technologies. One solution is to extend methods from automatic speech recognition and machine translation, and recruit linguists to provide narrow phonetic transcriptions and sentence-aligned translations. However, I believe that these approaches are not a good fit with the available data and skills, or with long-established practices that are essentially word-based. In seeking a more effective approach, I consider a century of transcription practice and a wide range of computational approaches, before proposing a computational model based on spoken term detection that I call "sparse transcription." This represents a shift away from current assumptions that we transcribe phones, transcribe fully, and transcribe first. Instead, sparse transcription combines the older practice of word-level transcription with interpretive, iterative, and interactive processes that are amenable to wider participation and that open the way to new methods for processing oral languages.

1. Introduction

Most of the world's languages only exist in spoken form. These oral vernaculars include endangered languages and regional varieties of major languages. When working with oral languages, linguists have been quick to set them down in writing: "The first [task] is to get a grip on the phonetics and phonology of the language, so that you can transcribe accurately. Otherwise, you will be seriously hampered in all other aspects of your work" (Bowern 2008, page 34).

There are other goals in capturing language aside from linguistic research, such as showing future generations what a language was like, or transmitting knowledge, or supporting ongoing community use. Within computational linguistics, goals range from modeling language structures, to extracting information, to providing speech or text interfaces. Each goal presents its own difficulties, and learning how to "transcribe accurately" may not be a priority in every case. Nevertheless, in most language situations, extended audio recordings are available, and we would like to be able to index their content in order to facilitate discovery and analysis. How can we best do this for oral languages?

The most common answer in the field of linguistics has been *transcription*: "The importance of the (edited) transcript resides in the fact that for most analytical procedures . . . it is the transcript (and not the original recording) which serves as the basis for further

Submission received: 16 July 2019; revised version received: 20 July 2020; accepted for publication: 13 September 2020.

https://doi.org/10.1162/COLI_a_00387

analyses" (Himmelmann 2006a, page 259). From this it follows that everything should be transcribed: "For the scientific documentation of a language it would suffice to *render all recordings* utterance by utterance in a phonetic transcription with a translation" (Mosel 2006, page 70, emphasis mine).

A parallel situation exists in the field of natural language processing (NLP). The "NLP pipeline" can be extended to cover spoken input by prefixing a speech-to-text stage. Given that it is easy to record large quantities of audio, and given that NLP tasks can be performed at scale, we have a problem known as the **transcription bottleneck**, illustrated in Figure 1.

Meanwhile, linguists have wondered for years whether methods from speech recognition could be applied to automatically transcribe speech in unwritten languages. In such cases there will not be a pronunciation lexicon or a language model, but it is becoming popular to automate at least the phone recognition stage on its own. For example, Michaud et al. report phone error rates in the 0.12–16 range, after training on 5 hours of transcribed audio from the Na language of China (Adams et al. 2018; Michaud et al. 2018). The idea is for humans to post-edit the output, in this case, correcting one out of every 6–8 characters, and then to insert word boundaries, drawing on their knowledge of the lexicon and of likely word sequences, to produce a word-level transcription. The belief is that by manually cleaning up an errorful phone transcription, and converting it into a word-level transcription, we will save time compared with entering a word-level transcription from scratch. To date, this position has not been substantiated.

Although such phone transcription methods are intended to support scalability, they actually introduce new problems for scaling: only linguists can provide phone transcriptions or graph-to-phone rules needed for training a phone recognizer, and only linguists can post-edit phone-level transcriptions.

To appreciate the severity of the problem, consider the fact that connected speech is replete with disfluencies and coarticulation. Thus, an English transcriber who hears *d'ya, d'ya see* might write *do you see* or /dou jʊ si/, to enable further analysis of the text. Instead, linguists are asked to transcribe at the phone level, i.e., [ʤəʤəsi]. We read this advice in the pages of *Language*: "field linguists [should modify] their [transcription] practice so as to assist the task of machine learning" (Seifart et al. 2018, page e335); and in the pages of *Language Documentation and Conservation*: "linguists should aim for exhaustive transcriptions that are faithful to the audio . . . mismatches result in high error rates down the line" (Michaud et al. 2018, page 12). Even assuming that linguists comply with these exhortations, they must still correct the output of the recognizer while re-listening to the source audio, and they must still identify words and produce a word-level transcription. It seems that the transcription bottleneck has been made more acute.

Three commitments lie at the heart of the transcription bottleneck: transcribing phones, transcribing fully, and transcribing first. None of these commitments is necessary, and all of them are problematic:

1. Transcribing phones. It is a retrograde step to build a phone recognition stage into the speech processing pipeline when the speech technology community has long moved away from the "beads-on-a-string" model. There is no physical basis for steady-state phone-sized units in the speech stream: "Optimizing for accuracy of low-level unit recognition is not the best choice for recognizing higher-level units when the low-level units are sequentially dependent" (Ostendorf 1999, page 79).

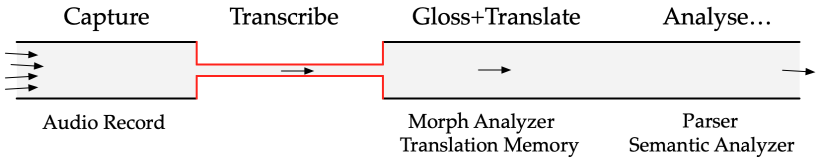


Figure 1
Transcription Bottleneck: the last frontier in computerizing oral languages?

2. Transcribing fully. The idea that search and analysis depend on written text has led to the injunction to transcribe fully: transcriptions have become the data. Yet no transcription is transparent. Transcribers are selective in what they observe (Ochs 1979). Transcriptions are subject to ongoing revision (Crowley 2007, pages 139f). "It would be rather naive to consider transcription exclusively, or even primarily, a process of mechanically converting a dynamic acoustic signal into a static graphic/visual one. Transcription involves interpretation..." (Himmelmann 2018, page 35). In short, *transcription is observation*: "a transcription, whatever the type, is always the result of an analysis or classification of speech material. Far from being the reality itself, transcription is an abstraction from it. In practice this point is often overlooked, with the result that transcriptions are taken to be the actual phonetic 'data' " (Cucchiarini 1993, page 3).

3. Transcribing first. In the context of language conservation, securing an audio recording is not enough by itself. Our next most urgent task is to capture the meaning while speakers are on hand. Laboriously re-representing oral text as written text has lower priority.

What would happen if we were to drop these three commitments and instead design computational methods that leverage the data and skills that are usually available for oral languages? This data goes beyond a small quantity of transcriptions. There will usually be a larger quantity of translations, because translations are easier to curate than transcriptions (cf. Figure 3). There will be a modest bilingual lexicon, because lexicons are created as part of establishing the distinct identity of the language. It will usually be straightforward to obtain audio for the entries in the lexicon. Besides the data, there are locally available skills, such as the ability of speakers to recognize words in context, repeat them in isolation, and say something about what they mean.

This leads us to consider a new model for large scale transcription that consists of identifying and cataloging words in an open-ended speech collection. Part of the corpus will be densely transcribed, akin to glossed text. The rest will be sparsely transcribed: words that are frequent in the densely transcribed portion may be detectable in the untranscribed portion. By confirming the system's guesses, it will get better at identifying tokens, and we leverage this to help us with the orthodox task of creating contiguous transcriptions.

I elaborate this "Sparse Transcription Model" and argue that it is a good fit to the task of transcribing oral languages, in terms of the available inputs, the desired outputs, and the available human capacity. This leads to new tasks and workflows that promise to accelerate the transcription of oral languages.

This article is organized as follows. I begin by examining how linguists have worked with oral languages over the past century (Section 2). Next, I review existing computational approaches to transcription, highlighting the diverse range of input and output data types and varying degrees of fit to the task (Section 3). I draw lessons from these

linguistic and computational contributions to suggest a new computational model for transcription, along with several new tasks and workflows (Section 4). I conclude with a summary of the contributions, highlighting benefits for flexibility, for scalability, and for working effectively alongside speakers of oral languages (Section 5).

2. Background: How Linguists Work with Oral Languages

The existing computational support for transcribing oral languages has grown from observations of the finished products of documentary and descriptive work. We see that the two most widely published textual formats, namely, phonetic transcriptions and interlinear glossed text, correspond to the two most common types of transcription tool (cf. Section 2.3). However, behind the formats is the process for creating them:

> No matter how careful I think I am being with my transcriptions, from the very first text to the very last, for every language that I have ever studied in the field, I have had to re-transcribe my earliest texts in the light of new analyses that have come to light by the time I got to my later texts. Not infrequently, new material that comes to light in these re-transcribed early texts then leads to new ways of thinking about some of the material in the later texts and those transcriptions then need to be modified. You can probably expect to be transcribing and re-transcribing your texts until you get to the final stages of your linguistic analysis and write-up (Crowley 2007, pages 139f).

Put another way, once we reach the point where our transcriptions do not need continual revision, we are sufficiently confident about our analysis to publish it. By this time the most challenging learning tasks—identifying the phonemes, morphemes, and lexemes—have been completed. From here on, transcription is relatively straightforward. The real problem, I believe, is the task discussed by Crowley, which we could call "learning to transcribe." To design computational methods for this task we must look past the well-curated products to the processes that created them (cf. Norman 2013).

For context, we begin by looking at the *why* (Section 2.1) before elaborating on the *how* (Section 2.2), including existing technological support (Section 2.3). We conclude with a set of requirements for the task of learning to transcribe (Section 2.4).

2.1 Why Linguists Transcribe

Linguists transcribe oral languages for a variety of reasons: to preserve records of linguistic events and facilitate access to them, and to support the learning of languages and the discovery of linguistic structures. We consider these in turn.

2.1.1 Preservation and Access. The original purpose of a transcription was to document a communicative event. Such "texts" have long been fundamental for documentation and description (Boas 1911; Himmelmann 2006b).

For most of history, writing has been the preferred means for inscribing speech. In the early decades of modern fieldwork, linguists would ask people to speak slowly so they could keep up, or take notes and reconstruct a text from memory. Today we can capture arbitrary quantities of spontaneous speech, and linguists are exhorted to record as much as possible: "Every chance should be seized immediately, for it may never be repeated…the investigator should not hesitate to record several versions of the same story" (Bouquiaux and Thomas 1992, page 58).

We translate recordings into a widely spoken language to secure their interpretability (Schultze-Berndt 2006, page 214). Obtaining more than one translation increases the likelihood of capturing deeper layers of meaning (Bouquiaux and Thomas 1992, page 57; Evans and Sasse 2007; Woodbury 2007).

Once we have recorded several linguistic events, there comes the question of access: how do we locate items of interest? The usual recommendation is to transcribe everything: "If you do a time-aligned transcription . . . you will be able to search across an entire corpus of annotated recordings and bring up relevant examples" (Jukes 2011, page 441). Several tools support this, representing transcriptions as strings anchored to spans of audio (e.g., Bird and Harrington 2001; Sloetjes, Stehouwer, and Drude 2013; Winkelmann and Raess 2014).

Because transcriptions are seen as the data, we must transcribe fully. Deviations are noteworthy: "we are also considering not transcribing everything..." (Woodbury 2003, page 11). The existence of untranscribed recordings is seen as dirty laundry: "Elephant in the room: Most language documentation and conservation initiatives that involve recording end up with a backlog of unannotated, 'raw' recordings" (Cox, Boulianne, and Alam 2019). There is broad consensus: preservation and access are best served by transcribing fully. Nevertheless, we will establish a way to drop this requirement.

2.1.2 Learning and Discovery. The task of learning to transcribe can be difficult when the sounds are foreign and we don't know many words. The situation is nothing like transcribing one's own language. At first we hear a stream of sounds, but after a while we notice recurring forms. When they are meaningful, speakers can readily reproduce them in isolation, and offer a translation, or point at something, or demonstrate an action. As Himmelmann also notes, "transcription necessarily involves hypotheses as to the meaning of the segment being transcribed" (Himmelmann 2018, page 35). We recognize more and more words over time (Newman and Ratliff 2001; Rice 2001). As a consequence, "transcription also involves language learning" (Himmelmann 2018, page 35), and transcribers inevitably "learn the language by careful, repeated listening" (Meakins, Green, and Turpin 2018, page 83).

For readers who are not familiar with the process of transcribing an oral language, I present an artificial example from the TIMIT Corpus (Garofolo et al. 1986). Consider the sentence *she had your dark suit in greasy wash water all year*. At first, we recognize nothing other than a stream of phones (1a).

(1) a. ʃiɦɑɡɨdɑrksüdn̩grisiwɑʃwɑrəʔɔlyɪə (a stream of sounds)
b. ʃi ɦɑɡɨ dɑrk südn̩grisiwɑʃ wɑrə ʔɔlyɪə (some words recognized)
c. ʃi ɦæd jɔr dɑrk sut ɪn grisi wɑʃ wɑtər ɔl yɪə (all words recognized)

If we were to transcribe the same audio a few days later, once we recognized common words like *she*, *dark*, and *water*, we might write (1b), which includes a residue where we can still only write phonetically. Once we recognize more words, we transition to writing canonical lexical representations (1c). The sequence in (1b) approximates the state of our knowledge while we are learning to transcribe.

The story is a little more complex, as there are places where we equivocate between writing or omitting a phone, such as in rapid utterances of *half a cup*, [hafəkʌp] ∼ [hafkʌp]. It requires some knowledge of a language to be able to go looking for something that is heavily reduced. Additionally, there are places where we might entertain more than one hypothesis. After all, language learners routinely mis-parse speech (Cairns et al. 1997), and even fluent adult speakers briefly recognize words that

do not form part of the final hypothesis (e.g., recognizing *bone* en route to *trombone*, Shillcock 1990). This phenomenon of multiple hypotheses occurs when transcribing field recordings (Hermes and Engman 2017), and it is recommended practice to keep track of them:

> Don't be afraid of writing multiple alternative transcriptions for a word you hear, especially when the language is new to you... Similarly, it is wise not to erase a transcription in your notebook; simply put a line through it and write the correction next to it. Make note of any transcriptions you are unsure about. Conversely, keep a record of all those you are confident about (Meakins, Green, and Turpin 2018, page 99).

In summary, learning to transcribe is a discovery process with an indeterminate endpoint: we come across new words in each new spoken text; we encounter variability in pronunciation; we run into allophony and allomorphy; we are tripped up by disfluency and coarticulation; and we guess at the contents of indistinct passages.

2.2 How Linguists Transcribe

Over an extended period, linguists have primarily transcribed at the word level. Often, translation was given higher priority than transcription. There have been many efforts to delegate transcription to speakers. We consider these points in turn.

2.2.1 Transcribing Words. Since the start of the modern period of linguistic description, linguists have transcribed at the word level. Working in the Arctic in the 1880s, Franz Boas "listened to stories and wrote down words" (Sanjek 1990, page 195). His early exposure to narratives only produced a wordlist: "my glossary is really growing." Once he gained facility in the language, Boas would transcribe from memory, or back-translate from his English notes (Sanjek 1990, pages 198f). This is hardly a way to capture idiosyncratic pronunciations. Thus, the texts that have come down to us from this period are not transparent records of speech (Clifford 1990, page 63).

Half a century later, a similar practice was codified in phonemic theory, most notably in Kenneth Pike's *Phonemics: A Technique for Reducing Language to Writing* (Pike 1947). The steps were to (a) collect words, (b) identify "minimal pairs," (c) establish the phonemic inventory, and (d) collect texts using this phonemic orthography. Ken Hale describes a process that began with a version of the phonemic approach and continued with Boas' practice of building up more vocabulary in sentential contexts:

> In starting work on Ulwa, I decided to follow the procedure I have used elsewhere – North America, Mexico, Australia – in working on a "new" language. The first session, for example, would involve eliciting basic vocabulary – I usually start with body part terms – with a view, at this early point, of getting used to the sounds of the language and developing a way of writing it. And I would proceed in this manner through the basic vocabulary (of some 500 items) ... until I felt enough at ease with the Ulwa sound system to begin *getting the vocabulary items* in sentences rather than in isolation (Hale 2001, page 85, emphasis mine).

When approaching the task of transcription, both linguists and speakers come to the table with naive notions of the concept of *word*. Any recognized "word" is a candidate for later re-interpretation as a morpheme or multiword expression. Indeed, all levels of segmentation are suspect: texts into sentences, sentences into words, and words into morphemes. When it comes to boundaries, our typography—with its periods, spaces,

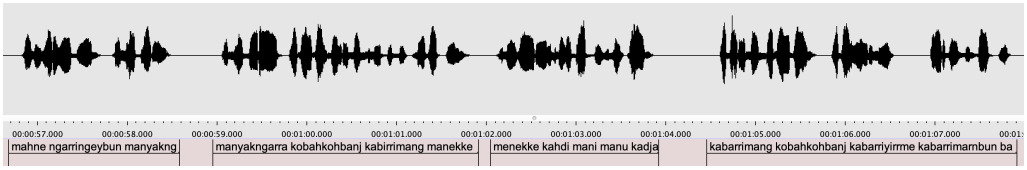


Figure 2
Transcription from the author's fieldwork with Kunwinjku [gup], aligned at the level of breath groups (Section 2.3).

and hyphens—makes it easy to create textual artifacts which are likely to be more precise as to boundaries than they are accurate.

A further issue arises from the "articulatory overlap" of words (Browman and Goldstein 1989), shown in (2). The desire to segment phone sequences into words encourages us to *modify* the phone transcription:

(2) a. tɛm̲pɪn ∼ tɛn pɪn 'ten pin' (assimilation)
b. ɦæd̻̲ʲi ∼ ɦæd jɨ 'had your' (palatalization)
c. tɛnt̲sɛnts ∼ tɛn sɛnts 'ten cents' (intrusive stop)
d. lɔr̲ænd ∼ lɔ ænd 'law and (order)' (epenthesis)

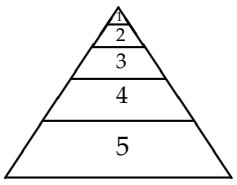
As a result, there is more to word segmentation than inserting spaces into the phone sequence. On the descriptive side, this means that if we ever reassign a word boundary by moving the whitespace, our phone transcription may preserve a trace of our earlier guesswork. On the computational side, this means that segmenting words by inserting boundaries into a phone sequence is unsound, which carries implications for the "word segmentation task" (Section 3).

There are just two coherent, homogenous representations: phonetic sequences (1a), and canonical lexical sequences (1c). We can insist on the former, but writing out the idiosyncratic detail of individual tokens is arduous, and subject to unknown inter-annotator agreement (Himmelmann 2018, page 36), and in view of this fact, no longer accepted as the basis for research in phonetics (Valenta et al. 2014; Maddieson 2001, page 213). This leaves the latter, and we observe that writing canonical lexical sequences does not lose idiosyncratic phonetic detail when there is a time-aligned speech signal. It is common practice to align at the level of easily-identifiable "breath groups" (Voegelin and Voegelin 1959, page 25), illustrated in Figure 2.

2.2.2 Prioritizing Translation Over Transcription. The language preservation agenda involves capturing linguistic events. Once we have a recording, how are we to prioritize transcription and translation? Recent practice in documentary linguistics has been to transcribe then translate, as Chelliah explains:

> Once a recording is made, it must be transcribed and translated to be maximally useful, but as is well-known, *transcription is a significant bottleneck to translation.* . . . For data gathering and analysis, the following workflow is typical: recording language interactions or performances > collecting metadata > *transcribing* > *translating* > annotating > archiving > disseminating (Chelliah 2018, pages 149, 160, emphasis mine).

In fact, transcription is only a bottleneck for translation if we assume that translation involves written sources. Mid-century linguists made no such assumption, believing



(1) **Core Corpus**: a central body of data, the basis of a complete linguistic analysis;
(2) **Indexed Corpus**: equipped with a complete lexicon, an indexed list of morphemes with glosses and morphological classifications;
(3) **Transcribed Corpus**: transcriptions prepared as soon as possible after the recording to reduce the frustrations of cold notes;
(4) **Translated Corpus**: not transcribed but translated into some familiar language, with indications of the social contexts;
(5) **Raw Corpus**: unprocessed recordings.

Figure 3
The tapered corpus. The quantity of data at each level follows a power law based on the amount of curation required (after Samarin 1967, page 70; Twaddell 1954, page 108). A similar situation has been observed in NLP (Abney and Bird 2010).

that far more material would be translated than we could ever hope to transcribe (Figure 3). The same was true fifty years earlier, when Boas prioritized translation over transcription (Sanjek 1990, pages 198f). Fifty years later it was still considered best practice to prioritize translation over transcription:

> At least a rough word-for-word translation must be done immediately along with a free translation. . . Putting off the transcription may spare the narrator's patience. . . (Bouquiaux and Thomas 1992).

It remains likely for endangered languages that more materials will be translated than transcribed. Translation is usually quicker, easier, and more important for the documentary record (Section 2.1.1). Moreover, translation sits better with speakers who assume that linguists would be more interested in the content of their stories, than in how they would be represented in a written code (cf. Maddieson 2001, page 215; Bird 2020).

2.2.3 Working with Speakers. Language documentation involves substantial collaboration with local people (Austin 2007; Rice 2011). Often, these people are bilingual in a language of wider communication. Local bilinguals could be retired teachers or government employees. They could be students who meet the visiting linguist in the provincial capital and escort her to the village. They might have moved from the ancestral homeland to a major city where they participate in urban fieldwork (Kaufman and Perlin 2018). Whatever the situation, these people may be able to provide transcriptions, perhaps by adapting the orthography of another language (Figure 4). Sometimes, one finds speakers who are proficient in the official orthography of the language, even though that orthography might not be in widespread use.

Literate speakers have some advantages over linguists when it comes to transcription. They have a comprehensive lexicon and language model. They can hold conversations in the language with other speakers to clarify nuances of meaning. Their professional work may have equipped them with editorial and keyboarding skills.

In many places, employing locals is inexpensive and delivers local economic benefits. There are many initiatives to train these people up into "community linguists" (Dobrin 2008; Rice 2009; Bird and Chiang 2012; Yamada 2014; Sapién 2018). This is suggested as a solution to the transcription bottleneck:

> Ideally we should be getting transcriptions of all the recordings, but that is not always feasible or affordable... Most funders will pay for transcription work in the heritage

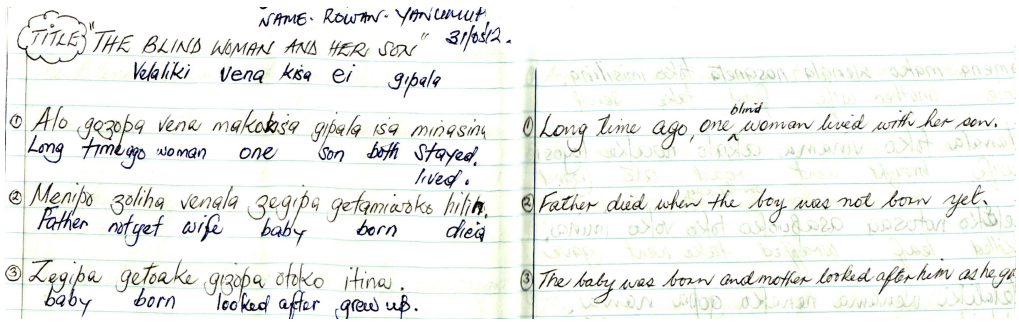


Figure 4
Transcription and glossing performed by a speaker (Bird and Chiang 2012).

> language, so add that line item to your budget to get more recordings transcribed by someone else (King 2015, page 10).

When speakers are not literate—in the narrow western sense—they can still identify words in connected speech, supporting linguists as they transcribe (Gudschinsky 1967, page 9; Nathan and Fang 2009, page 109; Meakins, Green, and Turpin 2018, page 230).

Another way to involve speakers is through oral transcription or **respeaking**: "starting with hard-to-hear tapes and asking elders to 'respeak' them to a second tape slowly so that anyone with training in hearing the language can make the transcription" (Woodbury 2003, page 11). Respeaking is tantamount to dictation to future transcribers (Abney and Bird 2010; Sperber et al. 2013). Respeaking reduces the number of transcription mistakes made by non-speakers (Bettinson 2013). Of note for present purposes, respeaking reproduces *words*. Speakers do not reproduce disfluencies or mimic dialects. Thus, the respeaking task supports transcription at the word level.

This concludes our discussion of how linguists typically work with oral languages, showing that the most prevalent transcriptional practice takes place at the word level.

2.3 Technological Support for Working with Oral Languages

For many years, the Linguist's Shoebox (later, Toolbox) was the mainstay of language data collection, replacing the traditional shoebox of file cards with hand-written lexical entries and cultural notes (Buseman, Buseman, and Early 1996). It stored language data in a text file as a sequence of blankline-delimited records, each containing a sequence of newline-delimited fields, each consisting of a field name such as `\lx` followed by whitespace followed by text content. This format, known as SIL Standard Format, is supported in the Natural Language Toolkit (Robinson, Aumann, and Bird 2007), to facilitate processing of texts and lexicons coming from linguistic fieldwork (Bird, Klein, and Loper 2009, §11.5).

SIL Standard Format was an early version of semistructured data, for which XML was later devised (Abiteboul, Buneman, and Suciu 2000). Fieldworks Language Explorer (FLEx, Butler and Volkinburg 2007; Moe 2008) switched to XML in order to benefit from schemas, validation, stylesheets, and so forth. FLEx, like Shoebox, is "especially useful for helping researchers build a dictionary as they use it to analyze and interlinearize text" (SIL Language Technology 2000). This functionality—updating a lexicon while glossing text—parallels the approach to transcription described in this article.

To accelerate the collection of lexical data, the Rapid Words Method was devised, organizing teams of speakers and collecting upwards of 10,000 words over a ten-day period (Rapidwords 2019). The Rapid Words Method is implemented in WeSay (Albright and Hatton 2008), which permits lexical data to be interchanged with FLEx.

The curation of text data in FLEx is a more involved process. In FLEx, the "baseline" of a glossed text is imported from a text file, i.e., a transcription. The user works through the words adding glosses, cf. (3).

(3)

ngalwanjdjuk	korroko	ngalkohbanj	ba	di	ba	yimi.	yoh,	ba	rei	mandjai,	ba	buni
emu	long.ago	old.woman	she	was	she	said	yes	she	went	?	she	was.hitting

With each new word, the transcriber adds a lexical entry. Each time that same word is encountered, the gloss is filled in from the lexicon (Rogers 2010). If an existing word occurs with a new sense, the user adds this to the lexical entry. This interface encourages consistent spelling, because the gloss is only auto-filled if the transcriber adopts a canonical spelling for each word. Internally, FLEx represents texts as a sequence of pointers, cf. (4).

(4) 42 46 39 47 48 47 37 90 47 . . .

:		
45	korroko	just.now
46	korroko	long.ago
47	ba	he, she, it;
48	di	be
49	di	stand
:		

If a lexical entry is updated, for example, with a revised spelling or gloss, this update appears everywhere. The display in (3) is merely a presentation format for the data structure in (4). Therefore, when we see displays like (3) we have the choice of viewing them as a sequence of pairs of strings, or a sequence of *sense-disambiguated lexical identifiers*. FLEx builds in the latter viewpoint.

FLEx supports morphological segmentation and glossing, as shown in (5). The baseline includes allophonic detail. The second row shows a phonemic transcription with morphological segmentation. The third row contains morphological glosses with information about tense, aspect, person, number, and noun class. The last row shows the phrasal translation.

(5)

alwaɲɟuk	gɔrɔgɔ	algɔʔbaɲ	bari	bajimi.	wou,	baɪei	anɟai,	babuni
ŋal-waɲɟuk	gɔdɔgɔ	ŋal-gɔʔbaɲ	ba-di	ba-jim-i.	woh,	ba-ɪe-i	man-ɟai,	ba-bu-ni
II-emu	before	II-old.person	3P-beP	3P-say-PI	yes	3P-go-PI	III-cane.grass	3/3P-hit-PI

Long ago, Emu was an old woman. Yes, she would go off to get cane grass.

FLEx builds up a morphological analysis during transcription and uses this to analyze successive words: "the user gradually tells the system what s/he knows about the grammar, receiving as a reward increasingly automated analysis of text" (Black and Simons 2008, page 44). Thus, all the information in (5) is projected from the lexicon. The display of rows can be turned on and off, supporting the practice of showing different amounts of detail depending on the audience (Bowern 2008, page 60).

There is a qualitative distinction between (3) and (5) aside from the level of detail. The first format is a documentary artifact. It is a work in progress, a sense-disambiguated word-level transcription where the segmentation and the glosses are volatile. I will call this IGT_1. The second format is an analytical product. It is a completed work, illustrating a morphophonemic and morphosyntactic analysis (cf. Evans 2003, page 663). I will call this IGT_2.

In this view, IGT_1 is a kind of word-level transcription where we take care not to collapse homographs. Here, words are not ambiguous grapheme strings, but unique lexical identifiers. Where words are morphologically complex, as we see in (5), we simply expect each morph-gloss pair to uniquely identify a lexical or grammatical morpheme. Getting from IGT_1 to IGT_2 involves analytical work, together with refactoring the lexicon, merging and splitting entries as we discover the morphology. No software support has been devised for this task.

To accelerate the collection of text data, the most obvious technology is the audio recording device, coupled with speech recognition technology. This promises to deliver us the "radically expanded text collection" required for language documentation (Himmelmann 1998). However, these technological panaceas bring us back to the idea of the processing pipeline (Figure 1) with its attendant assumptions of transcribing first and transcribing fully, and to the transcription bottleneck. Tedlock offers a critique:

> Those who deal with the spoken word . . . seem to regard phonography as little more than a device for moving the scene of alphabetic notation from the field interview to the solitude of an office... The real analysis begins only after a document of altogether pre-phonographic characteristics has been produced... The alphabet continues to be seen as an utterly neutral, passive, and contentless vehicle (Tedlock 1983, page 195).

Here, the possibility of time-aligned annotation offers a solution, cf. Figure 2 (Bird and Harrington 2001; Jacobson, Michailovsky, and Lowe 2001; Sloetjes, Stehouwer, and Drude 2013; Winkelmann and Raess 2014). We do not need to view transcriptions as the data, but just as annotations of the data:

> The importance of the (edited) transcript resides in the fact that for most analytical procedures . . . it is the transcript (and not the original recording) which serves as the basis for further analyses. Obviously, whatever mistakes or inconsistencies have been included in the transcript will be carried on to these other levels of analysis... This problem may become somewhat less important in the near future inasmuch as it will become standard practice to link transcripts line by line (or some other unit) to the recordings, which allows direct and fast access to the original recording whenever use is made of a given segment in the transcript (Himmelmann 2006a, page 259).

Unfortunately, the available software tools fall short, representing the time-aligned transcription of a phrase as a character string. None of the analysis and lexical linking built into FLEx and its predecessors is available. With some effort it is possible to export speech annotations to FLEx, and then to import this as a text for lexical linking and interlinear glossing (Gaved and Salffner 2014). However, this brings us back to the pipeline, and to text standing in for speech as primary data.

As a result, there is a fundamental shortcoming in the technological support for working with oral languages. We have tools for linking unanalyzed texts to speech, and tools for linking analyzed texts to the lexicon. However, to date, there is no transcription tool that supports simultaneous linking to audio and to speech and to the lexicon. To

the extent that both speech and lexical knowledge inform transcription, linguists are on their own.

A recent development in the technological support for processing oral languages comes under the heading of Basic Oral Language Documentation (BOLD) (Bird 2010; Reiman 2010). Here, we stay in the oral domain, and speakers produce respeakings and spoken interpretations into a language of wider communication, all aligned at the granularity of breath groups to the source audio (cf. Figure 2). Tools that support the BOLD method include SayMore and Aikuma (Hatton 2013; Hanke and Bird 2013; Bird et al. 2014). SayMore incorporates automatic detection of breath groups, while Aikuma leaves this in the hands of users.

2.4 Requirements for Learning to Transcribe

In this section I have explored why and how linguists learn to transcribe, and the existing technological support for transcription, leading to six central requirements. First, documentary and descriptive practice has focused on *transcribing words*, not phones. We detect repeated forms in continuous speech, and construct an inventory (Section 2.2.1). Second, in the early stages we have no choice but to transcribe naively, not knowing whether a form is a morpheme, a word, or a multiword expression, only knowing that we are *using meaningful units* (Section 2.1.2). Third, transcription can proceed by *leveraging translations*. The translation task is more urgent and speakers find it easier. Thus, we can assume access to meanings, not just forms, during transcription (Sections 2.1.1, 2.2.2). Fourth, our representations and practices need to enable *transcribing partially*. There will always be regions of signal that are difficult to interpret given the low quality of a field recording or the evolving state of our knowledge (Section 2.1.2). Fifth, we are engaging with a speech community, and so we need to find effective ways of *working with speakers* in the course of our transcription work (Section 2.2.3). Finally, given our concern with consistency across transcribers, texts, and the lexicon, we want our tools to support *simultaneous linking* of transcriptions to both the source audio and to the lexicon (Section 2.3).

These observations shape our design of the Sparse Transcription Model (Section 4). Before presenting the model, we review existing computational approaches to transcription that go beyond the methods inspired by automatic speech recognition, and consider to what extent they already address the requirements coming from the practices of linguists.

3. Computation: Beyond Phone Recognition

If language documentation proceeds from a "radically expanded text collection," how can speech and language technology support this? Approaches inspired by automatic speech recognition draw linguists into laborious phone transcription work. Automatic segmentation delivers pseudowords, not actual words, and depends on the false assumption that words in connected speech do not overlap (Section 2.2.1). As they stand, these are not solutions to the word-level transcription task as practiced in documentary and descriptive linguistics, and as required for natural language processing.

In this section we consider approaches that go beyond phone transcription. Several approaches leverage the translations that are readily available for endangered languages (Section 2.2.2). They may segment phone transcriptions into pseudowords

then align these with translations (Section 3.1). Segmentation and alignment may be performed jointly (Section 3.2). Segmentation and alignment may operate directly on the speech, bypassing transcription altogether (Section 3.3). An entirely different approach to the problem is based on spoken term detection (Section 3.4). Next, we consider the main approaches to evaluation and their suitability for the transcription task (Section 3.5). The section concludes with a summary of approaches to computation and some high-level remarks (Section 3.6), and a discussion of how well these methods address the requirements for learning to transcribe (Section 3.7).

3.1 Segmenting and Aligning Phone Sequences

This approach involves segmenting phone sequences into pseudowords then aligning pseudowords with a translation. We construe the aligned words as glosses. At this point, we have produced interlinear glossed text of the kind we saw in (3). Now we can use alignments to infer structure in the source utterances (cf. Xia and Lewis 2007).

Phone sequences can be segmented into word-sized units using methods developed for segmentation in non-roman orthographies and in first language acquisition (Cartwright and Brent 1994; Goldwater, Griffiths, and Johnson 2006, 2009; Johnson and Goldwater 2009; Elsner et al. 2013). Besacier et al. took a corpus of Iraqi Arabic text with sentence-aligned English translations, converted the Arabic text to phone transcriptions by dictionary-lookup, then performed unsupervised segmentation of the transcriptions (Besacier, Zhou, and Gao 2006). In a second experiment, Besacier et al. replaced the canonical phone transcriptions with the output of a phone recognizer trained on 200 hours of audio. This resulted in a reasonable phone error rate of 0.15, and slightly lower translation scores. One of their evaluations is of particular interest: a human was asked to judge which of the automatically identified pseudowords were actual words, and then the system worked with this hybrid input of identified words with intervening phone sequences representing unidentified words, the same scenario discussed in Section 2.1.2.

Although the translation scores were mediocre and the size of training data for the phone recognizer was not representative for most oral languages, the experimental setup is instructive: it leverages prior knowledge of the phone inventory and a partial lexicon. We can be confident of having such resources for any oral language. Zanon Boito et al. used a similar idea, expanding a bilingual corpus with additional input pairs to teach the learner the most frequent 100 words, "representing the information a linguist could acquire after a few days" (Zanon Boito et al. 2017).

Unsupervised segmentation methods often detect sequences of phones that are unlikely to appear within morphemes. Thus, in English, the presence of a non-homorganic sequence [np] is evidence of a boundary. However, the ability of a system to segment [tɛnpɪn] as [tɛn pɪn] tells us little about its performance on more plausible input where coarticulation has removed a key piece of evidence: [tɛm̲pɪn] (2a). Unsupervised methods also assume no access to speakers and their comprehension of the input and their ability to respeak it or to supply a translation (Bird 2020). Thus, learning to transcribe is not so much language acquisition as cross-linguistic bootstrapping, on account of the available data and skills (Figure 3; Abney and Bird 2010).

Another type of information readily available for endangered languages is translation, as just mentioned. Several computational approaches have leveraged translations to support segmentation of the phone sequence. This implements the translation-before-transcription workflow that was discussed in Section 2.2.2.

3.2 Leveraging Translations for Segmentation

More recent work has drawn on translations to support segmentation. Neubig et al. (2012) were the first to explore this idea, leveraging translations to group consecutive phones into pseudowords. They drew on an older proposal for translation-driven segmentation in phrase-based machine translation in which contiguous words of a source sentence are grouped into phrases which become the units of alignment (Wu 1997).

Stahlberg et al. began with IBM Model 3, which incorporates a translation model for mapping words to their translations, and a distortion model for placing translated words in the desired position (Brown et al. 1993; Stahlberg et al. 2016). In Model 3P, distortion is performed first, to decide the position of the translated word. A target word length is chosen and then filled in with the required number of phones. Stahlberg et al. trained their model using 600k words of English–Spanish data, transliterating the source language text to canonical phone sequences and adding 25% noise, approximating the situation of unwritten languages. They reported segmentation F-scores in the 0.6–0.7 range. Stahlberg et al. do not report whether these scores represent an improvement over the scores that would have been obtained with an unsupervised approach.

Godard et al. applied Stahlberg's method using a more realistic data size, i.e., a corpus of 1.2k utterances from Mboshi with gold transcriptions and sentence-aligned French translations (Godard et al. 2016). They segmented transcriptions with the support of translations, reporting that it performed less well than unsupervised segmentation. Possible factors are the small size of the data, the reliance on spontaneous oral interpretations, and working across language families.

Godard et al. extended their work using a corpus of 5k Mboshi utterances, using phone recognizer output instead of gold transcriptions (Godard et al. 2018b). Although promising, their results demonstrate the difficulty of segmentation in the presence of noise, and sensitivity to the method chosen for acoustic unit discovery.

Adams et al. performed joint word segmentation and alignment between phone sequences and translations using pialign, for German text transliterated into canonical phone sequences (Adams et al. 2015). They extracted a bilingual lexicon and showed that it was possible to generate hundreds of bilingual lexical entries on the basis of just 10k translated sentences. They extended this approach to speech input, training a phone recognizer on Japanese orthographic transcriptions transliterated to gold phoneme transcriptions, and segmented and aligned the phone lattice output (Adams et al. 2016b). They evaluated this in terms of improvement in phone error rate.

This shift to speech input opens the way to a more radical possibility: bypassing transcription altogether.

3.3 Bypassing Transcription

If the downstream application does not require it, why would we limit ourselves to an impoverished alphabetic representation of speech when a higher-dimensional, or non-linear, or probabilistic representation would capture the input more faithfully? Segmentation and alignment tasks can be performed on richer representations of the speech input, for example, building language models over automatically segmented speech (Neubig et al. 2010), aligning word lattices to translations (Adams et al. 2016a), training a speech recognizer on probabilistic transcriptions (Hasegawa-Johnson et al. 2016), or translating directly from speech (Duong et al. 2016; Bansal et al. 2017; Weiss et al. 2017; Chung et al. 2019).

Anastasopoulos, Chiang, and Duong (2016) inferred alignments directly between source audio and text translations first for Spanish–English, then for 330 sentences of Griko speech with orthographic translations into Italian. The task is to "identify recurring segments of audio and cluster them while aligning them to words in a text translation." Anastasopoulos et al. (2017) took this further, discovering additional tokens of the pseudowords in untranslated audio, in experiments involving 1–2 hours of speech in Ainu and Arapaho.

There is another way to free up our idea of the required output of transcription, namely, to relax the requirement that it be contiguous. We turn to this next.

3.4 Spoken Term Detection

Spoken term detection, also known as keyword spotting, is a long-standing task in speech recognition (Myers, Rabiner, and Rosenberg 1980; Rohlicek 1995; Fiscus et al. 2007). The primary impetus to the development of this method was to detect a term in the presence of irrelevant speech, perhaps to perform a command in real time, or to retrieve a spoken document from a collection. Spoken term detection is much like early transcription in which the non-speaker transcriber is learning to recognize words. Some methods involve modeling the "filler" between keywords, cf. (1b) (Rohlicek 1995).

Spoken term detection methods have been applied to low-resource languages in the context of the IARPA Babel Program (e.g., Rath et al. 2014; Gales et al. 2014; Liu et al. 2014; Chen et al. 2015; Metze et al. 2015), but with far more data than we can expect to have for many oral languages, including 10+ hours of phone transcriptions and more-or-less complete lexicons.

Spoken term detection has been applied to detect recurring forms in untranscribed speech, forms that do not necessarily carry any meaning (Park and Glass 2008; Jansen, Church, and Hermansky 2010; Dunbar et al. 2017). However, this unsupervised approach is less well aligned to the goals of transcription where we expect terms to be meaningful units (Section 2.1.2; Fiscus et al. 2007, page 52).

3.5 Test Sets and Evaluation Measures

In evaluating the above methods, a popular approach is to take an existing large corpus with gold transcriptions (or with transcriptions that are simulated using grapheme-to-phoneme transliteration), along with translations. We simulate the low-resource scenario by giving the system access to a subset of the data with the possible addition of noise (e.g., Besacier, Zhou, and Gao 2006; Stahlberg et al. 2016). Some have compiled small corpora in the course of their work, for example, Griko (Boito et al. 2018) or Mboshi (Godard et al. 2018a; Rialland et al. 2018). Some collaborations have tapped a long-standing collection activity by one of the partners, for example, Yongning Na (Adams et al. 2017). Others have found small collections of transcribed and translated audio on the Web, for example, Arapaho and Ainu (Anastasopoulos et al. 2017).

Phone error rate is the accepted measure for phone transcriptions, and it is easy to imagine how this could be refined for phonetic or phonemic similarity. However, optimizing phone error rate comes at a high cost for linguists, because it requires that they "aim for exhaustive transcriptions that are faithful to the audio" (Michaud et al. 2018, page 12). Reducing phone error rate is not necessarily the most effective way to improve word-level recognition accuracy.

Word error rate is a popular measure, but it suffers from a fundamental problem of representation. Words are identified using character strings, and a transcription word is

considered correct if it is string-identical to a reference word. Yet these string identifiers are subject to ongoing revision. Our evolving orthography may collapse phonemic distinctions, creating homographs (cf. English [wɪnd] vs [waɪnd] both written *wind*). Our incomplete lexicon may omit homophones (cf. an English lexicon containing bat_1 *flying mammal* but not bat_2 *striking instrument*). Our orthographic practice might not enable us to distinguish homophones (cf. English [sʌn] written *son* vs *sun*). This reliance on inadequate word identifiers and an incomplete lexicon will cause us to *underestimate* word error rate. Conversely, a lack of understanding about dialect variation, or lack of an agreed standard spelling, or noise in word segmentation, all serve to multiply the spellings for a word. This variability will cause us to *overestimate* word error rate. These problems carry forward to type-based measures of lexicon quality such as precision at *k* entries (Adams et al. 2015).

Translation quality is another popular measure, and many of the above-cited papers report BLEU scores (Papineni et al. 2002). However, the data requirements for measuring translation quality amplify our sparse data problems. Data sparseness becomes more acute when we encounter mismatches between which concepts are lexicalized across source and target languages:

> Parallel texts only address standardized, universal stories, and fail to explore what is culture-specific, either in terms of stories or in terms of lexical items. Parallel bible or other corpora may tell us how to say 'arise!' or 'Cain fought with Abel.' But we will not encounter the whole subworld of lexical particularities that make a language unique, such as Dalabon *dalabborrord* 'place on a tree where the branches rub together, taken advantage of in sorcery by placing something that has been in contact with the victim, such as clothes, in such a way that it will be rubbed as the tree blows in the wind, gradually sickening and weakening the victim.' The thousands of fascinating words of this type are simply bracketed out from traditions of parallel translation (Evans and Sasse 2007, page 71).

Each "untranslatable word" leads to parallel texts that include lengthy interpretations, with a level of detail that varies according to the translator's awareness of the lexical resources of the target language, and beliefs about the cultural awareness of the person who elicits the translation. The problem runs deeper than individual lexical items. For example, Woodbury describes a class of suffixes that express affect in Cup'ik discourse, which translators struggle to render into English (Woodbury 1998, page 244). We expect great diversity in how these suffixes are translated. For these reasons, it seems far-fetched to evaluate translations with BLEU, counting *n*-gram overlap with reference translations. Perhaps the main benefit of translation models lies not in translation per se, but in their contribution to phone or word recognition and segmentation.

3.6 Summary

For linguists, transcription is a unitary task: listen to speech and write down words. Computational approaches, by contrast, show great diversity. Indeed, it is rare for two approaches to use the same input and output settings, as we see in Figure 5.

One source of diversity is the input: a large corpus of transcriptions; a small corpus of transcriptions augmented with material from other languages; linguist- or speaker- or non-speaker-supplied phonemic, phonetic, or orthographic transcriptions; transcriptions with or without word breaks; transcriptions derived from graph-to-phone rules applied to lexemes; partial transcriptions; or probabilistic transcriptions. Outputs are equally diverse: phones with or without word segmentation; a lexicon or not; associated

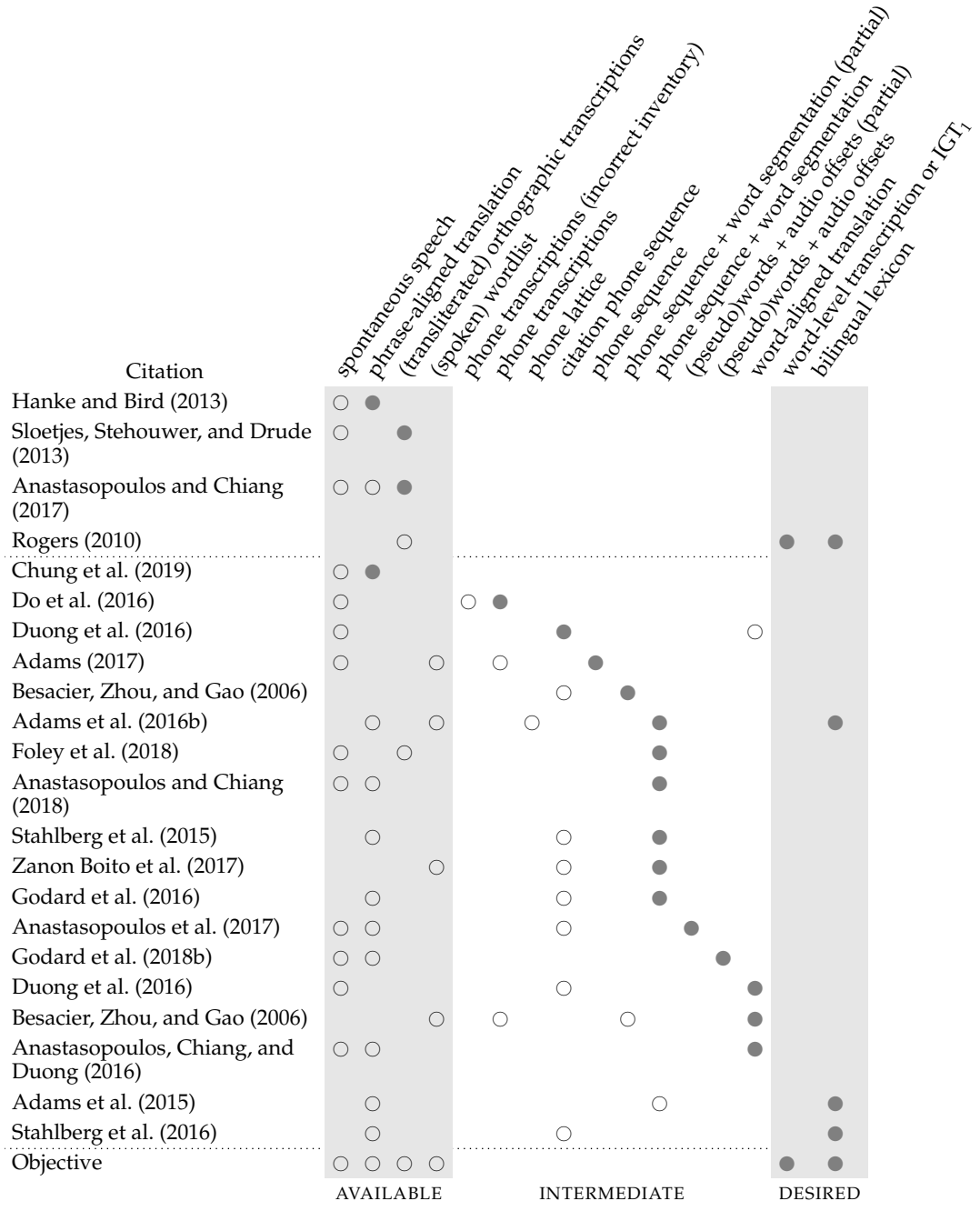


Figure 5
Summary of transcription methods: Inputs and outputs are indicated using ○ and ● respectively. Sometimes, an output type also serves as an input type for training. Shaded regions represent data whose existence is independently motivated. When a paper includes multiple experiments, we only document the most relevant one(s). I do not include models as inputs or outputs; if data X is used to train a model, and then the model is used to produce data Y, we identify a task mapping from X to Y. The first section of the table lists tools for use by humans; the second section lists automatic tools; the final section shows our objective, relying on independently motivated data.

meanings; meaning represented using glosses, alignments, or a bilingual lexicon. The computational methods are no less diverse, drawing on speech recognition, machine translation, and spoken term detection.

From the standpoint of our discussion in Section 2, the existing computational approaches incorporate some unwarranted and problematic assumptions. First, all approaches set up the problem in terms of inputs and outputs. However, linguists regularly access speakers during their transcription work. These people not only have a greater command of the language, they may be familiar with the narrative being transcribed, or may have been present during the linguistic performance that was recorded. Given how commonly a linguist transcriber has a speaker "in the loop," it would be worthwhile to investigate human-in-the-loop approaches. For instance, facilitating access to spoken and orthographic translations can support transcription by non-speakers (Anastasopoulos and Chiang 2017).

Second, the approaches inspired by speech recognition and machine translation proceed by segmenting a phone sequence into words. Evaluating these segmentations requires a gold standard that is not available in the early stages of language documentation. At this point we only have partial knowledge of the morphosyntax; our heightened awareness of form (the speech signal) and reduced awareness of meaning causes us to favor observable *phonological* words over the desired *lexical* words; we cannot be confident about the treatment of clitics and compounds as independent or fused; transcribers may use orthographic conventions influenced by another language and many other factors (Himmelmann 2006a; Cahill and Rice 2014). Add to all this the existence of coarticulation in connected speech: consecutive words may not be strictly contiguous thanks to overlaps and gaps in the phone sequence, with the result that segmentation is an unsound practice (cf. Section 2.2.1).

3.7 Addressing the Requirements

How well does this computational work address our six requirements (Section 2.4)?

1. Transcribing words. Much computational work has approached phone recognition in the absence of a lexicon. If we cannot be sure of having a lexicon, the interests of general-purpose methods seem to favor approaches that do not require a lexicon. However, the process of identifying a language and differentiating it from its neighbors involves eliciting a lexicon. Establishing a phoneme inventory involves a lexicon. Early transcription involves a lexicon. In short, *we always have a lexicon*. To the extent that accelerating the transcription of words means accelerating the construction of a lexicon, then it makes sense to devote some effort early on to lexical elicitation. There are creative ways to simulate partial knowledge of the lexicon, for example, by substituting system output with correct output for the n most frequent words (Besacier, Zhou, and Gao 2006), or by treating the initial bilingual lexicon as a set of mini parallel texts and adding them to the parallel text collection.

2. Using meaningful units. Much computational work has discovered "words" as a byproduct of detecting word boundaries or repeated subsequences, sometimes inspired by simulations of child language acquisition. The lexicon is the resulting accidental inventory of forms. However, transcribers access meanings, and even non-speakers inevitably learn some vocabulary and leverage this in the course of transcribing. It is premature to "bake in" the word boundaries before we know whether a form is meaningful.

3. Leveraging translations. Many computational methods leverage translations (Section 3.2), benefiting from the fact that they are often available (Section 2.2.2). This accords with the decision to view translations as independently motivated in Figure 5.

4. Transcribing partially. Sections of a corpus that need to be submitted to conventional linguistic analysis require contiguous or "dense" transcription. We can locate items for dense transcription by indexing and searching the audio using spoken term detection or alignment to translations. This aligns with the assumptions of mid-century linguists, and the present reality, that not everything will be transcribed (Section 2.2.2).

5. Working with speakers. We observed that there is usually some local capacity for transcription (Section 2.2.3). However, the phone recognition approaches demand a linguist's expertise in phonetic transcription. In contrast, the other approaches stand to make better use of speakers who are not trained linguists.

6. Simultaneous linking. Users of existing software must decide if they want their transcriptions to be anchored to the audio or the lexicon (Section 2.3). A major source of the difficulty in creating an integrated tool has been the reliance on phone transcription. However, if we transcribe at the word level, there should be no particular difficulty in simultaneously linking to the signal and the lexicon.

This concludes our analysis of existing computational approaches to the processing of oral languages. Next we turn to a new approach, sparse transcription.

4. The Sparse Transcription Model

The Sparse Transcription Model supports interpretive, iterative, and interactive processes for transcribing of oral languages while avoiding the transcription bottleneck (Figure 1). It organizes the work into tasks which can be combined into a variety of workflows. We avoid making premature commitments to details which can only be resolved much later, if at all, such as the precise locus of boundaries, the representation of forms as alphabetic strings, or the status of forms as morphemes, words, or multiword expressions. We privilege the locally available resources and human capacity in a strengths-based approach.

In this section I present an overview of the model (Section 4.1), followed by a discussion of transcription tasks (Section 4.2) and workflows (Section 4.3), and then some thoughts on evaluation (Section 4.4). I conclude the section by showing how the Sparse Transcription Model addresses the requirements for transcription (Section 4.5).

4.1 Overview

The Sparse Transcription Model consists of several data types, together serving as the underlying representation for IGT_1 (3). The model includes sparse versions of IGT_1 in which transcriptions are not contiguous. The data types are as follows:

Timeline: an abstraction over one or more audio recordings that document a linguistic event. A timeline serves to resolve time values to offsets in an audio file, or the audio track of a video file, following Bird and Liberman (2001, pages 43f). Timelines are per-participant in the case of conversational speech. We assume the existence of some independent means to identify breath groups per speaker in the underlying audio files.

Glossary Entry: a form-gloss pair. Figure 6 shows a fragment of a glossary. Rows 47–49 represent transient stages when we have identified a form but not its meaning (entry 47), or identified a meaning without specifying the form (entry 48), or where we use a stub to represent the case of detecting multiple tokens of a single yet-to-be-identified lexeme (entry 49). The string representation of a form is subject to revision.

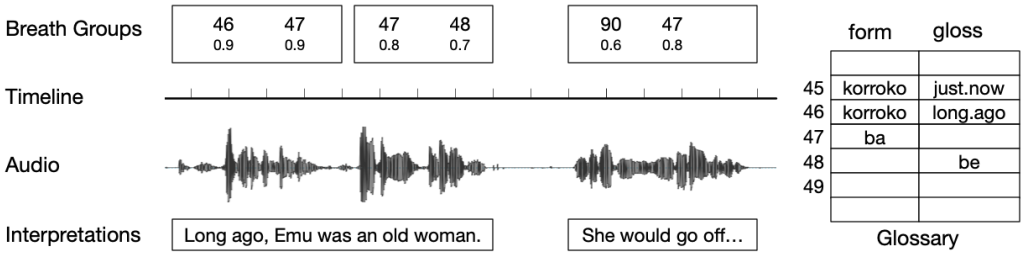


Figure 6
Sparse Transcription Model: Breath groups contain tokens which link to glossary entries; breath groups are anchored to the timeline and contained inside interpretations. Tokens are marked for their likelihood of being attested in a breath group, and optionally with a time, allowing them to be sequenced within this breath group; cf. (4).

This representation is related to OntoLex (McCrae et al. 2017). A form *korroko* stands in for a pointer to an OntoLex `Form`, and a gloss *long.ago* stands in for a pointer to a `LexicalSense`, and thence an `OntologyEntry`. A glossary entry is a subclass of `LexicalEntry`, or one of its subclasses `Word`, `MultiwordExpression`, or `Affix`. Glossary entries may contain further fields such as the morphosyntactic category, the date of elicitation, and so on. Entries may reference constituents, via a pointer from one `LexicalEntry` to another.

Token: a tuple consisting of a glossary entry and a breath group, along with a confidence value and an optional time. Examples of tokens are shown in Figure 6. A single time point enables us to sequence all tokens of a breath group. This avoids preoccupation with precise boundary placement, and it leaves room for later addition of hard-to-detect forms that appear between previously identified forms.

Breath Group: a span of a timeline that corresponds to a (partial) utterance produced between breath pauses. These are visible in an audio waveform and serve as the units of manual transcription (cf. Figure 2). Breath groups become the mini spoken documents that are the basis for spoken term detection. Breath groups are non-overlapping. Consecutive breath groups are not necessarily contiguous, skipping silence or extraneous content. Breath groups carry a size value that stores the calculated duration of speech activity.

Interpretation: a representation of the meaning of an utterance (one or more consecutive breath groups). Interpretations may be in the form of a text translation into another language, or an audio recording, or even an image. There may be multiple interpretations assigned to any timeline, and there is no requirement that two interpretations that include the same breath group have the same temporal extent.

4.2 Transcription Tasks

Glossary. Many workflows involve growing the glossary: reviewing existing glossaries with speakers; eliciting words from speakers; previewing some recordings and eliciting words; processing archived transcriptions and entering the words.

Task G (Growing the Glossary: Manual)
Identify a candidate entry for addition and check if a similar entry is already present. If so, possibly update the existing entry. If not, create a new entry *ge*, and elicit one or

more reference recordings for this item and add to A, and create the token(s) to link this to *ge*. Update our estimate of the typical size of this entry (duration of its tokens).

Task R (Refactoring the Glossary: Semi-automatic)
Identify entries for merging, based on clustering existing entries; or for splitting, based on clustering the tokens of an entry; or for decomposition, based on morphological discovery. Submit these for manual validation.

Breath Groups. Each audio timeline contains breath groups B. We store the amount of speech activity in a breath group. We mark fragments that will not be transcribed (e.g., disfluent or unintelligible speech).

Task B (Breath Group Identification: Semi-automatic)
Automatically identify breath groups and submit for manual review. Compute and store the duration of speech activity in a breath group. Permit material within a breath group to be excluded from transcription. Permit updating of the span of a breath group through splitting or merging.

Interpretation. Interpretations are any kind of semiotic material. This task elicits interpretations, and aligns existing interpretations to utterances of the source audio. We assume that the span of an interpretation $[t_1, t_2]$ corresponds to one or more breath groups.

Task I (Interpretation: Manual)
Associate interpretations with timeline spans that correspond to one or more consecutive breath groups.

Transcription. The first task is word spotting, or spoken term detection, with the requirement that terms can be assigned a gloss. This may leverage time-aligned translations if they are available. The second task is contiguous, or dense transcription, which leverages word spotting for allocating transcription effort to breath groups of interest, and then for accelerating transcription by guessing which words are present.

Task S (Word Spotting: Automatic)
Find possible instances of a given glossary entry and create a new token for each one. This token identifies a breath group and specifies a confidence value.

Task T (Contiguous Transcription: Manual)
Identify successive forms in a breath group, specifying a time, and linking them to the glossary, expanding the glossary as needed (Task G).

Completion. Here we focus on gaps, aiming to produce contiguous transcriptions. We need to be able to identify how much of a breath group still needs to be transcribed in order to prioritize work. Once a breath group is completed, we want to display a view

of the transcription, consisting of form-gloss pairs like (3) or a speech concordance, with support for navigating to the audio or the lexicon.

Task C (Coverage: Automatic)
Calculate the transcriptional coverage for the breath groups of an audio recording.

Task V (View: Automatic)
Generate transcript and concordance views of one or more audio recordings.

This concludes our formalization of transcription tasks. We envisage that further tasks will be defined, extending this set. Next we turn to the workflows that are built from these tasks.

4.3 Transcription Workflows

Classical Workflows. Following Franz Boas, Ken Hale, and countless others working in the intervening century, we begin by compiling a glossary, and we regularly return to it with revisions and enrichments. We may initialize it by incorporating existing wordlists and eliciting further words (Task G). As we grow the glossary, we ideally record each headword and add the recording to the corpus.

Now we begin collecting and listening to recordings. Speakers readily interpret them utterance by utterance into another language (Task I). These utterances will involve one or more breath groups, and this is an efficient point to create and review the breath groups, possibly splitting or merging them, and bracketing out any extraneous material (Task B). These two tasks might be interleaved.

Next, we leverage our glossary and interpretations to perform word spotting (Task S), to produce initial sparse transcription. We might prioritize transcription of a particular topic, or of texts that are already well covered by our sparse transcription.

Now we are ready for contiguous transcription (Task T), working one breath group at a time, paying attention to regions not covered by our transcription (Task C), and continually expanding the glossary (Task G). Now that we have gained evidence for new words we return to word spotting (our previous step).

We postpone transcription of any item that is too difficult. We add new forms to the glossary without regard for whether they are morphemes, words, or multiword expressions. We take advantage of the words that have already been spotted in the current breath group (Task S), and of system support that displays them according to confidence and suggests their location. Once it is complete, we visualize the transcription (Task V).

This workflow is notated in (6). Concurrent tasks are comma-separated, and there is no requirement to complete a task before continuing to a later task. We can jump back any number of steps to revise and continue from that point.

(6) $G \longrightarrow B, I \longrightarrow S \longrightarrow T, G, C \longrightarrow V$

Sociolinguistic Workflow. Example (7) represents a practice in sociolinguistics, where we identify words that reveal sociolinguistic variables and then try to spot them in our audio corpus. When we find some hits, we review the automatically-generated breath

groups, interpret, and transcribe. This task cuts across audio recordings and does not result in a normal text view, only a concordance view.

(7) $G \longrightarrow S \longrightarrow B, I, T \longrightarrow V$

Elicitation Workflow. Example (8) shows a workflow for elicitation. Here the audio recordings were collected by prompting speakers, perhaps using images. The first step is to align these prompts with the audio.

(8) $I \longrightarrow B \longrightarrow T, G$

Archival Workflows. Another workflow comes from the context of working with materials found in libraries and archives. We may OCR the texts that appear in the back of a published grammar, then locate these in a large audio collection with the help of grapheme-to-phoneme rules and a universal pronunciation model. We would probably extract vocabulary from the transcriptions to add to the glossary. We might forcibly align the transcriptions to the audio, and then create tokens for each word. Then begins the process of refactoring the glossary.

Glossary Workflows. Working on the glossary may be a substantial undertaking in itself, and may even be the driver of the sparse transcription activity. Early on, we will not know if a form is a morph, a word, or a multiword expression. The presence of allophony and allomorphy may initially multiply the number of entries. What originally looks like homonymy (two entries with a shared form) may turn out to be a case of polysemy (one entry with two glosses), or vice versa, leading us to refactor the glossary (Task R). Clues to word meaning may come from interpretations (if available), but they are just as likely to be elicited, a process that may generate extended commentary, itself a source text which may be added to the audio corpus. Word-spotting and the concordance view may support lexicographic research over a large audio corpus, which itself may never be transcribed more than required for the lexicographic endeavor.

Collaborative Workflows. Transcribing oral languages is collaborative, and the use of computational methods brings us into the space of computer-supported cooperative language documentation (Hanke 2017). We note that the Sparse Transcription Model is designed to facilitate asynchronous update: e.g., we can insert a token without needing to modify the boundaries of existing tokens; multiple transcribers can be spotting words in the same audio files concurrently; several hypotheses for the token at a given location can be represented. It is a separate task, outside the scope of this article, to model lifecycle workflows covering the selection of audio, allocation of transcription tasks to individuals, and various processes for approving content and making it available.

Transcribing First vs. Translating First. Note that interpretations and transcriptions are independently related to the timeline, and so the model is neutral as to which comes first. The similarity-based clustering of tokens or of lexical entries could be sensitive to interpretations when they are available.

Finally, we can observe that the notation for representing transcription workflows offers a systematic means for documenting transcription practices "in order to get the full picture of which strategies and which forms of knowledge are applied in transcription" (Himmelmann 2018, page 37).

4.4 Evaluation

When it comes to recordings made in field conditions, we need to take seriously the fact that there is no such thing as "gold transcription," as there are so many sources of significant variability (Cucchiarini 1993; Bucholtz 2007). I am not aware of spontaneous speech corpora where multiple human-validated phone transcriptions of the same audio are supplied; cf. the situation for machine translation where source texts may be supplied with multiple translations to support the BLEU metric (Papineni et al. 2002).

One proposal is to apply abductive inference in the presence of a comprehensive lexicon: "a transcription is 'good enough' when a human being looking at the transcription can work out what the words are supposed to be" (Bettinson 2013, page 57). This is the case of *inscription*, efficiently capturing an event in order to reconstruct it later.

The metric of word error rate is usually applied to orthographic words, which tend to collapse meaningful distinctions or introduce spurious distinctions (e.g., *wind* [wɪnd] *vs* [waɪnd]; *color vs*. *colour*). Instead, we propose "lexeme error rate," the obvious measure of the correctness of a lexeme-level transcription (Task T).

Word spotting (Task S) can be evaluated using the usual retrieval measures such as precision and recall. This generalizes to the use of word spotting for selecting audio recordings to transcribe. We have no need for measures of segmentation quality. Our measure of completeness (Task C) does not require segmentation. A measure of the quality of lexical normalization still needs to be considered. We must avoid the combinatorial explosion caused by multiplying out morphological possibilities, while retaining human judgments about the acceptability of morphologically complex forms.

4.5 Addressing the Requirements

The Sparse Transcription Model meets the requirements for transcription (Section 2.4). We transcribe at the word level, and these words are our meaningful units. We transcribe partially, leveraging translations. We draw on the knowledge and skills of speakers, and produce a result that is simultaneously linked to the audio and the lexicon. To reiterate, the model supports "standard practice to link transcripts [to] recordings, which allows direct and fast access" (Himmelmann 2006a, page 259), while taking seriously the fact that "transcription necessarily involves hypotheses as to the meaning of the segment being transcribed and the linguistic forms being used" (Himmelmann 2018, page 35).

Some significant modeling topics remain. First, the glossary may contain redundancy in the form of variants, constituents, and compositions. We assume there is computational support for refactoring the glossary. Second, nothing has been said about how word spotting should be performed. When identifying new tokens of a glossary entry, we would like to call on all available information, including elicited forms, contextual forms, differently inflected forms, unrelated forms which have syllables in common, and so forth. Available interpretations may contribute to this process. Many approaches are conceivable depending on the typology of the language: Does the language have syllable margins characterized by minimal coarticulation? How synthetic is the language, and what is the degree of agglutination or fusion?

Significant implementation topics remain open. First, we say little about the user interface except that it must be able to display sense-disambiguated word-level transcriptions. Each manual task could have software support of its own, for example, a mobile interface where users swipe right or left to confirm or disconfirm hits for a lexeme in phrasal context. Second, we have not articulated how such tools might be

integrated with established methods for language processing. These topics are left for further investigation.

5. Conclusion

Oral languages present a challenge for natural language processing, thanks to the difficulties of mapping from spontaneous speech to the expected text input. The most widely explored approach is to prefix a speech-to-text component to the existing NLP pipeline, keeping alive the hope for "a full-fledged automatic system which outputs not only transcripts, but glossed and translated texts" (Michaud et al. 2018, page 22).

However, this position ignores the interpretive nature of transcription, and it intensifies the transcription bottleneck by introducing tasks that can only be performed by the most scarce and expensive participants, namely linguists. Instead of a deficit framing where "unwritten languages" are developed into written languages, we can adopt a strengths-based model where "oral languages" present their own opportunities. We embrace the interests and capacities of the speech community and design tasks and tools that support local participation. After all, these people represent the main workforce and the main beneficiary of language work.

In developing a new approach, I have shown how a pipeline conception of the processing task has led us to make unhelpful assumptions, to transcribe phones, to transcribe fully, and to transcribe first. Long-established transcriptional practices demonstrate that these assumptions are unwarranted. Instead, I proposed to represent a transcription as a mapping from locations in the speech stream to an inventory of meaningful units. We use a combination of spoken term detection and manual transcription to create these tokens. As we identify more instances of a word, and similar sounding words, we get better at spotting them, and accelerate our work of orthodox, contiguous transcription.

Have we addressed the transcription bottleneck? The answer is a qualified yes. When transcriptions are not the data, we are freed from transcribing everything. When we avoid unnecessary and time-consuming tasks, we can allocate scarce resources more judiciously. When we use spoken term detection to identify recordings of interest, we go straight to our descriptive or analytical tasks. We identify linguistically meaningful units at an early stage without baking in the boundaries. We are more selective in where to "transcribe exhaustively." This is how we manage the transcription trap.

This approach offers new promises of scalability. It embraces the Zipfian distribution of linguistic forms, allocating effort according to frequency. It facilitates linguists in identifying the words, phrases, and passages of interest. It treats each additional resource—including untranscribed audio and interpretations—as further supervision to help us identify the meaningful units. It postpones discussion of orthographic representation. It offers diverse workflows and flexibility in the face of diverse language situations. Finally, it opens up new opportunities to engage with speakers. These people offer the deepest knowledge about the language and the best chance of scaling up the work. This is their language.

Acknowledgments

I am indebted to the Bininj people of the Kuwarddewardde "Stone Country" in Northern Australia for the opportunity to live and work in their community, where I gained many insights in the course of learning to transcribe Kunwinjku. Thanks to Steve Abney, Laurent Besacier, Mark Liberman, Maïa Ponsonnet, to my colleagues and students in the Top End Language Lab at Charles Darwin University, and to several anonymous reviewers for thoughtful feedback. This research has been supported by a grant from the Australian Research

Council, *Learning English and Aboriginal Languages for Work*.

References

Abiteboul, Serge, Peter Buneman, and Dan Suciu. 2000. *Data on the Web: From Relations to Semistructured Data and XML*. Morgan Kaufmann.

Abney, Steven and Steven Bird. 2010. The Human Language Project: Building a universal corpus of the world's languages. In *Proceedings of the 48th Meeting of the Association for Computational Linguistics*, pages 88–97, Uppsala.

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

Adams, Oliver, Trevor Cohn, Graham Neubig, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pages 3356–3365, Miyazaki.

Adams, Oliver, Trevor Cohn, Graham Neubig, and Alexis Michaud. 2017. Phonemic transcription of low-resource tonal languages. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 53–60, Brisbane.

Adams, Oliver, 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 246–255, Da Nang.

Adams, Oliver, Graham Neubig, Trevor Cohn, and Steven Bird. 2016a. Learning a translation model from word lattices. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, pages 2518–2522, San Francisco, CA. DOI: https://doi.org/10.21437/Interspeech.2016-862

Adams, Oliver, Graham Neubig, Trevor Cohn, Steven Bird, Quoc Truong Do, and Satoshi Nakamura. 2016b. Learning a lexicon and translation model from phoneme lattices. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 2377–2382, Austin, TX. DOI: https://doi.org/10.18653/v1/D16-1263

Albright, Eric and John Hatton. 2008. WeSay, a tool for engaging communities in dictionary building. In Victoria D. Rau and Margaret Florey, editors, *Documenting and Revitalizing Austronesian Languages*, number 1 in Language Documentation and Conservation Special Publication. University of Hawai'i Press, pages 189–201.

Anastasopoulos, Antonios, Sameer Bansal, David Chiang, Sharon Goldwater, and Adam Lopez. 2017. Spoken term discovery for language documentation using translations. In *Proceedings of the Workshop on Speech-Centric NLP*, pages 53–58, Copenhagen. DOI: https://doi.org/10.18653/v1/W17-4607

Anastasopoulos, Antonios and David Chiang. 2017. A case study on using speech-to-translation alignments for language documentation. In *Proceedings of the Workshop on the Use of Computational Methods in Study of Endangered Languages*, pages 170–178, Honolulu, HI. DOI: https://doi.org/10.18653/v1/W17-0123

Anastasopoulos, Antonios, David Chiang, and Long Duong. 2016. An unsupervised probability model for speech-to-translation alignment of low-resource languages. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 1255–1263, Austin, TX. DOI: https://doi.org/10.18653/v1/D16-1133

Anastasopoulos, Antonis and David Chiang. 2018. Leveraging translations for speech transcription in low-resource settings. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, pages 1279–1283, Hyderabad. DOI: https://doi.org/10.21437/Interspeech.2018-2162

Austin, Peter K. 2007. Training for language documentation: Experiences at the School of Oriental and African Studies. In D. Victoria Rau and Margaret Florey, editors, *Documenting and Revitalizing Austronesian Languages*, number 1 in Language Documentation and Conservation Special Issue, University of Hawai'i Press, pages 25–41.

Bansal, Sameer, Herman Kamper, Adam Lopez, and Sharon Goldwater. 2017. Towards speech-to-text translation without speech recognition. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 474–479, Valencia. DOI: https://doi.org/10.18653/v1/E17-2076, PMID: 28345436

Besacier, Laurent, Bowen Zhou, and Yuqing Gao. 2006. Towards speech translation of non written languages. In *Spoken Language*

Technology Workshop, pages 222–225, IEEE. DOI: https://doi.org/10.1109/SLT.2006.326795

Bettinson, Mat. 2013. The effect of respeaking on transcription accuracy. Honours Thesis, Department of Linguistics, University of Melbourne.

Bird, Steven. 2010. A scalable method for preserving oral literature from small languages. In *Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries*, pages 5–14, Gold Coast. DOI: https://doi.org/10.1007/978-3-642-13654-2_2

Bird, Steven. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain. To appear.

Bird, Steven and David Chiang. 2012. Machine translation for language preservation. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 125–134, Mumbai.

Bird, Steven, Florian Hanke, Oliver Adams, and Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–5, Baltimore, MD. DOI: https://doi.org/10.3115/v1/W14-2201

Bird, Steven and Jonathan Harrington, editors. 2001. *Speech Communication: Special Issue on Speech Annotation and Corpus Tools*, 33 (1–2). Elsevier. DOI: https://doi.org/10.1016/S0167-6393(00)00066-2

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media. DOI: https://doi.org/10.1016/S0167-6393(00)00068-6

Bird, Steven and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33:23–60.

Black, H. Andrew and Gary F. Simons. 2008. The SIL FieldWorks Language Explorer approach to morphological parsing. In Nicholas Gaylord, Stephen Hilderbrand, Heeyoung Lyu, Alexis Palmer, and Elias Ponvert, editors, *Computational Linguistics for Less-studied Languages: Proceedings of Texas Linguistics Society 10*, CSLI, pages 37–55.

Boas, Franz, editor. 1911. *Handbook of American Indian Languages*, volume 40 of *Smithsonian Institution Bureau of American Ethnology Bulletin*. Washington, DC: Government Printing Office.

Boito, Marcely Zanon, Antonios Anastasopoulos, Aline Villavicencio, Laurent Besacier, and Marika Lekakou. 2018. A small Griko-Italian speech translation corpus. In *6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 36–41, Gurugram. DOI: https://doi.org/10.21437/SLTU.2018-8

Bouquiaux, Luc and Jacqueline M. C. Thomas. 1992. *Studying and describing unwritten languages*. Dallas, TX: Summer Institute of Linguistics.

Bowern, Claire. 2008. *Linguistic Fieldwork: A Practical Guide*. Palgrave Macmillan. DOI: https://doi.org/10.1017/S0952675700001019

Browman, Catherine and Louis Goldstein. 1989. Articulatory gestures as phonological units. *Phonology*, 6:201–251.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311.

Bucholtz, Mary. 2007. Variability in transcription. *Discourse Studies*, 9:784–808. DOI: https://doi.org/10.1177/1461445607082580

Buseman, Alan, Karen Buseman, and Rod Early. 1996. *The Linguist's Shoebox: Integrated Data Management and Analysis for the Field Linguist*. Waxhaw NC: SIL.

Butler, Lynnika and Heather Van Volkinburg. 2007. Fieldworks Language Explorer (FLEx). *Language Documentation and Conservation*, 1:100–106.

Cahill, Michael and Keren Rice, editors. 2014. *Developing orthographies for unwritten languages*. SIL International.

Cairns, Paul, Richard Shillcock, Nick Chater, and Joe Levy. 1997. Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, 33:111–153. DOI: https://doi.org/10.1006/cogp.1997.0649, PMID: 9245468

Cartwright, Timothy A. and Michael R. Brent. 1994. Segmenting speech without a lexicon: the roles of phonotactics and speech source. In *Proceedings of the First Meeting of the ACL Special Interest Group in Computational Phonology*, pages 83–90, Las Cruces, NM.

Chelliah, Shobhana. 2018. The design and implementation of documentation projects for spoken languages. In *Oxford Handbook of Endangered Languages*. Oxford University Press. DOI: https://doi.org/10.1093/oxfordhb/9780190610029.013.9

Chen, Nancy F., Chongjia Ni, I-Fan Chen, Sunil Sivadas, Haihua Xu, Xiong Xiao, Tze Siong Lau, Su Jun Leow, Boon Pang Lim, Cheung-Chi Leung, et al. 2015. Low-resource keyword search strategies for Tamil. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 5366–5370, IEEE. DOI: https://doi.org/10.1109/ICASSP.2015.7178996 PMID: 26244568

Chung, Yu An, Wei-Hung Weng, Schrasing Tong, and James Glass. 2019. Towards unsupervised speech-to-text translation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 7170–7174, Brisbane. DOI: https://doi.org/10.1109/ICASSP.2019.8683550

Clifford, James. 1990. Notes on (field) notes. In Roger Sanjek, editor, *Fieldnotes: The Makings of Anthropology*. Cornell University Press, pages 47–70. DOI: https://doi.org/10.7591/9781501711954-004, PMID: 29994340

Cox, Christopher, Gilles Boulianne, and Jahangir Alam. 2019. Taking aim at the 'transcription bottleneck': Integrating speech technology into language documentation and conservation. Paper presented at the 6th International Conference on Language Documentation and Conservation, Honolulu, HI, https://instagram.com/p/Buho4Z0B7xT/

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

Cucchiarini, Catia. 1993. *Phonetic Transcription: A Methodological and Empirical Study*. Ph.D. thesis, Radboud University.

Do, Van Hai, Nancy F. Chen, Boon Pang Lim, and Mark Hasegawa-Johnson. 2016. Analysis of mismatched transcriptions generated by humans and machines for under-resourced languages. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*, pages 3863–3867, San Francisco, CA. DOI: https://doi.org/10.21437/Interspeech.2016-736

Dobrin, Lise M. 2008. From linguistic elicitation to eliciting the linguist: Lessons in community empowerment from Melanesia. *Language*, 84:300–324. DOI: https://doi.org/10.1353/lan.0.0009

Dunbar, Ewan, 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 323–330, Okinawa. DOI: https://doi.org/10.1109/ASRU.2017.8268953

Duong, Long, Antonios Anastasopoulos, David Chiang, Steven Bird, and Trevor Cohn. 2016. An attentional model for speech translation without transcription. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 949–959, San Diego, CA. DOI: https://doi.org/10.18653/v1/N16-1109

Elsner, Micha, Sharon Goldwater, Naomi Feldman, and Frank Wood. 2013. A joint learning model of word segmentation, lexical acquisition, and phonetic variability. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 42–54, Seattle, WA.

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

Evans, Nicholas and Hans-Jürgen Sasse. 2007. Searching for meaning in the library of babel: field semantics and problems of digital archiving. *Archives and Social Studies: A Journal of Interdisciplinary Research*, 1:63–123.

Fiscus, Jonathan G., Jerome Ajot, John S. Garofolo, and George Doddington. 2007. Results of the 2006 spoken term detection evaluation. In *Proceedings of the Workshop on Searching Spontaneous Conversational Speech*, pages 51–57, Amsterdam.

Foley, Ben, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, František Kratochví, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 205–209, Gurugram. DOI: https://doi.org/10.21437/SLTU.2018-42

Gales, Mark, 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, St. Petersburg.

Garofolo, John S., Lori F. Lamel, William M. Fisher, Jonathon G. Fiscus, David S. Pallett, and Nancy L. Dahlgren. 1986. *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM*. NIST.

Gaved, Tim and Sophie Salffner. 2014. Working with ELAN and FLEx together. https://www.soas.ac.uk/elar/helpsheets/file122785.pdf, accessed 21 March 2019.

Godard, Pierre, 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 3539–3543, San Francisco, CA. DOI: https://doi.org/10.21437/Interspeech.2016-886

Godard, Pierre, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noel Kouarata, Lori Lamel, Hélène Maynard, Markus Mueller, Annie Rialland, Sebastian Stueker, François Yvon, and Marcely Zanon-Boito. 2018a. A very low resource language speech corpus for computational language documentation experiments. In *Proceedings of the 11th Language Resources and Evaluation Conference*, pages 3366–3370, Miyazaki.

Godard, Pierre, Marcely Zanon Boito, Lucas Ondel, Alexandre Berard, François Yvon, Aline Villavicencio, and Laurent Besacier. 2018b. Unsupervised word segmentation from speech with attention. In *Proceedings of the 19th Annual Conference of the International Speech Communication Association*, pages 2678–2682, Hyderabad. DOI: https://doi.org/10.21437/Interspeech.2018-1308

Goldwater, Sharon, Thomas Griffiths, and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112:21–54. DOI: https://doi.org/10.1016/j.cognition.2009.03.008, PMID: 19409539

Goldwater, Sharon, Thomas L. Griffiths, and Mark Johnson. 2006. Contextual dependencies in unsupervised word segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 673–680, Sydney. DOI: https://doi.org/10.3115/1220175.1220260

Gudschinsky, Sarah. 1967. *How to Learn an Unwritten Language*. Holt, Rinehart and Winston. DOI: https://doi.org/10.1017/CBO9780511810206.005

Hale, Ken. 2001. Ulwa (Southern Sumu): The beginnings of a language research project. In Paul Newman and Martha Ratliff, editors, *Linguistic Fieldwork*. Cambridge University Press, pages 76–101.

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

Hanke, Florian and Steven Bird. 2013. Large-scale text collection for unwritten languages. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 1134–1138, Nagoya.

Hasegawa-Johnson, Mark A., Preethi Jyothi, Daniel McCloy, Majid Mirbagheri, Giovanni M di Liberto, Amit Das, Bradley Ekin, Chunxi Liu, Vimal Manohar, Hao Tang, and others. 2016. ASR for under-resourced languages from probabilistic transcription. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 25:50–63. DOI: https://doi.org/10.1109/TASLP.2016.2621659

Hatton, John. 2013. SayMore: Language documentation productivity. Paper presented at the Third International Conference on Language Documentation and Conservation, http://hdl.handle.net/10125/26153

Hermes, Mary and Mel Engman. 2017. Resounding the clarion call: Indigenous language learners and documentation. *Language Documentation and Description*, 14:59–87.

Himmelmann, Nikolaus. 1998. Documentary and descriptive linguistics. *Linguistics*, 36:161–195. DOI: https://doi.org/10.1515/ling.1998.36.1.161

Himmelmann, Nikolaus. 2006a. The challenges of segmenting spoken language. In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel, editors, *Essentials of Language Documentation*. Mouton de Gruyter, pages 253–274.

Himmelmann, Nikolaus. 2006b. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel, editors, *Essentials of Language Documentation*. Mouton de Gruyter, pages 1–30.

Himmelmann, Nikolaus. 2018. Meeting the transcription challenge. In *Reflections on Language Documentation 20 Years after Himmelmann 1998*, number 15 in Language Documentation and Conservation Special Publication, University of Hawai'i Press, pages 33–40.

Jacobson, Michel, Boyd Michailovsky, and John B. Lowe. 2001. Linguistic documents synchronizing sound and text. *Speech Communication*, 33:79–96. DOI: https://doi.org/10.1016/S0167-6393(00)00070-4

Jansen, Aren, Kenneth Church, and Hynek Hermansky. 2010. Towards spoken term discovery at scale with zero resources. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, pages 1676–1679, Chiba.

Johnson, Mark and Sharon Goldwater. 2009. Improving nonparameteric Bayesian inference: Experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*, pages 317–325, Boulder, CO. DOI: https://doi.org/10.3115/1620754.1620800

Jukes, Anthony. 2011. Researcher training and capacity development in language documentation. In *The Cambridge Handbook of Endangered Languages*. Cambridge University Press, pages 423–445. DOI: https://doi.org/10.1017/CBO9780511975981.021

Kaufman, Daniel and Ross Perlin. 2018. Language documentation in diaspora communities. In *Oxford Handbook of Endangered Languages*. Oxford University Press, pages 399–418. DOI: https://doi.org/10.1093/oxfordhb/9780190610029.013.20

King, Alexander D. 2015. Add language documentation to any ethnographic project in six steps. *Anthropology Today*, 31:8–12. DOI: https://doi.org/10.1093/oxfordhb/9780190610029.013.20

Liu, Chunxi, Aren Jansen, Guoguo Chen, Keith Kintzley, Jan Trmal, and Sanjeev Khudanpur. 2014. Low-resource open vocabulary keyword search using point process models. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association*, pages 2789–2793, Liu.

Maddieson, Ian. 2001. Phonetic fieldwork. In Paul Newman and Martha Ratliff, editors, *Linguistic Fieldwork*. Cambridge University Press, pages 211–229. DOI: https://doi.org/10.1017/CBO9780511810206.011

McCrae, John P., Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The Ontolex-Lemon model: Development and applications. In *Proceedings of the eLex Conference*, pages 19–21, Leiden.

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

Metze, Florian, Ankur Gandhe, Yajie Miao, Zaid Sheikh, Yun Wang, Di Xu, Hao Zhang, Jungsuk Kim, Ian Lane, Won Kyum Lee, et al. 2015. Semi-supervised training in low-resource ASR and KWS. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 4699–4703, Brisbane.

Michaud, Alexis, Oliver Adams, Trevor Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone Toolkit. *Language Documentation and Conservation*, 12:481–513.

Moe, Ron. 2008. *FieldWorks Language Explorer 1.0*. Number 2008–011 in SIL Forum for Language Fieldwork. SIL International.

Mosel, Ulrike. 2006. Fieldwork and community language work. In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel, editors, *Essentials of Language Documentation*. Mouton de Gruyter, pages 67–85.

Myers, Cory, Lawrence Rabiner, and Andrew Rosenberg. 1980. An investigation of the use of dynamic time warping for word spotting and connected speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 173–177, Denver, CO.

Nathan, David and Meili Fang. 2009. Language documentation and pedagogy for endangered languages: A mutual revitalisation. *Language Documentation and Description*, 6:132–160.

Neubig, Graham, Masato Mimura, Shinsuke Mori, and Tatsuya Kawahara. 2010. Learning a language model from continuous speech. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association*, pages 1053–1056, Chiba.

Neubig, Graham, 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 165–174, Jeju Island.

Newman, Paul and Martha Ratliff. 2001. Introduction. In Paul Newman and Martha Ratliff, editors, *Linguistic Fieldwork*. Cambridge University Press.

Norman, Don. 2013. *The Design of Everyday Things*. Basic Books.

Ochs, Elinor. 1979. Transcription as theory. *Developmental Pragmatics*, 10:43–72.

Ostendorf, Mari. 1999. Moving beyond the 'beads-on-a-string' model of speech. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pages 79–84, Keystone, USA.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA. DOI: https://doi.org/10.3115/1073083.1073135

Park, Alex and James Glass. 2008. Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 16:186–197. DOI: https://doi.org/10.1109/TASL.2007.909282

Pike, Kenneth L. 1947. *Phonemics: A Technique for Reducing Language to Writing*. Ann Arbor: University of Michigan Press.

Rapidwords. 2019. Rapid Word Collection. rapidwords.net, accessed 26 June 2019.

Rath, Shakti, Kate Knill, Anton Ragni, and Mark Gales. 2014. Combining tandem and hybrid systems for improved speech recognition and keyword spotting on low resource languages. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association*, pages 14–18, Singapore.

Reiman, Will. 2010. Basic oral language documentation. *Language Documentation and Conservation*, 4:254–268.

Rialland, Annie, Martine Adda-Decker, Guy-Noël Kouarata, Gilles Adda, Laurent Besacier, Lori Lamel, Elodie Gauthier, Pierre Godard, and Jamison Cooper-Leavitt. 2018. Parallel corpora in Mboshi (Bantu C25, Congo-Brazzaville). In *Proceedings of the 11th Language Resources and Evaluation Conference*, pages 4272–4276, Miyazaki.

Rice, Keren. 2001. Learning as one goes. In Paul Newman and Martha Ratliff, editors, *Linguistic Fieldwork*. Cambridge University Press, pages 230–249. DOI: https://doi.org/10.1017/CBO9780511810206.012

Rice, Keren. 2009. Must there be two solitudes? Language activists and linguists working together. In Jon Reyhner and Louise Lockhard, editors, *Indigenous language revitalization: Encouragement, guidance, and lessons learned*. Northern Arizona University, pages 37–59.

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

Robinson, Stuart, Greg Aumann, and Steven Bird. 2007. Managing fieldwork data with Toolbox and the Natural Language Toolkit. *Language Documentation and Conservation*, 1:44–57.

Rogers, Chris. 2010. Fieldworks Language Explorer (FLEx) 3.0. *Language Documentation and Conservation*, 4:78–84.

Rohlicek, Jan Robin. 1995. Word spotting. In Ravi P. Ramachandran and Richard J. Mammone, editors, *Modern Methods of Speech Processing*. Springer, pages 123–157. DOI: https://doi.org/10.1007/978-1-4615-2281-2_6

Samarin, William. 1967. *Field Linguistics: A Guide to Linguistic Field Work*. Holt, Rinehart and Winston.

Sanjek, Roger. 1990. The secret life of fieldnotes. In Roger Sanjek, editor, *Fieldnotes: The Makings of Anthropology*. Cornell University Press, pages 187–272. DOI: https://doi.org/10.7591/9781501711954

Sapién, Racquel María. 2018. Design and implementation of collaborative language documentation projects. In *Oxford Handbook of Endangered Languages*. Oxford University Press, pages 203–224. DOI: https://doi.org/10.1093/oxfordhb/9780190610029.013.12

Schultze-Berndt, Eva. 2006. Linguistic annotation. In Jost Gippert, Nikolaus Himmelmann, and Ulrike Mosel, editors, *Essentials of Language Documentation*. Mouton de Gruyter, pages 213–251.

Seifart, Frank, Harald Hammarstroöm, Nicholas Evans, and Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language*, 94:e324–e345. DOI: https://doi.org/10.1353/lan.2018.0070

Shillcock, Richard. 1990. Lexical hypotheses in continuous speech. In Gerry Altmann, editor, *Cognitive Models of Speech Processing*. MIT Press, pages 24–49.

SIL Language Technology. 2000. Shoebox. https://software.sil.org/shoebox/, accessed 26 April 2020.

Sloetjes, Han, Herman Stehouwer, and Sebastian Drude. 2013. Novel developments in Elan. Paper presented at the Third International Conference on Language Documentation and Conservation, Honolulu, HI, http://hdl.handle.net/10125/26154. DOI: https://doi.org/10.1093/oxfordhb/9780199571932.013.019

Sperber, Matthias, Graham Neubig, Christian Fügen, Satoshi Nakamura, and Alex Waibel. 2013. Efficient speech transcription through respeaking. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association*, pages 1087–1091, Lyon.

Stahlberg, Felix, Tim Schlippe, Stephan Vogel, and Tanja Schultz. 2015. Cross-lingual lexical language discovery from audio data using multiple translations. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 5823–5827, Brisbane. DOI: https://doi.org/10.1109/ICASSP.2015.7179088

Stahlberg, Felix, Tim Schlippe, Stephan Vogel, and Tanja Schultz. 2016. Word segmentation and pronunciation extraction from phoneme sequences through cross-lingual word-to-phoneme alignment. *Computer Speech and Language*, 35:234–261. DOI: https://doi.org/10.1016/j.csl.2014.10.001

Tedlock, Dennis. 1983. *The Spoken Word and the Work of Interpretation*. University of Pennsylvania Press. DOI: https://doi.org/10.9783/9780812205305

Twaddell, William. F. 1954. A linguistic archive as an indexed depot. *International Journal of American Linguistics*, 20:108–110. DOI: https://doi.org/10.1086/464261

Valenta, Tomáš, Luboš Šmídl, Jan Švec, and Daniel Soutner. 2014. Inter-annotator agreement on spontaneous Czech language. In *Proceedings of the International Conference on Text, Speech, and Dialogue*, pages 390–397, Brno. DOI: https://doi.org/10.1007/978-3-319-10816-2_47

Voegelin, Charles Frederick and Florence Marie Voegelin. 1959. Guide for transcribing unwritten languages in field work. *Anthropological Linguistics*, pages 1–28.

Weiss, Ron, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association*, pages 2625–2629, Stockholm. DOI: https://doi.org/10.21437/Interspeech.2017-503

Winkelmann, Raphael and Georg Raess. 2014. Introducing a web application for labeling, visualizing speech and correcting derived speech signals. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 4129–4133, Reykjavic.

Woodbury, Anthony C. 1998. Documenting rhetorical, aesthetic, and expressive loss in language shift. In Lenore Grenoble and Lindsay Whaley, editors, *Endangered Languages: Language Loss and Community Response*. Cambridge University Press, pages 234–258. DOI: https://doi.org/10.1017/CBO9781139166959.011

Woodbury, Anthony C. 2003. Defining documentary linguistics. *Language Documentation and Description*, 1:35–51.

Woodbury, Anthony C. 2007. On thick translation in linguistic documentation. *Language Documentation and Description*, 4:120–135.

Wu, Dekai. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, pages 377–403.

Xia, Fei and William D. Lewis. 2007. Multilingual structural projection across interlinearized text. In *Proceedings of the North American Chapter of the Association for Computational Linguistics*. ACL, pages 452–459, Rochester, NY.

Yamada, Racquel María. 2014. Training in the community-collaborative context: A case study. *Language Documentation and Conservation*, 8:326–344.

Zanon Boito, Marcely, 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 458–465, Okinawa. DOI: https://doi.org/10.1109/ASRU.2017.8268972