Technology and Language Revitalization: A Roadmap for the Mvskoke Language

Julia Mainzinger
University of Washington
jmainz@uw.edu

Abstract

Speaking a language is inherent to maintaining cultural identity and pride for many indigenous peoples of the Americas. For the Mvskoke people, a history of removal from tribal lands and colonialism has accelerated language loss. In recent years, there has been a resurgence of tribal members interested in reclaiming their language. This paper is a discussion of how natural language processing (NLP) can come alongside community efforts to aid in revitalizing the Mvskoke language. Presented here is an overview of available resources in Mvskoke, an exploration of relevant NLP tasks and related work in endangered language contexts, and applications to language revitalization.

1 Introduction

As NLP research matures and computational linguistic technologies enter the commercial realm, low-resource and endangered languages are seeing a greater gap between what is possible for high-resource languages, and what is available for endangered language communities. Work is needed to reverse this trend and include diverse languages in the forefront of language technology.

The technology cannot be an end in itself - rather the goal of developing NLP tools should be to support the ongoing work of the endangered language community. The Mvskoke[1] have several ongoing language revitalization efforts. This paper explores some of the techniques that have been employed in other low-data situations and how they might be helpful for the Mvskoke language.

1.1 Mvskoke Language Reclamation

The Mvskoke tribe has been invested in language reclamation efforts for decades. The Muscogee (Creek) Nation established the Mvskoke Language Program more than 25 years ago to collect and create language documentation and educational resources. The tribal college, the College of the Muscogee Nation (CMN), established in 2004, has a Mvskoke Language Studies certificate program, and more recently is offering a Mvskoke Language Teaching certificate.

Though these efforts have been ongoing, the COVID-19 pandemic was a pivotal time for increasing awareness among the broader Mvskoke community. Home isolation drove people to seek online interaction. Mvskoke speakers formed online groups and began holding Mvskoke-only chats. The CMN moved its classes online, allowing students from diverse locations beyond their service area to attend. These online activities provided better access for displaced tribal members and fostered connection within the tribe. Many people began learning their heritage language for the first time.

Since 2020, new initiatives are inviting long-term involvement from community members. In 2022, the inaugural Mvskoke Language Symposium was held. In fall 2023, the CMN established its master-apprentice program, in which eight committed students are studying the language full-time under three master-level first language teachers for a full academic year.

Some of these efforts could benefit from improved computational infrastructure and Mvskoke language NLP tools. Similar surveys of have been done for the Cherokee (Zhang et al., 2022) and Bodwéwadmimwen (Potawatomi) (Lewis, 2023). The hope is that this paper might increase visibility for the needs of the Mvskoke language revitalization community.

Finally, while this author is a citizen of the Muscogee Nation and has spent time in personal conversation and working groups within the tribe over the last several years, my views are by no means representative of the entire tribe. However,

[1]Following the trend of Mvskoke-led scholarship, I will use the spelling "Mvskoke" from the traditional orthography in all cases except when referring to certain legal entities such as the tribal government and tribal college.

Proceedings of the The Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 7–12

I have received feedback from tribal members and leadership on this paper, and I attempt here to summarize some trends within the community.

2 The Mvskoke Language

2.1 Background

The Mvskoke language (also spelled Muscogee or Muskogee) is a member of the Muskogean family, a group of several languages indigenous to the southeastern United States (Martin and Mauldin, 2000). The language is now spoken by residents of the Muscogee (Creek) Nation and Seminole Nation in Oklahoma, and members of the Seminole tribe of Florida. It is estimated that less than 300 first-language speakers remain, and nearly all are over the age of 60[2]. The loss of language was expedited by removal from tribal lands, residential schools, and U.S. federal policies. My grandparents, who spoke Mvskoke as their first language, were encouraged to raise their children exclusively in English in order to encourage assimilation into the "white" American world. Thus my mother, like most in her generation, grew up without speaking the language. Reversing this trend of language loss will require concerted effort in multiple disciplines.

2.2 Linguistic Description

The Mvskoke language has two writing systems. The traditional spelling is a Latin-based orthography of 20 letters, which was developed in the 1800s. A phonemic system was developed by the linguist Mary R. Haas in the 1930s and is used primarily by linguists (Martin and Mauldin, 2000). Most Mvskoke speakers are familiar with the traditional spelling.

Mvskoke is a subject–object–verb (SOV) language that is agglutinating and synthetic (Martin, 2011; Frye, 2020). Subjects and objects are marked by an affix. Verbs have many prefixes, suffixes, and internal grade changes that indicate person, tense, number, and duration, among other things. For example, the verb *liketv* "to sit" could appear in many forms including:

liketv	to sit
likis	"I have sat down"
likes	"S/he sat down"
kakes	"(Of two) They sit down"
vpokes	"(Of three or more) They sit down"
likepvs	"Have a seat!"
likvranis	"I will sit"
ohliketska	"Are you sitting (up there)?"

3 Mvskoke Language Resources

Mvskoke is a smaller language group than Cherokee, which has a growing NLP community, but has more speakers than Seneca, which is seeing great strides in both NLP research and language revitalization (Liu et al., 2021). We can take cues from other indigenous language communities in developing NLP tools for Mvskoke based on the size and type of resources available. This section contains an overview of available resources.

3.1 Text

Mvskoke has a rich history of language documentation, dating back to the 1730s (Frye, 2020). A few of the more recent documentation efforts are highlighted here. A series of 29 traditional stories was written in 1915 by Earnest Gouge. In the 1930s, Mary R. Haas conducted extensive fieldwork documentation, along with James Hill, who wrote down stories, songs, sermons, letters, and descriptions of Mvskoke cultural practices. Beginning in 1992, Dr. Jack Martin and Margaret Mauldin began preserving much of this linguistic work, and published the Earnest Gouge collection in 2004 and the Haas/Hill collection in 2014 (Gouge et al., 2004; Haas et al., 2015). The majority of these texts contain orthographic transcriptions, phonemic transcriptions, morpheme-by-morpheme glosses, and free translation into English. In 2000, Dr. Martin published a dictionary in print, and the FLEx data was published online in 2023[3] (Martin and Mauldin, 2000). The New Testament was translated in the early 1900s and republished in 2011 (Randall and Randall).

3.2 Audio Recordings

The Gouge stories, a portion of the Haas/Hill texts, as well as other stories and letters have been recorded by dictation, and the entire New Testament has been recorded, totalling about 50 hours of read speech. From 2015-2017, the Seminole Nation's Pumvhakv School conducted video interviews of fluent Seminole and Mvskoke speakers. This has led to a collection of nearly 14 hours of transcribed spontaneous speech in ELAN. Other untranscribed audio data includes a series of radio

[2]This estimate is from personal communication with a member of Ekvn-Yefolecv, a community of Mvskoke people.

[3]https://www.webonary.org/muscogee/

recordings from the 1990s as well as audio lessons and story tellings recorded by the Mvskoke Language Program. In the future, the CMN hopes to build a recording studio to conduct interviews and produce other Mvskoke-language media. These resources can be accessed on the Muskogee (Seminole/Creek) Documentation Project website [4].

3.3 Corpus Development and Archiving

The available resources have not yet been centralized into a corpus ready for NLP. Most of the resources exist in various file formats (Word, PDF, mp3, wav, etc) on hard drives, cloud folders, and websites. As part of my speech experiments, I am developing a labeled speech corpus; more information about the data preparation is in Section 4.2. A structured, searchable corpus would be useful for educators, as mentioned in Section 4.4.

Archival versions of many of the audio recordings are housed at the Sam Noble Museum at the University of Oklahoma. The Mvskoke National Library and Archives houses physical documents and cultural objects, as well as a growing digital collection, with moderated access for community members. In similar fashion, an NLP corpus would need to have appropriate viewing access in accordance with the wishes of the families represented.

3.4 Language Learning Technology

A Muscogee language learning app was published several years ago with some limited vocabulary lists, songs, and quiz games. Since 2020, there has been a marked increase in the effort to make learning materials available online. The Muscogee Nation has been releasing video recordings of online and community classes, and a Mvskoke language learning podcast has been proposed. The dictionary website and mobile app are in the final stages of publishing. In order to facilitate communicating in the language over text, I have built a Mvskoke keyboard with limited predictive text based on a simple unigram language model that is currently undergoing community evaluation[5]. NLP tools can support the growing work of Mvskoke language education and revitalization.

4 NLP Roadmap

The goal of any NLP tool development should contribute to the language community's goals. For the Mvskoke people, this includes empowering language educators, producing new language speakers, leading beginning speakers to fluency, and removing obstacles to sharing knowledge within the language community.

4.1 Morphological Analysis

Morphological analysis, the task of splitting words into morphemes, is helpful to many NLP systems. As an agglutinative and synthetic language in which verbs can have hundreds of forms and words can grow quite long, Mvskoke technology could be improved by morphological parsing. For example, a morphological parser could improve dictionary search by being able to split long queries into morphemes. Generative morphology can be implemented in language learning software, as in the case of Kanyen'k´eha (Mohawk) (Lessard et al., 2018). Furthermore, a morphological parser could be built into the mobile keyboard with predictive text, helping Mvskoke speakers use the language in their daily life.

Current machine learning approaches rely on corpora on the order of millions or more tokens. The amount of text data available for Mvskoke is relatively small in comparison. However, since Mvskoke has a wonderful dictionary and grammar, a rule-base approach may be advantageous over a data-driven approach. Finite-state transducer approaches have seen success in other morphologically complex languages, such as Yup'ik (Strunk and Bender, 2020) and Inuktitut (Farley, 2009). Deep learning techniques are also possible, as seen in the case of Innu-aimun (Le et al., 2022). If deep learning techniques were attempted for Mvskoke, training could be supervised by the interlinear glossed text (IGT) collected in the language documentation, but experimentation is needed to determine if the amount of data is sufficient.

4.2 Speech Recognition

Automatic Speech Recognition (ASR) is the process of using machine learning to produce written text from audio recordings. An ASR model could be used in applications such as speech-to-text input and language learning software. Classically, work in language documentation has viewed ASR as a way to overcome the "transcription bottleneck." Two community members specifically requested assistance with the burden of transcription. However, there are applications for ASR that can go beyond transcription, though transcriptions are im-

[4] https://muskogee.pages.wm.edu

[5] https://github.com/muscogee-language-foundation/muscogee-keyboard

	dev	test
WER	0.27	0.35
CER	0.09	0.06

Table 1: Results of ASR model fine-tuned on 1.12 hours of data of a single speaker.

portant for providing more training data and valuable language documentation. This author agrees with (Bird, 2020) that the goal of speech technology need not be full transcriptions but rather that the diversity of computational methods can facilitate a number of tasks that offer more participatory opportunities for speakers of the language.

One such example application is corpus search, which could allow users to find examples from recordings using an audio query. In low data settings, the high error rates of traditional ASR may render transcriptions unusable. Instead, spoken term detection is an area of research that could be used in place of traditional ASR, to detect isolated terms in a speech collection (Le Ferrand, 2023). Spoken term detection can be accomplished via either ASR with phone recognition, or dynamic time warping (DTW).

I have begun aligning recordings with transcripts at the word and phrase level using the Montreal Forced Aligner (McAuliffe et al., 2017). Only a small portion of the many hours of transcribed audio data is prepared in phrase-transcription pairs. Mvskoke can be easily mapped with a near 1-to-1 grapheme-to-phoneme conversion. Therefore, I am able to map Mvskoke to an English phone set and adapt an English acoustic model to Mvskoke with 2 hours of audio data. Resulting alignments on new transcripts require only minimal correction.

Initial ASR experiments are conducted by fine-tuning the Massively Multilingual Speech (MMS-1B-l1107) model, a pretrained wav2vec 2.0 model (Pratap et al., 2023; Baevski et al., 2020). The model is trained with 1.12 hours of low-noise data from one speaker. Preliminary results show promise, especially considering the lack of a language model. Forced alignment will allow for improvement by providing more data for supervision, and a language model can improve word accuracy. Results are shown in Table 1 and details of experimentation are in Appendix A.

4.3 Text to Speech

Two instructors at the College of the Muscogee Nation expressed interest in training a text-to-speech synthesis (TTS) model to support language learning software. In their view, engaging with the language outside the classroom would help keep students at the college motivated during school breaks. As Mvskoke language learners are usually surrounded by English on a day-to-day basis, time needs to be set aside for listening to the language. Existing recordings can serve as one resource, but being able to generate new examples can provide a more interactive experience. With so few speakers remaining, TTS can relieve some of the burden of recording elders for language learning applications, especially as recordings for posterity generally take priority.

In a low data situations, TTS traditionally requires careful rule-based construction (Koffi and Petzold, 2022; Chasaide et al., 2015). However, a TTS model for Ojibwe has been trained using a neural approach from scratch with about 12 hours of speech, and is deployed into a web-based language learning platorm (Hammerly et al., 2023), paving the way for other indigenous languages to follow suit in using TTS for language revitalization.

4.4 Corpus Search

Because there is no longer intergenerational language transmission for Mvskoke, we are dependent on teachers to continue passing on the language. Therefore NLP technology for Mvskoke needs to be built with language education in mind. There has been an increase in fluent speakers being willing to teach over the last few years, but pedagogy needs to improve in order to move beginning speakers to fluency (Frye, 2020). One way to support teachers is by providing access to a digital corpus. Currently, some resources are available for access through documentation and archival websites, or can be purchased in print form. But for teachers preparing lessons, an organized and searchable corpus would be hugely beneficial.

Teacher in the Loop was proposed in (Neubig et al., 2020) as an interface for educators to have access to textual language documentation resources. In this type of program, a teacher could not only search the corpus, but could also provide feedback during search to improve the system. Yup'ik developers have built a great example of a dictionary website with in-context examples that are linked to interviews and texts.[6]. Building a searchable corpus with these features would require at least

[6] https://www.yugtun.com/

a morphological analyzer, word embeddings, an indexed digital corpus, a search engine, and a user interface, and is therefore a significant undertaking. In the case of Mvskoke, the main obstacles are lack of funding and human resources.

My current experiments with ASR could aid corpus search in that audio results can be returned by matching text queries to automatically transcribed text. Even if word accuracy is low, spoken term detection techniques could be utilized to match queries to most likely examples as mentioned in section 4.2.

5 Ethical Considerations

Due to the painful history of colonization and even recent data misuse, indigenous people are wary of allowing sensitive documents and recordings to be viewed by those outside the tribe. Therefore, there is an active push for tribes to collect and maintain their own language data. Te Hiku Media, an organization active in Māori language revitalization, has written on concerns about the use of indigenous language data in training AI, and has created guidelines for communities to write their own data licenses (Mahelona et al., 2023; Media, 2022). A digital corpus of Mvskoke language materials would need to be treated with care and dignity, and any use of the data to train NLP systems should benefit the tribal community.

6 Conclusion

This survey presents the current state of resource availability and development of NLP tools for the Muscogee language. Also examined are challenges and steps towards the future development of NLP tools and how they can be applied to the goals of language revitalization in the Muscogee community.

Acknowledgements

I am thankful to faculty at the College of the Muscogee Nation, employees of the Muscogee (Creek) Nation, and Mvskoke friends and community members for their helpful comments. Thank you to my advisor, Gina-Anne Levow for her encouragement and guidance. Thank you to Jack Martin for providing resources and support. And I am especially indebted to our tribal elders who have passed on our language and culture through much adversity. *Mvto*.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Steven Bird. 2020. Sparse transcription. *Computational Linguistics*, 46(4):713–744.

Ailbhe Ní Chasaide, Neasa Ní Chiaráin, Harald Berthelsen, Christoph Wendler, and Andrew Murphy. 2015. Speech technology as documentation for endangered language preservation: The case of irish. In *ICPhS*, volume 2015, page 18th.

B. Farley. 2009. The uqailaut project.

Melanie Frye. 2020. Improving mvskoke (creek) language learning outcomes: A frequency-base approach. *Thesis, University of Oklahoma*.

Earnest Gouge, Edited, Translated by Jack B. Martin, and Juanita McGirt. 2004. *Totkv Mocvse / New Fire: Creek Folktales*. Norman: University of Oklahoma Press.

Mary R. Haas, James H. Hill, Jack B. Martin, Margaret McKane Mauldin, and Juanita McGirt. 2015. *Creek (Muskogee) Texts*. University of California Publications.

Christopher Hammerly, Sonja Fougère, Giancarlo Sierra, Scott Parkhill, Harrison Porteous, and Chad Quinn. 2023. A text-to-speech synthesis system for border lakes Ojibwe. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 60–65, Remote. Association for Computational Linguistics.

Ettien Koffi and Mark Petzold. 2022. A tutorial on formant-based speech synthesis for the documentation of critically endangered languages. *Linguistic Portfolios*, 11(3).

Ngoc Tan Le, Antoine Cadotte, Mathieu Boivin, Fatiha Sadat, and Jimena Terraza. 2022. Deep learning-based morphological segmentation for indigenous languages: A study case on innu-aimun. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 146–151.

Eric Le Ferrand. 2023. *Leveraging Speech Recognition for Interactive Transcription in Australian Aboriginal Communities*. Theses, Charles Darwin University.

Greg Lessard, Nathan Brinklow, and Michael Levison. 2018. Natural language generation for polysynthetic languages: Language teaching and learning software for Kanyen'kéha (Mohawk). In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 41–52, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Robert Lewis. 2023. A survey of computational infrastructure to help preserve and revitalize bodwéwadmimwen. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 44–50, Remote. Association for Computational Linguistics.

Zoey Liu, Robert Jimerson, and Emily Prud'hommeaux. 2021. Morphological segmentation for Seneca. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 90–101, Online. Association for Computational Linguistics.

Keoni Mahelona, Gianna Leoni, Suzanne Duncan, and Miles Thompson. 2023. OpenAI's whisper is another case study in colonisation.

Jack B. Martin. 2011. *A Grammar of Creek (Muskogee)*. University of Nebraska Press.

Jack B. Martin and Margaret McKane Mauldin. 2000. *A Dictionary of Creek/Muskogee*. University of Nebraska Press.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Te Hiku Media. 2022. Data sovereignty and the kaitiakitanga license.

Graham Neubig, Shruti Rijhwani, Alexis Palmer, Jordan MacKenzie, Hilaria Cruz, Xinjian Li, Matthew Lee, Aditi Chaudhary, Luke Gessler, Steven Abney, Shirley Anugrah Hayati, Antonios Anastasopoulos, Olga Zamaraeva, Emily Prud'hommeaux, Jennette Child, Sara Child, Rebecca Knowles, Sarah Moeller, Jeffrey Micher, Yiyuan Li, Sydney Zink, Mengzhou Xia, Roshan S Sharma, and Patrick Littell. 2020. A summary of the first workshop on language technology for language documentation and revitalization.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages.

Steve Randall and Monte Randall, editors. *Nak-cokv Mucvsat (The Bible)*. Wiyo Publishing Company.

Lonny Alaskuk Strunk and Emily M. Bender. 2020. A finite-state morphological analyzer for central alaskan yup'ik.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? a case study and roadmap for the Cherokee language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.

A Appendix. ASR Experimentation

Speech data for training and evaluation consists of low-noise recordings of read speech from one female speaker. Other data is significantly more noisy and therefore omitted for this initial experiment. The clarity of recordings likely lowers the error rate.

The data is separated into training and test sets, and the development set is automatically split at 10% of the training set during the preprocessing step. The training set consists of 1,703 utterances, and the test set consists of 141 utterances, with the lengths shown in the table below (h=hour, m=minute, s=second).

	train+dev	test
Total Length	1.12h	10.2m
Average Length	2.3s	4.4s

Implementation follows the steps detailed by Patrick von Platen to fine-tune the MMS adapter using Huggingface Transformers[7] (Wolf et al., 2019). I fine-tune MMS-1B-l1107 (Pratap et al., 2023), a wav2vec model (Baevski et al., 2020). The following parameters are used during training:

Learning rate: 1e-3
Training epochs: 4
Train batch size: 2
Eval batch size: 8
Gradient accumulation steps: 4

Example output:

Predict:	vtok**k**ehatt**e** vtokfenētk**e** vtoy**ē**hattē
Reference:	vtokyehattē vtokfenētkē vtoyehattē

Mvskoke has geminate consonants, in which clusters of like consonants are slightly longer than single consonants. The model correctly identifies the "tt" geminate but mis-identifies a nongeminate consonant "k". The model also has trouble distinguishing a long "ē" from a short "e", which can be difficult even for speakers.

[7] https://huggingface.co/blog/mms_adapters