

Automatic Speech Recognition for Supporting Endangered Language Documentation

Emily Prud'hommeaux
Boston College

Robbie Jimerson
Rochester Institute of Technology

Richard Hatcher
University at Buffalo

Karin Michelson
University at Buffalo

Generating accurate word-level transcripts of recorded speech for language documentation is difficult and time-consuming, even for skilled speakers of the target language. Automatic speech recognition (ASR) has the potential to streamline transcription efforts for endangered language documentation, but the practical utility of ASR for this purpose has not been fully explored. In this paper, we present results of a study in which both linguists and community members, with varying levels of language proficiency, transcribe audio recordings of an endangered language under timed conditions with and without the assistance of ASR. We find that both time-to-transcribe and transcription error rates are significantly reduced when correcting ASR for language learners of all levels. Despite these improvements, most community members in our study express a preference for unassisted transcription, highlighting the need for developers to directly engage with stakeholders when designing and deploying technologies for supporting language documentation.

1. Introduction¹ A substantial percentage of the world's languages are endangered and likely to fall out of use in this century (Krauss 1992; Moseley 2010; Seifart et al. 2018; Simons 2019). Although language preservation is not necessarily a priority for all speakers of endangered languages (Ladefoged 1992; Mufwene 2017), many endangered language communities in North America are actively working to create a permanent textual record of their language (Himmelman 1998) as a means to document their culture, to reclaim their heritage, to unify their communities, and to serve as a resource for language revitalization efforts,

¹ We are grateful for the cooperation and support of the Seneca Nation of Indians, especially Sandy Dowdy who continues to share her knowledge, voice, and enthusiasm for the Seneca language with our research team. This material is based upon work supported by the National Science Foundation under Grant No. 1761562. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Many aspects of language documentation are time consuming when the documentarian is not a fluent speaker of the language, which is often the case in endangered language documentation efforts. As the transcription of spontaneous speech has come to be the focus of much documentary linguistic work, one particular challenge is the “transcription bottleneck” (Seifart et al. 2018; Himmelman 2018). A recent survey of documentary linguists found that transcribing one hour of audio can take from 40 hours for word-level transcripts (Foley et al. 2018) to 60 hours for phone-level transcripts (Michaud et al. 2018). Automatic speech recognition (ASR) technology – variously known as voice recognition, speech-to-text, or dictation software – has the potential to break through this barrier by automatically generating text transcripts of speech samples.

Although ASR technology has been commercially available for decades and has long been used for specialized transcription tasks and as an assistive technology, it is only in the last few years that it has achieved error rates low enough for it to be used as a substitute for manual text entry by the general population. Indeed, ASR technology is routinely used today on smartphones and personal assistant devices by speakers of high-resource languages such as English and Mandarin. ASR for these high-resource languages has achieved “near-human” levels of accuracy not only because of the introduction of the much-touted technological advances associated with deep learning but also because of the large amount of labelled audio data currently available in these languages for training deep learning models. Large corpora of this sort have existed in academic settings for some time for English, Mandarin, Arabic, and a few European languages (Godfrey et al. 1992; Canavan & Zipperlen 1996; Canavan, Graff, & Zipperlen 1997; Canavan, Zipperlen, & Graff 1997; Cieri et al. 2004). Corporations like Google, Microsoft, Amazon, and Baidu have leveraged their ability to collect and access data to create substantial additional text and audio training corpora for these and other politically or economically important languages. The majority of the world’s languages, however, lack the quantities of training data necessary to train highly accurate ASR models. Although low-resource ASR is a growing area of interest in the speech signal processing research communities (Gelas et al. 2012; Thomas et al. 2013; Besacier et al. 2014; Nguyen et al. 2014; Metze et al. 2015; Gauthier et al. 2016; Scharenborg et al. 2017; Bansal et al. 2019; Stoian et al. 2020), the performance we see in our interactions with Siri or Alexa in a high-resource language is not what we can expect for a language with a few dozen hours of acoustic training data or for an endangered language with only a few hours of data.

Despite relatively high error rates, low-resource ASR can offer some utility for language documentation. Computer-assisted transcription, in which ASR output serves as a first-pass transcript that is then corrected or edited by a human, has been used for medical transcription for decades (Rosenthal et al. 1998; Borowitz 2001; Mohr et al. 2003; Hodgson & Coiera 2016; Goss et al. 2019). This same approach has the potential to be deployed for language documentation purposes. Despite a growing interest in developing tools, software, and frameworks that can be used by community members or field linguists engaged in language documentation (Strunk et al. 2014; Ćavar et al. 2016; Adams et al. 2018; Foley et al. 2018;

Michaud et al. 2018), there have been few empirical studies of the practical utility of ASR-based tools and technologies for this purpose.

We report here on an empirical study that demonstrates the efficacy of including ASR in a pipeline for transcribing recordings of Seneca (Onödowá'ga:), a language indigenous to what is now the western area of New York State in the United States and neighboring parts of Ontario, Canada. Seneca is classified by Ethnologue (Eberhard et al. 2021) as level 8a (“moribund”), with roughly 50 elderly first-language speakers and 100 or more second-language speakers, many of whom have participated in language immersion programs for both adults and children. We compare a traditional transcription method, in which transcribers incrementally listen to recordings and type what they hear, with an ASR-assisted approach. In this latter approach, raw speech recordings are first fed through a Seneca ASR system to produce a rough hypothesis of the transcription with the timestamps of the boundaries between utterances. The ASR output is then corrected by the transcriber within the speech analysis software, Praat. We find that correcting the output of even a relatively error-prone ASR system dramatically reduces both the time required to transcribe audio and the number of word-level and character-level errors in those transcriptions. These results hold for language learners at all language levels, from novice linguists to knowledgeable L2 community members. Despite the improvements in both the speed and the quality of transcription that ASR can provide, some participants in our study expressed a preference for unassisted transcription. Our results underscore the importance of considering the preferences of the community of speakers and the need for a flexible approach to the task of endangered language transcription.

2. Background

2.1 A note on documentation We note that the output of our work is time-stamped orthographic transcriptions of spontaneous speech, recorded during storytelling sessions and conversations with L2 learners. Because ASR acoustic models are trained to associate properties of the acoustic signal with specific phones, phone(me) labels with reasonable boundaries can also be trivially derived from the ASR output. We recognize that utterance-level transcripts are not necessarily the desired end product that a descriptive linguist or field linguist might aim to produce. We are not explicitly generating a lexicon or producing English glosses or translations, nor are we using ASR to transcribe individually elicited grammatical forms or paradigms.

We focus specifically on speech corpus documentation – the transcription of spontaneous speech samples – for several reasons. First, members of the Seneca community have expressed an interest in collecting precisely this kind of data. There are existing grammars and lexica of Seneca that are sufficient for teaching and understanding the morphology and syntax of Seneca, but there are few recordings or transcripts of Seneca speakers naturally using their language to communicate. This community believes that creating a record of the language as it is spoken will be more culturally valuable and more useful for developing instructional materials for their language immersion programs. Secondly, even for linguists who are interested in generating grammars and tables, there is no particular reason that transcription

and analysis must be carried out simultaneously. Many linguists produce a first-pass transcription and later analyze and gloss the transcription, returning to their consultants to ask about specific forms or constructions. Finally, we note that collecting and transcribing samples of spontaneous speech is now generally considered, if not a substitute for, then at least a supplement or complement to traditional paradigmatic elicitation (Himmelman 2018; Rice 2018; Seifart et al. 2018).

2.2 Automatic speech recognition (ASR) A full description of automatic speech recognition is beyond the scope of this paper, but we very briefly describe here the two frameworks that are commonly used in speech recognition today: the statistical framework relying on Gaussian mixture models (GMMs) and hidden Markov models (HMMs) dominant throughout the 1990s and early 2000s, and deep neural architectures, which are dominant today, particularly in high-resource scenarios. Both frameworks typically combine an acoustic model (i.e., the model that associates acoustic properties of the speech signal with labelled speech segments) and a language model (i.e., the model that is used to predict the sequence of textual segments, such as characters, morphemes, or words). The acoustic model is trained on speech recordings and transcripts of those recordings, ideally, though not necessarily, with segment labels and time stamps. The language model is trained on any available text data of the language, which in most cases far exceeds the amount of transcribed audio data. The framework used in this study is described below in Section 4.2.

The metric typically used to evaluate ASR performance is word error rate (WER), the length-normalized Levenshtein distance between the output and a ground-truth transcript. To calculate WER, an alignment between the output of the ASR system (the hypothesis) and a verified human-generated transcript (the reference) is derived by identifying the word deletions, insertions, and substitutions that would need to be made to the hypothesis to make it identical to the reference. From that alignment, the minimum number of word insertions, deletions, and substitutions is tallied; this sum is divided by the total number of words in the reference; and the result is multiplied by 100. The closer this number is to 0, the fewer errors were produced and, hence, the more accurate the ASR system can be said to be. In the example below, there are two substitutions, denoted in all caps in the hypothesis; and one deletion and one insertion, both denoted with asterisks in one text and all caps in the other. The WER of this output is $4/6 * 100 = 66.7\%$.

```
REF: This is a SHORT example *** sentence
HYP: THUS is AN *** example OF sentence
```

This same metric can be calculated for characters, phones, morphs, or any other subword unit, depending on the goals of the research and the characteristics of the target language. Before the introduction of deep neural architectures, state-of-the-art ASR for the US English conversational telephone speech dataset Switchboard (Godfrey et al. 1992) had plateaued at a WER of 20-25% (Hinton et al. 2012). WER on that same dataset has now reached close to 5% (Han et al. 2017; Saon et al. 2017;

Xiong et al. 2018), which is roughly the WER observed when trained native speakers transcribe this particular dataset (Glenn et al. 2010; Xiong et al. 2018).

Although the application of deep neural networks is certainly responsible for the enormous improvements in ASR accuracy over the past decade, the use of deep neural architectures would not have been possible without very large amounts of acoustic model training data. Switchboard includes 2000 hours of transcribed acoustic training data; Baidu's proprietary ASR technology reportedly is trained on 40,000 hours of acoustic training data. The amount of data required to train a robust ASR system with human-level accuracy is so large that languages with tens of millions of speakers like Haitian Creole and Bengali are considered "low-resource" languages for ASR purposes. Needless to say, while such languages have only dozens or hundreds of hours of *readily* available transcribed audio data, it would be relatively easy to acquire additional data. In addition, there are likely to be many millions of words of available text data for training the language model, which can have a significant impact on recognition accuracy (Chelba et al. 2012).

The amount of data available for training an ASR system for a typical endangered language is likely to be on an entirely different scale, with perhaps a handful of hours of transcribed audio and a few thousand words of text. Gathering additional data is likely to be difficult and time consuming given the small number of speakers and, in some cases, their reluctance to share their language with outsiders. An ASR system trained on such a small corpus will generally have a high WER, regardless of the architecture used to build the models. ASR systems for languages with fewer than 20 hours of acoustic training data yield WERs above 15% in the best of circumstances (Juan et al. 2015; Adams et al. 2019). When recording quality is low, the semantic domain is varied, the morphology is complex, or the writing system is not a more or less 1-to-1 character-to-phone or -syllable mapping, WERs rise substantially (Gelas et al. 2012; Gauthier et al. 2016; Adams et al. 2019).

2.3 ASR-assisted transcription ASR has been used in the service of improving the efficiency of speech transcription since the 1990s, with medical report transcription among the first practical applications of this technology. When ASR was first proposed for medical transcription, it was assumed that automated transcription would soon replace medical transcriptionists entirely, enabling clinicians themselves to dictate their reports directly into digital medical records, making only minor edits to the ASR output (e.g., Rosenthal et al. 1998). In the ensuing decades, studies exploring the utility of ASR for streamlining medical transcription have reported conflicting results, with relatively little investigation into the relationships between typical evaluation metrics, such as speed, accuracy, and keystroke savings, and variables such as transcriptionist skill, personal preference, and ASR word error rate (Goss et al. 2019). In one broad and careful study, Mohr et al. (2003) compared the efficiency of correcting ASR output to that of unassisted transcription, finding that correcting ASR output generally required more time and that medical transcriptionists preferred standard transcription. Similar results were reported by David et al. (2009), who reported that the majority of medical transcriptionists surveyed were unenthusiastic about using ASR and found that the task of correcting ASR was more

cognitively taxing than standard transcription, even if it reduced the time to transcribe. As the quality of ASR output for English has improved in the past several years, however, clinicians have noted higher satisfaction using ASR to dictate their reports (Goss et al. 2019).

More recently, ASR correction has been investigated for general purpose transcription, often with a focus on designing interfaces for displaying ASR hypotheses and optimizing methods for selecting among these hypotheses (Rodríguez et al. 2007; Revuelta-Martínez et al. 2012; Laurent et al. 2011). Bazillon et al. (2008) reported that editing ASR output is faster than transcribing manually but that the gains in speed vary according to whether the speech is spontaneous or prepared. Similarly, Akita et al. (2009) found that correcting ASR output was faster than unassisted transcription of lectures but that time to transcribe increased when error rates increased from 10% to 25% and that the transcriptionists expressed frustration when correcting inaccurate ASR output. In a very careful and comprehensive study comparing transcription from scratch with correcting ASR under various user-interface conditions, Sperber et al. (2016) found that correcting ASR was generally more efficient than transcribing from scratch but that the overhead associated with editing (e.g., backspacing, adjusting the cursor location) resulted in substantial reductions in speed when the number of character errors in a word to be corrected was large. In a follow-up study that explored the interaction between speed, error rate, and transcriber baseline efficiency, Sperber et al. (2017) reported that correcting ASR output was faster and more accurate than from-scratch transcription, that slow transcribers benefitted the most from ASR, and that these improvements were larger when word error rates were low.

Although the results from studies of general-domain ASR-assisted transcription suggest that there is potential utility in deploying ASR for endangered language documentation, there are a number of differences in the two tasks. In all of the prior studies, participants were transcribing a language in which they were expected to be proficient. Despite the lessons learned from the work on medical transcription, these studies generally did not include a discussion of user preference. Finally, the word error rates of the ASR systems used were much lower than what would typically be expected from an ASR for a truly low-resource endangered language.

2.4 ASR-based technologies for supporting language documentation There is substantial support in the documentary linguistics research community for using ASR and ASR-adjacent technologies to support language documentation (Blokland et al. 2015; Thieberger 2017; Gessler 2019; van Esch et al. 2019). Much of the previous work involving ASR for this purpose has focused on phone transcription and alignment. Prior work on forced alignment has found that models trained on either small amounts of carefully aligned data in the target language or large amounts of data from unrelated high-resource languages can be effectively leveraged to generate accurate phone segmentations for Yoloxóchitl Mixtec (DiCanio et al. 2012; DiCanio et al. 2013; Strunk et al. 2014), Chatino (Ćavar et al. 2016; Adams et al., 2018), Bribri (Coto-Solano & Solórzano 2017), Yongning Na (Michaud et al. 2018; Adams et al. 2019), and Australian Kriol (Jones et al. 2019). Other work in using ASR for

documentation has focused on developing user-friendly front ends for ASR model building, which typically requires significant computational expertise (Foley et al. 2018; Foley et al. 2019). Although these studies demonstrate the potential utility of speech recognition technology, whether at the phone or word level, for language documentation, none of these studies reports on the practical outcomes of deploying these systems for transcription by linguists or language community members.

2.5 Linguistic characteristics of Seneca (Onödowá'ga:) Seneca has a relatively small phonemic inventory, consisting of consonants /th, t, kh, k, ʔ, s, ʃ, h, tsh, ts, tʃh, tʃ, n, w/ and vowels /i, u, e, o, ē, ̄, æ, a/. The orthography that our Seneca speakers use and that is preferred by most users of the language represents these phones in a mostly unambiguous one-to-one mapping as, respectively: t, d, k, g, ʔ, s, ʃ, h, ts, dz, tʃ, j, n, w and i, u, e, o, ē, ̄, ä, a. Although many of our speakers of Seneca argue that vowel length is not phonemic, vowel length is typically indicated in the orthography using a colon.

Traditionally, the parts of speech of Seneca are nouns, verbs, and particles. More recently, kinship terms have been recognized as a fourth part of speech (Koenig & Michelson 2010). Nouns, verbs, and kinship terms are inflected, while particles generally occur in only one form. Verbs, kinship terms, and most nouns must have a pronominal prefix. The pronominal system of Seneca and of all the Iroquoian languages is highly complex. Seneca has 58 pronominal prefixes that identify the agent or both agent and patient arguments of the verb. Some analysis into meaningful elements is possible, but for the most part, these prefixes are treated as being synchronically unanalyzable. Verbs can also have one or more prepronominal prefixes. Verbs must have an aspect suffix, and most nouns have a noun suffix. Kinship terms have neither an aspect suffix nor a noun suffix, but most have an ending that can probably be identified as a diminutive ending (Chafe 2007).

Noun incorporation, or the compounding of a noun and verb stem to derive a verb stem, is extremely productive in Seneca. Verb stems can be derived from other verbs stems also by means of the reflexive/reciprocal and the semi-reflexive (or middle) prefixes, or by the benefactive or dative applicative, causative, dislocative or andative, distributive, inchoative, instrumental applicative, and reversative suffixes. Most morphemes have several allomorphs in the Iroquoian languages, and this is especially so for Seneca because it has undergone a large number of phonological changes relative to most of the other Iroquoian languages.

Verbs are pervasive in the Iroquoian languages, but the very low incidence of words with nominal morphology is a little misleading in that a striking property of the Iroquoian languages is the extent to which entities are referred to with words that are morphologically indistinguishable from verb forms. The Seneca word for 'chokecherry' is *deyagonyá'thā:s*, literally, 'it chokes people'; the word for 'school bus' is *hadiksa'danēhguis*, literally, 'they deliver children'. There is little, if any, evidence for formal syntactic constraints in Seneca. This is not to say that there are not syntactic constructions in which words must appear in a particular linear order, but in general, the relative order of verbs and nominal words, when they occur, is accounted for by pragmatic principle (Koenig & Michelson 2014).

3. Experimental design

3.1 Participants Five Seneca community members, all current or former participants in the adult language immersion program, with varying levels of fluency in Seneca (participants S1 through S5) and five linguists with varying degrees of exposure to Seneca or other Iroquoian languages (participants L1 through L5) participated in this study. Table 1 lists the participants, ordered by their Seneca proficiency, from advanced to novice. All participants were adults between 18 and 50 years of age.

Table 1. Study participants with IDs and description of Seneca language level

ID	Seneca Language Level
S1	Advanced Seneca language apprentice
S2	Advanced Seneca language apprentice
S3	Advanced Seneca language apprentice
S4	Intermediate Seneca language apprentice
S5	Intermediate Seneca language apprentice
L1	Linguistics PhD student with some knowledge of Seneca and other Iroquois languages
L2	Linguistics undergrad with some knowledge of Seneca
L3	Linguistics undergrad with some knowledge of Seneca
L4	Linguistics master's degree with some prior exposure to Seneca
L5	Linguistics undergrad with some prior exposure to Oneida, another Iroquois language

3.2 ASR framework Our ASR system is built using the open-source Kaldi (Povey et al. 2011) ASR toolkit. We begin with the basic Kaldi tutorial recipe², which uses as features the usual 13-dimensional cepstral mean-variance normalized MFCCs, plus their first and second derivatives. The recipe² was extended to apply LDA transformation and Maximum Likelihood Linear Transform to the features. Other training techniques included boosted Maximum Mutual Information (bMMI) and Minimum Phone Error (MPE). Both bMMI and MPE were trained over 4 iterations, and bMMI used a boost weight of 0.5. The language model used for decoding was a word-level trigram model built with KenLM (Heafield 2011).

The acoustic model was trained on approximately 270 minutes of recorded speech transcribed orthographically at the word level with utterance boundary time stamps. The recordings were produced by three women and four men, all first-language Seneca speakers and over the age of 60. The bulk of the data was transcribed by young adult Seneca learners and produced by two elders of the community who

² https://kaldi-asr.org/doc/kaldi_for_dummies.html

gave IRB-approved verbal consent to participate in this data collection project. Other recordings were captured as part of earlier documentation efforts, in particular those overseen by the Wallace Chafe, author of a linguistic grammar of the Seneca language (Chafe 2014). These recordings were made under a variety of conditions with varying equipment over several decades. The small number of early recordings on magnetic tape, which consist of folktales and personal narratives, were digitized as 16-bit 16kHz WAV files approximately ten years ago. The majority of the recordings, which include both conversations and personal narratives, were captured as 16-bit 44.1kHz WAV files within the last five years with a hand-held recorder or a smartphone using the built-in microphone. All recordings were made in casual settings, such as homes and community centers. All files were converted to mono and downsampled, where necessary, to 16kHz to be compatible with the expected format of audio training files in the Kaldi ASR toolkit.

The language model was trained on 1843 utterances, including the transcriptions of the above recordings and a small number of texts gathered during previous documentation efforts for which there are no corresponding recordings. The lexicon extracted from this text includes 3498 unique words. This ASR system yielded a WER of 50.7 and a character error rate of 31.0 on the excerpt used for this study. Note that this is a very high WER. State-of-the-art ASR systems for English, trained with thousands of hours of audio and millions of words of text, yield word error rates of around 5%, comparable to human transcription performed by skilled transcriptionists (Lippmann 1997).

3.3 Transcription conditions Two approximately 45-second audio clips were chosen from a previously untranscribed audio recording of Seneca elder Sandra Dowdy, who founded the Seneca language immersion program for young children. Information about the two audio clips can be found in Table 2.³

Table 2. Characteristics of the two 45-second audio clips used for testing: number of words, number of characters, word error rate (WER), and character error rate (CER) where applicable.

	# words	# chars	WER	CER
Unassisted audio	70	402	N/A	N/A
ASR-assisted audio	73	453	50.7	31.0

³The recordings and transcripts used in this study are available on our project website at <http://cs.bc.edu/~prudhome/ADEL/products.html>. In accordance with the preferences of the Seneca elders, a subset of the data used to train our ASR models will be archived with the Native American Languages collection at the Sam Noble Museum at the University of Oklahoma.

One audio clip was designated to be transcribed by the participants orthographically in an unassisted fashion (i.e., without the use of ASR) using the software of their preference (generally Praat). The second audio clip was passed through the above-described ASR system. The ASR output, with hypothesized words and with hypothesized utterance boundary time stamps, was processed to create a TextGrid file. Participants opened the audio clip and TextGrid file together in Praat and then corrected the ASR output to produce an orthographic word-level transcript. All participants were experienced Praat users, except for one Seneca community member who was given a brief in-person training before completing the tasks.

Participants were first assigned to complete the unassisted transcription task under timed conditions. After a break, the participants completed the ASR-assisted transcription task under timed conditions. Participants self-reported the time required to transcribe each audio clip. In addition, the word error rate (WER) and character error rate (CER) of each transcription was calculated, using as a reference a careful manual transcription produced by a near-fluent L2 Seneca speaker who did not participate in the study. After completing both transcription tasks, participants completed a brief survey in which they reflected on their experience transcribing with and without the assistance of ASR.

4. Results

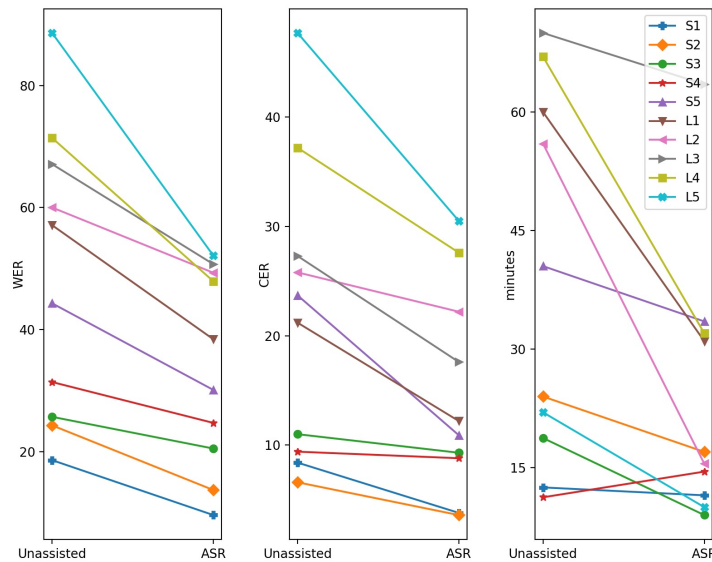


Figure 1. From L to R: word error rate (WER), character error rate (CER), and time to transcribe in minutes for each participant (S1–S5, L1–L5) under the two transcription conditions.

4.1 Time-to-transcribe and error rates As shown in the leftmost panel of Figure 1, transcription word error rates (WER) were reduced among all participants when correcting ASR output. On average, WER decreased by 15.2 (range: 5.2 to 36.5) with ASR-assisted transcription, with a mean 31.4% relative reduction. Linguists decreased their WER by 21.2 (29.8% relative) on average, and Seneca community members decreased their WER by an average of 9.1 (33.1% relative). The reduction in WER across all participants was significant as determined by a paired t-test ($t=5.12$, $p<0.001$, Cohen's $d=0.79$).

The center panel of Figure 1 shows that character error rates (CER) were reduced for all participants when performing ASR-assisted transcription. On average, CER was reduced by 7.2 (range: 0.6 to 17.2) using ASR-assisted transcription, with a mean 33% relative reduction. Linguists decreased their CER by 9.8 (30.8% relative) on average, while Seneca community members decreased their CER by 4.5 (35.2% relative). The reduction in CER across all participants was significant as determined by a paired t-test ($t=4.25$, $p<0.01$, Cohen's $d=0.65$).

Other than the one participant who had not previously used Praat (S4), all participants required less time when correcting ASR output than when transcribing unassisted, as shown in the rightmost panel of Figure 1. Transcription time was reduced by an average of 14.5 minutes (range: -3.25 to 40.5 minutes), with an average 31.4% relative reduction in time required to transcribe. All who had previously used Praat benefited from using ASR output to produce transcriptions, with linguists decreasing their times by 24.6 minutes (47.3%) on average and community members decreasing their times by an average of 4.3 minutes (15.5%). Time to transcribe was significantly reduced ($t=3.06$, $p<0.05$, Cohen's $d=0.75$) under the ASR-assisted condition.

As noted above, the WER rate of the ASR system used was 50.7. All participants but one – the linguistics student with the least knowledge of Seneca – achieved a WER at or lower than this baseline. Although it is concerning that this linguist introduced new word errors (a phenomenon that Sperber et al. (2016) also observed in ASR post-editing), all participants, including this linguist, achieved a CER lower than the ASR baseline of 31. Notably, one linguist whose WER remained at the baseline saw a dramatic improvement over the CER baseline (31 to 17.6), indicating that she identified the ASR word errors and made mostly appropriate corrections.

We found a large negative ($r=-0.68$), though non-significant, correlation between the reduction in time to transcribe and the reduction in both WER and CER among the linguists but not among the community members, suggesting that linguists who took their time correcting the ASR output generally produced higher quality transcriptions. We note that while WER and CER were nearly universally higher for linguists than community members, the most experienced linguist, when transcribing with ASR, achieved a WER and CER within the range of the Seneca community members when transcribing unassisted. This bodes well for the use of ASR for documentation carried out by field linguists with sufficient background in the target language.

4.1 Participant perceptions and preferences After completing the transcription tasks, each participant responded verbally or in writing to two open-ended questions about their experience: *Which transcription method was easier, and why?* and *Which transcription method did you prefer, and why?* Four of the five linguists found correcting ASR output to be easier and preferred ASR-assisted transcription to unassisted transcription. Although all but one of the Seneca community members preferred transcribing unassisted, three of the five found ASR-assisted to be the easier transcription method. Tables 3 and 4 provide selected excerpts from the participants' responses describing their observations about ease of use and their preferences.

Table 3. Excerpts from the written and verbal feedback from participants (S=Seneca, L=Linguist) describing why they found one method of transcription easier than the other

ID	Response to <i>Which transcription method was easier, and why?</i>
S1	<i>With the ASR it's a lot easier to hear the words that are being said because it's written right there in front of you.</i>
S2	<i>The ASR did a good job with the small words which reduced the amount of typing.</i>
S3	<i>With the ASR program it was easier because then I could just plainly listen to the recording and go based off of what I understood and fix the errors the ASR had output.</i>
S4	<i>Praat is not always easy to use and has a steep learning curve.</i>
S5	<i>Although the time with the ASR was shorter, I felt like I had to go through the ASR and double-check the output it provided.</i>
L1	<i>ASR took care of some of the necessary steps, e.g., segmenting speech into utterances.</i>
L2	<i>I felt as though I was much more capable of "transcribing" with less errors with the help of the ASR than just on my own (which felt like a more daunting task).</i>
L3	<i>I had a difficult time with the ASR.</i>
L4	<i>Transcription from ASR was much easier for someone like me with very little knowledge of the morphology.</i>

Table 4. Excerpts from the written and verbal response from participants (S=Seneca, L=Linguist) describing why they preferred one transcription method over the other

ID	Response to <i>Which transcription method did you prefer, and why?</i>
S1	<i>I would prefer to do the transcribing on my own – it's more of a challenging task and really tests my knowledge and understanding of the language.</i>
S2	<i>In some ways I preferred using ASR since I was surprised at how much faster it was to correct the ASR.</i>
S3	<i>I personally prefer transcribing from scratch, I like being able to listen and understand the conversation or story.</i>
S5	<i>As a second language Seneca learner, I feel it is more beneficial for me at the moment to “gain a good ear” for learning the writing system and honing my skills as a listener and writer.</i>
L1	<i>I preferred ASR because I can be more certain when I'm comparing than when I'm perceiving.</i>
L2	<i>Using ASR, I was able to focus on comparing the audio to the transcription rather than trying to perceive what was being said.</i>
L3	<i>I spent more time cross-checking the transcription than actually just transcribing so I preferred transcribing from scratch.</i>
L4	<i>I preferred correcting ASR. It very accurately transcribed particles and discourse markers which I regularly misidentified as parts of other words when I did the unassisted transcription.</i>

5. Discussion and future work Our findings suggest that editing and correcting ASR output, even when the output has a very large number of errors, is significantly faster and more accurate than transcribing from scratch for language learners of all levels. Linguists, particularly those with little prior knowledge of the language, generally preferred ASR-assisted transcription and found it easier. While the majority of Seneca community members acknowledged that correcting ASR output was faster and required less effort, all but one stated a clear preference for transcribing without the use of ASR.

Recently, our team began the arduous process of using OCR to digitize and correct scans of typed and hand-written Seneca texts collected in pre-digital times by linguists and missionaries of the 19th and 20th centuries. These texts were produced without the benefit of technology, requiring hundreds or perhaps thousands of hours of careful manual work, and it will take many more hours to complete our digitization project. Although most speech transcription for language documentation today is carried out natively on computers, often with helpful tools such as Praat or ELAN, speech transcription by language learners, whether linguists or language community

members, continues to be a time-consuming task. While technologies like automatic speech recognition have the potential to transform this process by reducing the time required to transcribe and the number of transcription errors, computer science researchers hoping to deploy these systems must remember that accuracy and speed are but two of many possible metrics for evaluating the utility of such systems for this task.

Community-driven language documentation efforts often have objectives beyond simply producing as much text as possible in the least amount of time (Czaykowska-Higgins 2009). This is particularly true when these efforts are part of a larger language revitalization project in which language learners are tasked with transcription in order to improve their own language skills. One of our Seneca team members noted that although ASR-assisted transcription did seem faster and easier, “if a person is dedicated enough to struggle through it...I believe [transcribing without assistance] will make that person a stronger speaker in the end”.

Unassisted transcription can also offer community members the opportunity to engage more deeply with the recordings they are transcribing as heritage artifacts. One Seneca apprentice mentioned that her preference for unassisted transcription was dependent on the content of the speech: “I know we have recordings of children’s short stories translated into Seneca. For those I would lean more on the ASR output. For the ceremonial kinds of recordings, I would prefer to listen and transcribe from scratch. That way I’m fully absorbing the meaning.”

In the time since we carried out this experiment, a total of 12 hours of Seneca recordings have been collected and transcribed. In addition, we have explored a variety of deep neural ASR architectures, including the available neural models in Kaldi, DeepSpeech (Amodei et al. 2016), wav2vec (Schneider et al. 2019), as well as our own fully convolutional architecture (Thai et al. 2020), both with and without transfer learning and data augmentation. Our best system now yields word error rates close to 25%. In our future work, we plan to carry out a second timed transcription study in order to compare the utility of high- vs. low-WER ASR output for technology-assisted language documentation. Given the already low baseline transcription WER of many of the Seneca language apprentices who participated in our study, we do not necessarily anticipate large gains in this group, but we expect that more accurate ASR output will further improve the transcription speed and accuracy for linguists. We will also attempt to recruit additional linguists with knowledge of Seneca and other Iroquois languages to participate in our future study since our results suggest that linguists are more likely than language community members to choose to use ASR for transcription. By recognizing the strengths and preferences of the full range of stakeholders engaged in endangered language documentation, we can help to shape the design of our own future technology-supported documentation and transcription projects as well as those of other research groups and language communities.

References

- Adams, Oliver, Trevor Cohn, Graham Neubig, & Alexis Michaud. 2017. Phonemic transcription of low-resource tonal languages. In *Proceedings of the Australasian Language Technology Association Workshop*, Brisbane, 6–8 December, 53–60. (<https://www.aclweb.org/anthology/U17-1006/>)
- Adams, Oliver, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, & Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, 7–12 May, 3356–3365. (<https://www.aclweb.org/anthology/L18-1530/>)
- Adams, Oliver, Matthew Wiesner, Shinji Watanabe, & David Yarowsky. 2019. Massively multilingual adversarial speech recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, 2–7 June, 96–108. (<https://www.aclweb.org/anthology/N19-1009/>)
- Akita, Yuya, Masato Mimura, & Tatsuya Kawahara. 2009. Automatic transcription system for meetings of the Japanese National Congress. In *Proceedings of the Annual Conference of the International Speech Communication Association*, Brighton, UK, 6–10 September, 84–87. (<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.159.8637&rep=rep1&type=pdf>)
- Amodei, Dario, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, et al. 2016. DeepSpeech 2: End-to-end speech recognition in English and Mandarin. In *Proceedings of the 33rd International Conference on Machine Learning*, New York, 19–24 June, 173–182. (<https://proceedings.mlr.press/v48/amodei16.pdf>)
- Bansal, Sameer, Herman Kamper, Karen Livescu, Adam Lopez, & Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, 2–7 June, 58–68. (<https://www.aclweb.org/anthology/N19-1006.pdf>)
- Bazillon, Thierry, Yannick Estève, & Daniel Luzzati. 2008. Manual vs. assisted transcription of prepared and spontaneous speech. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, Marrakech, 26 May – 1 June, 1067–1071. (http://www.lrec-conf.org/proceedings/lrec2008/pdf/277_paper.pdf)
- Besacier, Laurent, Etienne Barnard, Alexey Karpov, & Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication* 56. 85–100. (<https://www.sciencedirect.com/science/article/abs/pii/S0167639313000988>)

- Blokland, Rogier, Marina Fedina, Ciprian Gerstenberger, Niko Partanen, Michael Rießler, & Joshua Wilbur. 2015. Language documentation meets language technology. In *Proceedings of the First International Workshop on Computational Linguistics for Uralic Languages*, Tromsø, Norway, 16 January, 8–18. (<https://www.diva-portal.org/smash/get/diva2:898279/ATTACHMENT01>)
- Borowitz, Stephen M. 2001. Computer-based speech recognition as an alternative to medical transcription. *Journal of the American Medical Informatics Association* 8(1). 101–102. (<https://academic.oup.com/jamia/article/8/1/101/722991?login=true>)
- Canavan, Alexandra, David Graff, & George Zipperlen. 1997. *CALLHOME American English speech LDC97S42*. Philadelphia: Linguistic Data Consortium. doi:10.35111/exq3-x930
- Canavan, Alexandra & George Zipperlen. 1996. *CALLHOME Mandarin Chinese speech LDC96S34*. Philadelphia: Linguistic Data Consortium. doi:10.35111/at8b-ct20
- Canavan, Alexandra, George Zipperlen, & David Graff. 1997. *CALLHOME Egyptian Arabic speech LDC97S45*. Philadelphia: Linguistic Data Consortium. doi:10.35111/d8yb-9m13
- Čavar, Małgorzata E., Damir Cavar, & Hilaria Cruz. 2016. Endangered language documentation: Bootstrapping a Chatino speech corpus, forced aligner, ASR. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Portorož, Slovenia, 23–28 May, 4004–4011. (<https://www.aclweb.org/anthology/L16-1632.pdf>)
- Chafe, Wallace. 2007. *Handbook of the Seneca language* (New York State Museum and Science Service, Bulletin Number 388). Albany, NY: The University of the State of New York.
- Chafe, Wallace. 2014. *A grammar of the Seneca language* (University of California Publications in Linguistics 149). Oakland, CA: University of California Press.
- Chelba, Ciprian, Dan Bikel, Maria Shugrina, Patrick Nguyen, & Shankar Kumar. 2012. *Large scale language modeling in automatic speech recognition*. arXiv:1210.8440. (<https://arxiv.org/pdf/1210.8440>)
- Cieri, Christopher, David Miller, & Kevin Walker. 2004. The Fisher Corpus: A resource for the next generations of speech-to-text. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, Lisbon, 26–28 May, 69–71. (<http://www.lrec-conf.org/proceedings/lrec2004/pdf/767.pdf>)
- Coto-Solano, Rolando & Sofía Flores Solórzano. 2017. Comparison of two forced alignment systems for aligning Bribri speech. *CLEI Electronic Journal* 20(1). 2:1–2:13. (<http://www2.clei.org/cleiej/papers/v20i1p2.pdf>)
- Czaykowska-Higgins, Ewa. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian Indigenous communities. *Language Documentation & Conservation* 3(1). 15–50. (<https://scholarpace.manoa.hawaii.edu/bitstream/10125/4423/czaykowskahiggins.pdf>)

- David, Gary C., Angela Cora Garcia, Anne Warfield Rawls, & Donald Chand. 2009. Listening to what is said – transcribing what is heard: The impact of speech recognition technology on the practice of medical transcription. *Sociology of Health & Illness* 31(6). 924–938. (<https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9566.2009.01186.x>)
- DiCanio, Christian T., Hosung Nam, Douglas H. Whalen, H. Timothy Bunnell, Jonathan D. Amith, & Rey Castillo Garcia. 2012. Assessing agreement level between forced alignment models with data from endangered language documentation corpora. In *Proceedings of the Annual Conference of the International Speech Communication Association*, Portland, Oregon, 9–13 September, 130–133. (https://www.isca-speech.org/archive/archive_papers/interspeech_2012/i12_0130.pdf)
- DiCanio, Christian T., Hosung Nam, Douglas H. Whalen, H. Timothy Bunnell, Jonathan D. Amith, & Rey Castillo García. 2013. Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *The Journal of the Acoustical Society of America* 134(3). 2235–2246. (<https://asa.scitation.org/doi/10.1121/1.4816491>)
- Eberhard, David M., Gary F. Simons, & Charles D. Fennig. 2021. *Ethnologue: Languages of the world*. Dallas: SIL International. (<http://www.ethnologue.com>)
- Foley, Ben, Joshua T. Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, & Janet Wiles. 2018. Building speech recognition systems for language documentation: The CoEDL endangered language pipeline and inference system (ELPIS). In *Proceedings of the Sixth International Workshop on Spoken Language Technology for Under-resourced Languages*, Gurugram, India, 29–31 August, 205–209. (https://www.isca-speech.org/archive/SLTU_2018/abstracts/Ben.html)
- Foley, Ben, Alina Rakhi, Nicholas Lambourne, Nicholas Buckeridge, & Janet Wiles. 2019. Elpis, an accessible speech-to-text tool. In *Proceedings of the Annual Conference of the International Speech Communication Association*, Graz, 15–19 September, 4624–4625. (https://www.isca-speech.org/archive/Interspeech_2019/pdfs/8006.pdf)
- Gauthier, Elodie, Laurent Besacier, Sylvie Voisin, Michael Melese, & Uriel Pascal Elingui. 2016. Collecting resources in sub-Saharan African languages for automatic speech recognition: A case study of Wolof. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Portorož, Slovenia, 23–28 May, 3863–3867. (<https://www.aclweb.org/anthology/L16-1611.pdf>)
- Gelas, Hadrien, Laurent Besacier, & François Pellegrino. 2012. Developments of Swahili resources for an automatic speech recognition system. In *Proceedings of the Conference on Spoken Language Technologies for Under-Resourced Languages*, Cape Town, 7–9 May, 94–101. (https://www.isca-speech.org/archive/sltu_2012/papers/su12_094.pdf)

- Gessler, Luke. 2019. Developing without developers: Choosing labor-saving tools for language documentation apps. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Honolulu, 26–27 February, 6–13. (<https://www.aclweb.org/anthology/W19-6002.pdf>)
- Glenn, Meghan Lammie, Stephanie M. Strassel, Haejoong Lee, Kazuaki Maeda, Ramez Zakhary, & Xuansong Li. 2010. Transcription methods for consistency, volume and efficiency. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, Malta, 17–23 May, 2915–2920. (http://www.lrec-conf.org/proceedings/lrec2010/pdf/849_Paper.pdf)
- Godfrey, John J., Edward C. Holliman, & Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, San Francisco, 23–26 March, 517–520. (<https://ieeexplore.ieee.org/abstract/document/225858/>)
- Goss, Foster R., Suzanne V. Blackley, Carlos A. Ortega, Leigh T. Kowalski, Adam B. Landman, Chen-Tan Lin, Marie Meteer, Samantha Bakes, Stephen C. Gradwohl, David W. Bates, & Li Zhou. 2019. A clinician survey of using speech recognition for clinical documentation in the electronic health record. *International Journal of Medical Informatics* 130. 103938. (<https://www.sciencedirect.com/science/article/abs/pii/S1386505619304733>)
- Han, Kyu J., Seongjun Hahm, Byung-Hak Kim, Jungsuk Kim, & Ian Lane. 2017. Deep learning-based telephony speech recognition in the wild. In *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association*, Stockholm, 20–24 August, 1323–1327. (https://www.isca-speech.org/archive/Interspeech_2017/pdfs/1695.PDF)
- Heafield, Kenneth. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, 30–31 July, 187–197. (<https://www.aclweb.org/anthology/W11-2123.pdf>)
- Himmelman, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–196.
- Himmelman, Nikolaus P. 2018. Meeting the transcription challenge. In McDonnell, Bradley, Andrea L. Berez-Kroeker, & Gary Holten (eds.), *Reflections on language documentation 20 years after Himmelman 1998* (Language Documentation and Conservation Special Publication 15), 33–40. Honolulu: University of Hawai'i Press. (<https://scholarspace.manoa.hawaii.edu/bitstream/10125/24806/ldc-sp15-himmelman.pdf>)
- Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, & Brian Kingsbury. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine* 29(6). 82–97. (<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6296526>)

- Hodgson, Tobias & Enrico Coiera. 2016. Risks and benefits of speech recognition for clinical documentation: a systematic review. *Journal of the American Medical Informatics Association* 23. e169–e179. (<https://academic.oup.com/jamia/article-abstract/23/e1/e169/2379888>)
- Jones, Caroline, Weicong Li, Andre Almeida, & Amit German. 2019. Evaluating cross-linguistic forced alignment of conversational data in North Australian Kriol, an under-resourced language. *Language Documentation & Conservation* 13. 281–299. (<https://scholarspace.manoa.hawaii.edu/handle/10125/24869>)
- Juan, Sarah Samson, Laurent Besacier, Benjamin Lecouteux, & Mohamed Dyab. 2015. Using resources from a closely-related language to develop ASR for a very under-resourced language: A case study for Iban. In *Proceedings of the Annual Conference of the International Speech Communication Association*, Dresden, 6–10 September, 1270–1274. (<https://ieeexplore.ieee.org/document/7051423>)
- Koenig, Jean-Pierre & Karin Michelson. 2010. Argument structure of Oneida kinship terms. *International Journal of American Linguistics* 76(2). 169–205. (<https://www.journals.uchicago.edu/doi/abs/10.1086/652265?journalCode=ijal>)
- Koenig, Jean-Pierre & Karin Michelson. 2014. Deconstructing syntax. In *Proceedings of the 21st International Conference on Head-Driven Phrase Structure Grammar*, Buffalo, 27–29 August, 114–134. (<http://web.stanford.edu/group/csli-publications/csli-publications/HPSG/2014/hpsg2014.pdf#page=114>)
- Krauss, Michael. 1992. The world's languages in crisis. *Language* 68(1). 4–10. doi:10.2307/416368
- Kurimo, Mikko, Seppo Enarvi, Ottokar Tilk, Matti Varjokallio, André Mansikkaniemi, & Tanel Alumäe. 2017. Modeling under-resourced languages for speech recognition. *Language Resources & Evaluation* 51(4). 961–987. (<https://link.springer.com/article/10.1007/s10579-016-9336-9>)
- Ladefoged, Peter. 1992. Another view of endangered languages. *Language* 68(4). 809–811. (<https://www.jstor.org/stable/416854>)
- Laurent, Antoinem Sylvain Meignier, Teva Merlin, & Paul Deleglise. 2011. Computer-assisted transcription of speech based on confusion network reordering. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Prague, 22–27 May, 4884–4887. (<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5947450>)
- Lippmann, Richard P. 1997. Speech recognition by machines and humans. *Speech Communication* 22(1). 1–15. (<https://www.sciencedirect.com/science/article/pii/S0167639397000216>)
- Metze, Florian, Ankur Gandhe, Yajie Miao, Zaid Sheikh, Yun Wang, Di Xu, Hao Zhang, Jungsuk Kim, Ian Lane, Won Kyum Lee, Sebastian Stücker, & Markus Müller. 2015. Semi-supervised training in low-resource ASR and KWS. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, 19–24 April, 4699–4703. (<https://ieeexplore.ieee.org/document/7178862>)


- Michaud, Alexis, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, & Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. *Language Documentation & Conservation* 12. 393–429. (<https://scholarspace.manoa.hawaii.edu/bitstream/10125/24793/michaud.pdf>)
- Mitra, Vikramjit, Andreas Kathol, Jonathan D. Amith, & Rey Castillo García. 2016. Automatic speech transcription for low-resource languages—the case of Yo-*loxóchtl* Mixtec (Mexico). In *Proceedings of the Annual Conference of the International Speech Communication Association*, San Francisco, 8–12 September, 3076–3080.
- Mohr, David N., David W. Turner, Gregory R. Pond, Joseph S. Kamath, Cathy B. De Vos, & Paul C. Carpenter. 2003. Speech recognition as a transcription aid: a randomized comparison with standard transcription. *Journal of the American Medical Informatics Association* 10(1). 85–93. (<https://academic.oup.com/jamia/article-abstract/10/1/85/816884>)
- Moseley, Christopher. 2010. *Atlas of the world's languages in danger*, 3rd edn. Paris: UNESCO Publishing.
- Mufwene, Salikoko S. 2017. Language vitality: The weak theoretical underpinnings of what can be an exciting research area. *Language* 93(4). e202–e223. (<https://muse.jhu.edu/article/680465>)
- Nanjo, Hiroaki & Tatsuya Kawahara. 2006. Towards an efficient archive of spontaneous speech: Design of computer-assisted speech transcription system. *The Journal of the Acoustical Society of America* 120. 3042. (<https://asa.scitation.org/doi/pdf/10.1121/1.4787220>)
- Nguyen, Quoc Bao, Jonas Gehring, Markus Müller, Sebastian Stükerk, & Alex Waibel. 2014. Multilingual shifting deep bottleneck features for low-resource ASR. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Florence, 4–9 May, 5607–5611. (<https://ieeexplore.ieee.org/document/6854676>)
- Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, & Karel Veselý. 2011. *The Kaldi speech recognition toolkit*. (Paper presented at the IEEE Workshop on Automatic Speech Recognition and Understanding, Big Island, HI, 11–15 December.) (http://www.daniel-povey.com/files/2011_asru_kaldi.pdf)
- Rasipuram, Ramya & Mathew Magimai-Doss. 2015. Acoustic and lexical resource constrained ASR using language-independent acoustic model and language-dependent probabilistic lexical model. *Speech Communication* 68. 23–40. (<https://www.sciencedirect.com/science/article/abs/pii/S0167639314000995>)
- Reuelta-Martínez, Alejandro, Luis Rodríguez, & Ismael García-Varea. 2012. A computer assisted speech transcription system. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, 23–27 April, 41–45. (<https://www.aclweb.org/anthology/E12-2009.pdf>)

- Rice, Sally. 2018. Reflections on documentary corpora. In McDonnell, Bradley, Andrea L. Berez-Kroeker, & Gary Holten (eds.), *Reflections on language documentation 20 years after Himmelmann 1998* (Language Documentation and Conservation Special Publication 15), 157–172. Honolulu: University of Hawai'i Press.
- Rodríguez, Luis, Francisco Casacuberta, & Enrique Vidal. 2007. Computer assisted transcription of speech. In *Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis*, Girona, Spain, 6–8 June, 241–248. (https://link.springer.com/chapter/10.1007/978-3-540-72847-4_32)
- Rosenthal, Daniel I., Felix S. Chew, Damian E. Dupuy, Susan V. Kattapuram, William E. Palmer, Renee M. Yap, & Leonard A. Levine. 1998. Computer-based speech recognition as a replacement for medical transcription. *American Journal of Roentgenology* 170. 23–25. (<https://www.ajronline.org/doi/pdfplus/10.2214/ajr.170.1.9423591>)
- Saon, George, Gakuto Kurata, Tom Sercu, Kartik Audhkhasi, Samuel Thomas, Dimitrios Dimitriadis, Xiaodong Cui, Bhuvana Ramabhadran, Michael Picheny, Lynn-Li Lim, Bergul Roomi, & Phil Hall. 2017. English conversational telephone speech recognition by humans and machines. In *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association*, Stockholm, 20–24 August, 132–136. (https://www.isca-speech.org/archive/Inter-speech_2017/pdfs/0405.PDF)
- Scharenborg, Odette, Francesco Ciannella, Shruti Palaskar, Alan Black, Florian Metze, Lucas Ondel, & Mark Hasegawa-Johnson. 2017. Building an ASR system for a low-resource language through the adaptation of a high-resource language ASR system: Preliminary results. In *Proceedings of the International Conference on Natural Language, Signal and Speech Processing*, Casablanca, 5–6 December, 26–30. (http://www.isle.illinois.edu/speech_web_lg/pubs/2017/scharenborg17ic-nlssp.pdf)
- Schneider, Steffen, Alexei Baevski, Ronan Collobert, & Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association*, Graz, 15–19 September, 3465–3469. (<https://arxiv.org/pdf/1904.05862>)
- Seifart, Frank, Nicholas Evans, Harald Hammarström, & Stephen C. Levinson. 2018. Language documentation twenty-five years on. *Language* 94(4). e324–e345. (<https://muse.jhu.edu/article/712110/pdf>)
- Simons, Gary F. 2019. Two centuries of spreading language loss. In *Proceedings of the Linguistic Society of America 2019 Annual Meeting*, New York, 5–8 January, 27:1–12. (<http://www.journals.linguisticsociety.org/proceedings/index.php/PLSA/article/viewFile/4532/4126>)
- Sperber, Matthias, Graham Neubig, Satoshi Nakamura, & Alex Waibel. 2016. Optimizing computer-assisted transcription quality with iterative user interfaces. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Portorož, Slovenia, 23–28 May, 1986–1992. (<https://www.aclweb.org/anthology/L16-1314.pdf>)


- Sperber, Matthias, Graham Neubig, Jan Niehues, Satoshi Nakamura, & Alex Waibel. 2017. Transcribing against time. *Speech Communication* 93. 20–30. (<https://www.sciencedirect.com/science/article/pii/S0167639316301972>)
- Stahlberg, Felix, Tim Schlippe, Stephan Vogel, & Tanja Schultz. 2014. Towards automatic speech recognition without pronunciation dictionary, transcribed speech and text resources in the target language using cross-lingual word-to-phoneme alignment. In *Proceedings of the Fourth International Workshop on Spoken Language Technologies for Under-Resourced Languages*, St. Petersburg, 14–16 May, 73–80. (<http://mica.edu.vn/sltu2014/proceedings/SLTU2014Proc.pdf>)
- Stoian, Mihaela C., Sameer Bansal, & Sharon Goldwater. 2020. Analyzing ASR pretraining for low-resource speech-to-text translation. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, 4–8 May, 7909–7913. (<https://ieeexplore.ieee.org/iel7/9040208/9052899/09053847.pdf>)
- Strunk, Jan, Florian Schiel, & Frank Seifart. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using WebMAUS. In *Proceedings of the Language Resources and Evaluation Conference*, Reykjavik, 26–31 May, 3940–3947. (<https://www.aclweb.org/anthology/L14-1123/>)
- Thai, Bao, Robbie Jimerson, Raymond Ptucha, & Emily Prud'hommeaux. 2020. Fully convolutional ASR for less-resourced endangered languages. In *Proceedings of the Workshop on Spoken Language Technologies for Under-resourced Languages*, Marseilles, 11 May, 126–130. (<https://www.aclweb.org/anthology/2020.sltu-1.17.pdf>)
- Thieberger, Nick. 2017. LD&C possibilities for the next decade. *Language Documentation & Conservation* 11. 1–4. (<https://scholarspace.manoa.hawaii.edu/bitstream/10125/24722/thieberger.pdf>)
- Thomas, Samuel, Michael L. Seltzer, Kenneth Church, & Hynek Hermansky. 2013. Deep neural network features and semi-supervised training for low resource speech recognition. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, 26–31 May, 6704–6708. (<https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6638959>)
- Tüske, Zoltán, Pavel Golik, David Nolden, Ralf Schlüter, & Hermann Ney. 2014. Data augmentation, feature combination, and multilingual neural networks to improve ASR and KWS performance for low-resource languages. In *Proceedings of the Annual Conference of the International Speech Communication Association*, Singapore, 14–18 September, 1420–1424. (https://www.isca-speech.org/archive/archive_papers/interspeech_2014/i14_1420.pdf)
- van Esch, Daan, Ben Foley, & Nay San. 2019. Future directions in technological support for language documentation. In *Proceedings of the Third Workshop on Computational Methods for Endangered Languages*, Honolulu, 26–27 February, 14–22. (<https://www.aclweb.org/anthology/W19-6003.pdf>)
- Woodbury, Anthony C. 2003. Defining documentary linguistics. *Language Documentation & Description* 1. 35–51. (<https://cpb-us-w2.wpmucdn.com/voices.uchicago.edu/dist/1/140/files/2013/11/woodbury-defining-documentary-linguistics.pdf>)

Xiong, Wayne, Lingfeng Wu, Fil Alleva, Jasha Droppo, Xuedong Huang, & Andreas Stolcke. 2018. The Microsoft 2017 conversational speech recognition system. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, 15–20 April, 5934–5938. (<https://arxiv.org/abs/1708.06073>)

Emily Prud'hommeaux
prudhome@bc.edu

 orcid.org/0000-0003-3318-892X

Karin Michelson

 orcid.org/0000-0002-8404-6801