A Survey of Computational Infrastructure to Help Preserve and Revitalize Bodwéwadmimwen

Robert E. Lewis Jr.
Bodwéwadmimwen Ėthë ték, Inc.
(The Center for Potawatomi Language)
robert.bodwe@gmail.com

Abstract

This paper surveys the previous language initiatives that Potawatomi language communities have undertaken with technology to help them preserve and revitalize Bodwéwadmimwen (the Potawatomi language) and provides a menu of newer computational tools that could supplement previous initiatives or create new ones. The most immediate steps that could be taken are to establish unicode numbers for several characters, compile a usable corpus of language materials, and build a morphological parser that could be implemented on top of an online dictionary. The creation of automatic speech recognition and language models would have to wait for a usable language corpus to be created. Additionally, the utility of automatic speech recognition is largely contingent on whether language community workers perceive it as a crutch or something more. While the focus of this paper is on archival materials, a necessary step to improve any of these computational tools is to continue to learn and document Bodwéwadmimwen.

1 Introduction

Computational linguistics allows one to create useful tools for interacting with technology to preserve and revitalize Bodwéwadmimwen (the Potawatomi language). Several Potawatomi communities sought out these tools relatively early on; however, they have not taken full advantage of the tools computational linguistics has to offer. To do computational linguistics on Bodwéwadmimwen, one needs to know how to program, needs to know about the linguistic principles of Bodwéwadmimwen, and needs linguistic data to do computational linguistics on. This paper has two goals. The first goal is to survey the previous language initiatives that Potawatomi communities have undertaken with technology to help them preserve and revitalize Bodwéwadmimwen. The second goal is to provide a menu of newer computational tools that could supplement previous initiatives as well as describe how new initiatives could be undertaken with new computational tools to preserve and revitalize Bodwéwadmimwen.

2 Bodwéwadmimwen and Technology

Potawatomi is called Bodwéwadmimwen by its speakers. Bodwéwadmimwen was historically spoken around the Great Lakes; but due to United States, Canadian, and Mexican treaties, Potawatomi communities have been split up and removed from the Great Lakes. There are now seven federally recognized Potawatomi Nations in the United States spread from Michigan and Wisconsin to Kansas and Oklahoma. In addition, there are Potawatomi communities and people of Potawatomi heritage in Canada and Mexico. Bodwéwadmimwen is a critically endangered language. It is classified as stage 8a (Moribund) on the Expanded Graded Intergenerational Disruption Scale (EGIDS) (Lewis and Simons, 2010) as there are only five first language speakers who are all elders. There is also a small dedicated group of adult second language learners in every Potawatomi community.

Bodwéwadmimwen is a polysynthetic language, highly inflective, with agglutinative morphology. It is an Algonquian language closely related to Ojibwe; however, Bodwéwadmimwen has had a substantial amount of borrowing from the Fox Branch languages (Sauk, Meskwaki, and Kickapoo). There are six writing systems that have been used by Potawatomi people: the Dot system, the Traditional writing system, the Wisconsin Native American Language Program (WNALP) writing system, the BWAKA nonprofit writing system, the Prairie Band Potawatomi Nation (PBPN) writing system, and the Learners writing system. Of these, the Traditional, WNALP, PBPN, and Learners writing systems are regularly used by language communities. There are several additional writing systems used my missionaries in religious texts.

Bodwéwadmimwen is an extremely low resource language. Language documentation consists of a mix of texts, recordings, and videos. Excluding dictionaries and word lists, documentation includes several religious texts and a grammar from the 1800s (Lykins, 1844; Hoecken, 1844, 1846a,b; Gailland, 1868a,b), linguistic field notes transcribed by Charles F. Hockett in the 1930s and 1940s, Hockett's dissertation and a series of articles in the *International Journal of American Linguistics* (Hockett, 1939, 1948a,b,c,d), and more recently community documentation held by Potawatomi Language and Culture Departments arising out of their own documentation efforts, e.g. texts in the appendix of Lockwood (2017). Some of this documentation is not transcribed and/or translated.

The fact that Bodwéwadmimwen is a low-resource language and the current state of a text and audio corpus will dictate the type of computational tools that can be built for it in the short term.

3 Previous initiatives

3.1 Font maintenance

The Traditional and WNALP writing systems are supported on most platforms. All of the letters and symbols to write in the Traditional writing system are on the standard QWERTY keyboard. Additionally, Languagegeek makes available a keyboard download for the WNALP's writing system that works for Windows and Mac computers: https://www.languagegeek.com/. The only character that the WNALP writing system uses that is not on the QWERTY keyboard is *é*. Note that these are only two of the six writing systems, and at least four of the writing systems are regularly used by community members (i.e. Traditional, WNALP, PBPN, and Learners).

3.2 Online dictionaries

Bodwéwadmimwen was one of the first endangered languages to make use of an online dictionary as chronicled in (Buszard-Welcher, 2001). The website and the Bodwéwadmimwen nonprofit BWAKA were created by Prairie Band Potawatomi member Smokey McKinney to document the Potawatomi of his father and other Potawatomi people. The online dictionary contains more than 1,200 words and is hosted by the Kansas Heritage Group: http://www.kansasheritage.org/PBP/homepage.html. Since this first website, other publicly available online dictionaries have been created. The Citizen Potawatomi Nation (CPN) has an online dictionary with around 10k words: https://www.potawatomiheritage.com/language/. The Pokagon Band of Potawatomi also has an online dictionary with 741 words: https://wiwkwebthegen.com/dictionary. All three of these dictionaries use different writing systems. Smokey McKinney's dictionary uses the BWAKA writing system, the CPN uses the WNALP writing system, and the Pokagon Band use the Learners writing system.

4 Where to go from here?

4.1 Font maintenance

Not all systems support the PBPN and Learners writing systems. The characters *ė*, *ś*, *t̄*, and *ťt* in the PBPN writing system and the characters *ė* and *ê* in the Learners writing system are difficult for community members to write and not supported by many fonts. Since the PBPN and Learners writing systems are also regularly used in speech communities, a necessary step would be to have the special characters given their own unicode numbers.

4.2 Automatic speech recognition

Language documentation that only exists in audio or video format needs to be transcribed. A text transcript allows one to quickly identify what is in video and audio files. Manual transcription, however, is a time consuming process. Just a few minutes of audio can take up to an hour or more to transcribe. With lots of recordings and little staff this could lead to a transcription bottleneck (Seifart et al., 2018; Himmelmann, 2018).

Automatic speech recognition (ASR) software could be used to speed up the creation of text files for the video and audio that do not have transcripts. This technology could exist with a fairly small Bodwéwadmimwen text and audio corpus. Coto-Solano et al. (2022) use a transcript of 5,033 sentences and 3 hrs 57 mins of audio to do automatic speech recognition on Cook Island Māori, and find Kaldi (Povey et al., 2011) and XLSR (Conneau et al., 2020) the most effective models. Even with a relatively small corpus size, it may be difficult to compile about four hours of audio only in Bodwéwadmimwen for training the software on. To the best of my knowledge, outside of community documentation initiatives, there exists few monolingual Bodwéwadmimwen recordings. Most recordings, such as classroom recordings, contain large amounts of code switching.

Additionally, automatic speech recognition would likely only be able to provide a rough transcription. Coto-Solano et al. (2022) finds a word error rate (WER) of 18±2 and a character error rate (CER) of 7.5±0.8 for Kaldi and a WER of 23±2 and a CER of 6.1±0.6 for XLSR after training. The error rates increase when tested on held-out speakers to a WER of 46.4±15.6 and a CER of 14.9±7.2 averaged across the speakers. Thus, automatic speech recognition may be good for a first pass transcription to speed up the work of manual transcription, but it certainly would still be necessary to go back and hand verify the transcript.

Moreover, Cook Island Māori is not as complex as Bodwéwadmimwen in terms of phonotactics and morphology. It follows that these models have less phonotactic and morphological variation to learn for Cook Island Māori than they would for Bodwéwadmimwen. It may be that Bodwéwadmimwen's linguistic complexities would increase the WER and CER for Bodwéwadmimwen. In fact, this is what is found for the Iroquoian language Seneca which is closer to Bodwéwadmimwen in terms of morphological complexity. Prud'hommeaux et al. (2021) finds an average WER of 50.7 and CER of 31 when using Kaldi ASR on Seneca with 270 minutes of audio and 1843 utterances (3498 unique words).

In sum, using ASR technology would increase the size of Bodwéwadmimwen textual documentation while reducing the time to do so, but it would likely require pooling audio from across Potawatomi communities to make it work.

4.3 Corpus

A corpus of Bodwéwadmimwen language material must be created. One cannot do computational linguistics without a corpus of language data.

The workings of a usable corpus already exist based on current documentation of Bodwéwadmimwen. Several narratives have been published from Hockett's field notes (Buszard-Welcher, 2003; Lewis, 2020) and a community based project (Lockwood, 2017).

Additionally, I have begun the work of creating a larger corpus. I have copies of religious texts from the 1800s and Hockett's field notes from the 1930s and 1940s. These documents are handwritten. So far I have typed up Hockett's field notes as text files, but I need to go back through and hand verify the corpus. The text of Hockett's field notes consists of seventy-four narratives and is of roughly 4,913 word types. This number will likely fluctuate as the corpus is edited for accuracy.

Based on my experience with Hockett's field notes and the size of the current documentation of Bodwéwadmimwen, it would take a substantial financial and personnel commitment to make it possible to convert all existing data to a useable corpus. For example, it took me about ten years to type up and edit Hockett's field notes. I had to acquire a substantial amount of knowledge to be able to accurately transcribe and edit Bodwéwadmimwen. If a sizable commitment was made though, it would likely take less than a decade to do the work for the remaining current documentation.

The next step is to process the religious texts. The religious texts are in several different writing systems. They, as well as any other textual materials added to the corpus, would have to be converted into a phonemic writing system. One foreseeable problem is whether a copy of these materials can be found that is sufficiently legible to scan or even transcribe. Smokey McKinney, who has already typed up several books of the Bible on his website, notes that the copy of Gailland's Gospel According to Matthew is illegible in a number of places (http://www.kansasheritage.org/PBP/books/home.html).

The harder, and more time consuming, task is to transcribe documentation that only exists in audio and video format. Unless automatic speech recognition software is leveraged, and works, transcription is likely to be the most time consuming and expensive process of converting existing data to a usable corpus. It is worth noting here that community language workers sometimes prefer to transcribe audio from scratch rather than being assisted by ASR (Prud'hommeaux et al., 2021). To echo Prud'hommeaux et al. (2021)'s findings, there is value in using ASR to speed up transcription, but there is also value in engaging in tasks that will help language learners improve their language knowledge.

Finally, there are gaps in the documentation and description of Bodwéwadmimwen that need to be closed soon before a number of tech tools can be implemented. For example, there is little conversation based documentation as pointed out by Buszard-Welcher (2015). This is likely to skew any language models of Bodwéwadmimwen. While we are a long way off from getting a computer to find an answer to a question in Bodwéwadmimwen, we want to make sure we document these types of language in order to one day create such a computational tool.

4.4 Online dictionaries

There are gaps in the documentation and description of Bodwéwadmimwen at the dictionary level. The Pokagon Band dictionary contains 741 words. Smokey McKinney's online Kansas Potawatomi dictionary contains more than 1,200 words. The Forest County Potawatomi Community's print dictionary contains about 5k words (Forest County Potawatomi Community, 2014), and the CPN online dictionary contains about 10k words. The size of these lexicons are still considerably small compare to the mental lexicon of first language speakers. Materials from Hockett's field notes, religious texts, and untranscribed audio and video files are likely to increase these numbers.

Moreover, Smokey McKinney's online Kansas Potawatomi dictionary is woefully out of date. The late Smokey McKinney pointed this out to me. While the online dictionaries for the CPN and the Pokagon Band of Potawatomi are periodically updated, one cannot search in their dictionary with an inflected Potawatomi word. Technically, this has to do with the website's structure and information retrieval and not the dictionary. A morphological parser is needed on the back-end. That is, to make any online dictionary's website able to search for a definition with an inflected word (e.g. finding the same lexical entry for the inflected verbs *gi-zgebwé* 's/he bite' and *ngi-zgebwé* 'I bite'), stemming and lemmatization needs to be done.

Beyond stemming and lemmatization, dictionaries only provide so much information to learners. They do not provide a breakdown of the verb. This is relevant for Potawatomi because most verbs are made up of three lexical components (an initial, medial, and final). The internal structure of stems could also be added to an online dictionary's website. Loftis and Lewis (2021) is an aid in this process that simply needs to be embedded in the appropriate data structure. The data structure use for the online Ojibwe People's Dictionary is a good example.

Additionally, to the extent that a dictionary is meant to be accessible to language learners from various Potawatomi communities, it would be possible to write a programming script that moves from one writing system to another. A feature that the late Smokey McKinney had shared his desire for with me before he passed on. Recall that not every Potawatomi writing system is phonemic. Therefore, documentation only written in a phonemic writing system would need to be rendered into a phonetic one for certain communities' use. It may be possible to write sound rules that take a phonemic writing system to a phonetic one, but it is up for debate whether the same sound rules are at play for each dialect. There are also other dialect differences that would need to be addressed to make a single dictionary accessible to all communities (e.g. differences in vowel raising/lowering, vowel diphthongization/monothongization, and consonant lenition). Most of these dialect differences hold across the board but some are sure to be optional, exceptional, and perhaps even idiosyncratic.

4.5 Morphological parser

There is not currently a stand alone morphological parser for Bodwéwadmimwen. Fortunately, there has been work done on other Algonquian languages that may speed up the creation of a parser for Bodwéwadmimwen (Meskwaki (Luka, 1996), Plains Cree (Harrigan et al., 2017; Schmirler et al., 2018), Odawa (Bowers et al., 2017), and Michif (Davis et al., 2021)).

Bodwéwadmimwen makes frequent use of vowel syncope (and to a lesser extent palatalization) which can be difficult to write a parser for. Again fortunately, Odawa does too. Perhaps, Bowers et al. (2017) could be easily adapted and implemented for the Bodwéwadmimwen data since Bodwéwadmimwen is closest morphologically and phonologically to Odawa of the Algonquian languages. Additionally, Luka (1996)'s parser could be useful for accounting for the Fox Branch morphology in Bodwéwadmimwen (e.g.

the negative *bwa-*).

Of course, in addition to these two parsers, Potawatomi specific rules would need to be created to fully implement a morphological parser for Bodwéwadmimwen, but the future looks promising. In fact, most of the inflectional paradigms have been documented (Hockett, 1939; Lockwood, 2017), so implementation is only a time consuming matter of creating a corpus and lexical database, writing the code, and testing with a corpus. The inflectional paradigms do need to be separated out into dialect groups though, which heretofore has not been done.

Moreover, it is important to think about how much vocabulary does Potawatomi share with Ojibwe or with the Fox Branch languages when building a morphological parser. Loftis and Lewis (2021) begin to address this question by providing a par-wise comparison of Potawatomi verbal components to those in Ojibwe and the Fox Branch languages (mainly the Meskwaki language). They find initials to be the most different of the three core components of a verb (i.e. initials, medials, and finals) between Potawatomi and Ojibwe. This has important theoretical import as it may result in a Bodwéwadmimwen morphological parser that looks different from an Ojibwe (Odawa) one for handling initials of a verb stem.

To complicate this picture, Ojibwe too has had a substantial amount of borrowing from Cree, so it is not always clear whether Potawatomi or Ojibwe was the one to borrow a word. A comparison of Ojibwe, Cree, and Bodwéwadmimwen morphological parsers may shed light on this question of borrowing.

4.6 Language models

Language models statistically model a well-formed sentence. They can be used for things like spell-check and auto complete on phones and computers. Language models are not out of the question for Bodwéwadmimwen, but they remain off the table unless a sizable corpus is made available. This is because language models require a large amount of data to build a statistical model of a sentence. In other words, a language model must be trained on data. The more data a model is trained on, the better the model is likely to be. The state-of-the-art language models are neural models. Neural models require a large amount of text which Bodwéwadmimwen does not have. Therefore, rule based applications are the way to go in the short term.

It, however, depends on the task one needs the language model for. A relatively small corpus could suffice in some cases. Flor et al. (2019) shows that 5% of a dataset of (20-500K words) is robust enough to build a spell-check tool. The change from 5% of a dataset to 75% only increased the accuracy by 1%. Therefore, if a Bodwéwadmimwen corpus can be built to at least meet that threshold, then tools like a spell-checker are viable.

5 Conclusion

This paper surveyed the previous language initiatives that Potawatomi language communities have undertaken with technology to help them preserve and revitalize the Potawatomi language. The paper also provided a menu of newer computational tools that could supplement previous initiatives as well as described how new initiatives could be undertaken with new computational tools to preserve and revitalize the Potawatomi language.

The lowest hanging fruit would be to explore unicode numbers for difficult characters, continue to compile a usable corpus of language materials, and build a morphological parser that could be implemented on top of an online dictionary. While ASR offers to speed up the work of transcription, it may be spurred by community language workers if it is viewed as a crutch. Additionally, only after all archival materials are put into a usable Bodwéwadmimwen corpus, will it become clear whether there is a sufficient amount data to effectively train language models.

Meanwhile, a necessary step is to continue to learn and document Bodwéwadmimwen, especially conversational language.

References

Dustin Bowers, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2017. A Morphological Parser for Odawa. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–9, Honolulu. Association for Computational Linguistics.

Laura Buszard-Welcher. 2001. Can the web help save my language? In Leanne Hinton and Ken Hale, editors, *The Green Book of Language Revitalization in Practice*, pages 331–348. Academic Press, San Diego, California.

Laura A. Buszard-Welcher. 2003. *Constructional Polysemy and Mental Spaces in Potawatomi Discourse*. Ph.D. thesis, University of California, Berkeley, Berkeley, California.

Laura A. Buszard-Welcher. 2015. Challenges of Editing and Presenting the Corpus of Potawatomi Stories Told by Jim and Alice Spear to Charles Hockett. In David J. Costa, editor, *New Voices for Old Words: Algonquian Oral Literatures*, pages 454–489. University of Nebraska Press.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, abs/2006.13979.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer, and Isaac Feldman. 2022. Development of automatic speech recognition for the documentation of Cook Islands Māori. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3872–3882, Marseille, France. European Language Resources Association.

Fineen Davis, Eddie Antonio Santos, and Heather Souter. 2021. On the computational modelling of Michif verbal morphology. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2631–2636, Online. Association for Computational Linguistics.

Michael Flor, Michael Fried, and Alla Rozovskaya. 2019. A benchmark corpus of English misspellings and a minimally-supervised model for spelling correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 76–86, Florence, Italy. Association for Computational Linguistics.

Forest County Potawatomi Community. 2014. *Ézhebmadzimgek gdebodwéwadmi-zheshmomenan: Potawatomi Dictionary*. Forest County Potawatomi Community, Crandon, Wisconsin.

Maurince Gailland, S.J. 1868a. Grammaire de la langue potêvatémie. In *Laura Buszard-Welcher Papers on the Potawatomi language*. Survey of California and Other Indian Languages, University of California, Berkeley.

Maurince Gailland, S.J. 1868b. *Nemëmiseniükin ipi Nemënigamowinin*. Wrightson & Co., Cincinnati, Ohio.

Atticus G. Harrigan, Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of Plains Cree verbs. *Morphology*, 27(4):565–598.

Nikolaus P. Himmelmann. 2018. Meeting the transcription challenge. *Language Documentation & Conservation*, 15:33–40.

Charles F. Hockett. 1939. *The Potawatomi Language*. Ph.D. thesis, Yale University, New Haven, Connecticut.

Charles F. Hockett. 1948a. Potawatomi I: Phonemics, Morphophonemics, and Morphological Survey. *International Journal of American Linguistics*, 14(1):1–10.

Charles F. Hockett. 1948b. Potawatomi II: Derivation, Personal Prefixes, and Nouns. *International Journal of American Linguistics*, 14(2):63–73.

Charles F. Hockett. 1948c. Potawatomi III: The Verb Complex. *International Journal of American Linguistics*, 14(3):139–149.

Charles F. Hockett. 1948d. Potawatomi IV: Particles and Sample Texts. *International Journal of American Linguistics*, 14(4):213–225.

Christian Hoecken. 1844. *Potewatemi nememissinoikan*. W.J. Mullin, Saint Louis, Missouri.

Christian Hoecken. 1846a. Pewani ipi Potewatemi missinoikan, eyowat nemadjik, catholiqus endjik. In *Laura Buszard-Welcher Papers on the Potawatomi language*. Survey of California and Other Indian Languages, University of California, Berkeley.

Christian Hoecken. 1846b. Potewatemi nememissinoikan ewiyowat nemadjik catholiques endjik. In *Early Publications in American Indian Languages Collection*. Smithsonian Libraries and Archives, Washington DC.

M Paul Lewis and Gary F Simons. 2010. Assessing endangerment: expanding Fishman's GIDS. *Revue roumaine de linguistique (RRL)*, 55(2):103–120.

Robert E. Lewis, Jr. 2020. *Potawatomi Discourse Markers*. Ph.D. thesis, The University of Chicago, Chicago, Illinois.

Hunter T. Lockwood. 2017. *How the Potawatomi Language Lives: A Grammar of Potawatomi*. Ph.D. thesis, University of Wisconsin-Madison, Madison, Wisconsin.

Thomas Loftis and Robert E. Lewis, Jr. 2021. *Building Potawatomi Verbs*. Self-published.

Barbara Luka. 1996. PC-Kimmo for Fox: A Computational Tool for the Morphological Parsing of Fox Texts. In *Papers of the Twenty-Seventh Algonquian Conference*, pages 180–194, Winnipeg. University of Manitoba.

Johnston Lykins. 1844. *The Gospel According to Matthew and the Acts of the Apostles — Oti ere Mnoahemowun kaonuperuk Mrto*. W. C. Buck, Louisville, Kentucky.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, and et al. 2011. The Kaldi Speech Recognition Toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation & Conservation*, 15:491–513.

Katherine Schmirler, Antti Arppe, Trond Trosterud, and Lene Antonsen. 2018. Building a constraint grammar parser for Plains Cree verbs and arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2018. Language documentation 25 years on. *Language*, 94:e324–e345.