Creating Digital Learning and Reference Resources for Southern Michif

Heather Souter, Olivia Sammons
Prairies to Woodlands
Indigenous Revitalization Circle
{hsouter,osammons}@p2wilr.org

David Huggins-Daines
Independent Researcher
dhd@ecolingui.ca

Abstract

Minority and Indigenous languages are often under-documented and under-resourced. Where such resources do exist, particularly in the form of legacy materials, they are often inaccessible to learners and educators involved in revitalization efforts, whether due to the limitations of their original formats or the structure of their contents. Digitizing such resources and making them available on a variety of platforms is one step in overcoming these barriers. This is a major undertaking which requires significant expertise at the intersection of documentary linguistics, computational linguistics, and software development, and must be done while walking alongside speakers and language specialists in the community. We discuss the particular strategies and challenges involved in the development of one such resource, and make recommendations for future projects with a similar goal of mobilizing legacy language resources.

1 Introduction

Michif, ma-laañg-inaan, katawashishin[1] (Heather Souter). Southern Michif (ISO 639-3: `crg`; hereafter "Michif"), is one of three language varieties spoken by the Métis (Bakker, 1997; Sammons, 2019). It is a contact language combining elements from Algonquian languages—Plains Cree and the Saulteaux dialect of Ojibwe—with Métis French. Michif has traditionally been spoken in small, diasporic communities across western Canada and the northern United States, mainly on the Prairies. Reliable census data regarding the current number of Michif speakers are unavailable, largely due to ambiguity around the use of the label "Michif", However, Southern Michif speakers and community members who are actively involved in community-based language revitalization informally estimate that there are likely fewer than 100 speakers today (Chew et al., 2023). Intergenerational transmission of the language has ceased, and all but one or two mother-tongue speakers are over 70 years of age. Despite growing revitalization activities in Métis communities in western Canada, few print and digital resources based on best practices in lexicography, language documentation, and second language acquisition are available to support those efforts.

The primary aim of this project was to digitize and make accessible an out-of-print Michif dictionary (Laverdure et al., 1983), while also developing local capacity in technologies for Indigenous language documentation and revitalization. With the assistance of Michif first-language speakers, community-based language workers, project partners, and computational linguists, we have developed the Michif Talking Dictionary,[2] a digital spoken version of this important print resource. This dictionary is now available as a progressive web application, adapted to a wide variety of screen sizes, as shown in Figure 1. The application does not require an Internet connection to search and browse once accessed. Its source code, along with the code used to process the text and annotated speech data for the dictionary, is publicly available under an open-source license.[3]

Another major goal of this project was to develop capacity through the training of emerging Métis community linguists, language workers, and scholars in the areas of audio recording, application of speech technologies, and annotation. Between September 2019 and May 2021, one workshop on recording and five workshops on annotation were held in Brandon, Manitoba, Ottawa, Ontario, and online via Zoom.[4]

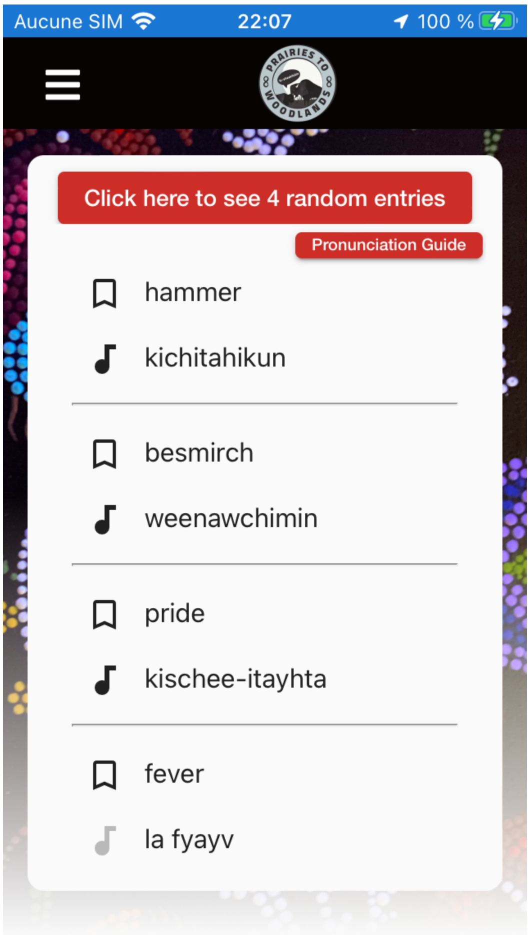
The original book, *The Michif Dictionary: Tur-*

[1] Michif, our language, is beautiful.

[2] https://dictionary.michif.org/

[3] https://github.com/p2wilrc/mtd-michif/

[4] After the outbreak of COVID-19 and resulting restrictions on travel and gathering.



hammer	kichitahikun
besmirch	weenawchimin
pride	kischee-itayhta
fever	la fyayv

Figure 1: Mobile dictionary on iPhone SE

tle Mountain Chippewa Cree, is recognized for its valuable contribution to Michif language documentation. However this now out-of-print resource is largely inaccessible to learners of Michif unless purchased used at a high price, and is rarely available for purchase. While other Michif dictionaries that include audio from native speakers have been published and made available in electronic format (e.g., Rosen et al., 2016; Gabriel Dumont Institute, 2012), both of which are based primarily on Michif as it is spoken in Manitoba and Saskatchewan), this dictionary is exceptional in its degree of coverage of lexical items and example sentences. In addition, many important linguistic studies of Michif (e.g. Bakker, 1997), as well as the lexical resources mentioned above, have relied to varying degrees on the contents of the original Turtle Mountain Dictionary as one of their primary sources. The Turtle Mountain Dictionary is also an important historical resource, as many Métis community members in Canada have kinship ties to Belcourt, ND, where the dictionary was created, and because it includes the speech of an under-represented dialect of Michif. For all of these reasons, multiple Elders and community members identified the creation of an electronic edition of this dictionary as a priority, as it is viewed as a resource that is much too valuable to remain inaccessible, but should rather be put into the hands of Michif language learners and educators.

Permission was granted by Turtle Mountain Community College, the dictionary's copyright holder, to the project team to create a digital version of the dictionary for online, offline, and mobile use. This "new" version retains all of the original content, but will also allow for the inclusion audio recordings of headwords and example sentences, as well as further enrichment in the eventual addition of alternate orthographies and grammatical information for lexical entries.

2 Recording

For the dictionary, 181 hours of high-quality audio recordings were collected from four separate speakers. One speaker, Verna DeMontigny, recorded the entire dictionary from cover to cover, while others recorded selected portions of it. Thus, all entries have been recorded by at least one speaker, with some entries being recorded by two or more speakers.

As shown in Table 1, multiple Michif varieties are represented in these recordings. It was particularly important for the Belcourt, ND variety to be represented here, as the original creators of the dictionary spoke this variety.

All the recordings, backed up regularly on multiple hard drives and on Dropbox, were named according to a consistent file-naming process. Metadata for each session, such as speaker name, location, and covered pages of the dictionary, was tracked and shared among team members via a Google spreadsheet. As we will discuss below, the management of metadata was one of several challenges we faced in the production of the dictionary; for example, the information in this spreadsheet ultimately diverged from that contained in the annotation files. In our discussion of these challenges we hope to identify pitfalls and propose solutions for other groups involved in a similar endeavour. In this case, in the absence of a content management system for the recordings, this problem could have been partially mitigated with the "data validation" feature, similar to the use of controlled vocabularies in ELAN.

Speaker	Michif Variety	Hours Recorded
Verna DeMontigny	The Corner, Manitoba	143h14m34.45
Sandra R. Houle	Belcourt, North Dakota	12h40m47.14
Albert Parisien	Belcourt, North Dakota	15h31m40.16
Connie Henry	Boggy Creek, Manitoba	10h00m00.97
	TOTAL	181h27m02.72

Table 1: Dictionary recording hours by speaker

3 Annotation

All audio recordings were annotated using ELAN (Wittenburg et al., 2006) to produce time-aligned transcripts. First, each recording was segmented into pause-delimited utterances automatically using a Deep Neural Network (DNN) voice activity detection service that was developed within the VESTA-ELAN project by the Centre de Recherche Informatique de Montréal (Gupta and Boulianne, 2022). This auto-segmentation saved an immeasurable amount of time in the annotation process.

To support remote annotators with heterogeneous Internet connections and computer hardware, hosting of the annotations was switched to Google Drive from Dropbox. As per the requirements of earlier versions of ELAN, it was necessary to provide WAV files to visualize waveforms, which were critical for annotators to be able to see and correct the automatic segmentation. However, dissemination of the 'master' WAV files was a challenge, given their large size. To address this, we first down-sampled the original audio from 48 kHz, 24-bit WAV into two different formats: (1) high-quality MP3 files (44.1kHz, 16-bit, 128kbps), which were used for playback; and (2) low-quality WAV files (8kHz, 8-bit), which were provided only for waveform visualization in ELAN, and were never used in playback. This approach made it feasible to share the entire audio collection with annotators over a cloud-based service, enabling them to both listen to high-quality versions of the audio and to display the corresponding waveforms in ELAN. The master recordings were maintained separately and later used as the source of the audio that was included in the dictionary.

The paper dictionary was scanned and converted to text using the Tesseract 4 optical character recognition engine. An ELAN template was created with tiers for English headwords, Michif definitions, and example sentences, and these were then integrated from the OCR text of the dictionary into these transcripts by a team of Indigenous and non-Indigenous language workers who contributed to the project as paid contract employees, volunteers, and, in one case, as a student in a for-credit independent study course in applied linguistics.

In most cases, the speakers recorded multiple instances of each word and example sentence. The annotators were therefore instructed to select the best recording for "export" to the talking dictionary. Due to the slow and careful speaking style used, the example sentences and definitions were frequently split into multiple segments, which had to be reassembled in the construction of the talking dictionary. Annotators were also instructed to adjust the boundaries of these segments to ensure that no words were cut off. In some cases, it was necessary to splice together different instances in order to obtain an audio clip without false starts or mispronunciations.

Because of the dialect variation which exists in Southern Michif, as well as the fact that the recordings were made nearly 40 years after the creation of the print dictionary, the speakers often diverge from the original text, or in some cases, provide a corrected version of a dictionary entry. Annotators were thus instructed to flag partial matches as well as novel forms. In the initial version of the the talking dictionary, we have attempted to remain faithful to the original text as much as possible, with the exception of typos and misspellings. A revised version is in development which will present these variant and corrected forms along with relevant grammatical information.

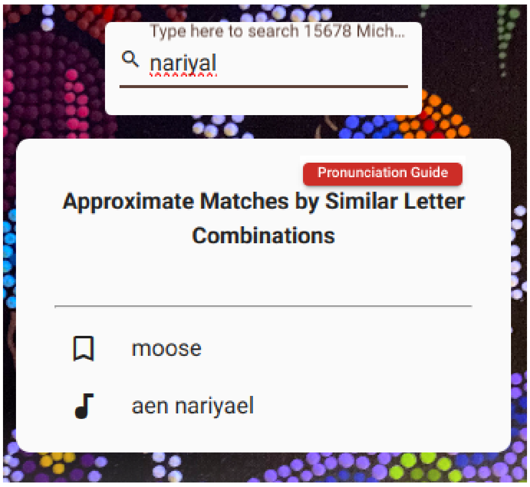
Manual review and corrections of the text of the dictionary was performed by 14 undergraduate students as part of a Community Service-Learning project in an Indigenous Languages of Canada course in winter 2021. Students in this course used Transkribus Lite, a web-based interface to functions of the Transkribus transcription platform (Kahle et al., 2017), to identify and address errors in the computer-readable text of the dictionary that were introduced by the previously applied OCR methods (e.g., correcting misspelled words, entering words or lines that were present on the page but missed by the OCR software, etc.). Errors were found and corrected on a total of 1600 lines of text, or 8.5% of the dictionary. However, there remained a large number of systematic OCR errors, such as ambiguity between `l`, `1`, and `I`, which were corrected semi-automatically in the dictionary build.

Since different parts of the project were conducted simultaneously, technical issues arose from the ordering of this work. For instance, the post-correction of the dictionary text took place *after* the start of the annotation process, resulting in a divergence between the text in the ELAN annotations and the text dictionary. Likewise, while the dictionary entries from the original OCR output were separated into definitions and examples when creating the ELAN files, and sometimes also corrected by the annotators afterwards, these modifications were not synchronized or linked in any way to the dictionary text. Because it was infeasible to correct these discrepancies manually, it was necessary to develop a complex data extraction process using heuristic matching algorithms to align dictionary text and annotations.

4 Dictionary Construction

The electronic dictionary was produced using a customized version of the MotherTongues (Littell et al., 2017) platform. This well-documented open-source tool provides Web and mobile applications with a flexible and configurable approximate search feature, shown in Figure 2, along with a tool to automate the conversion of dictionaries from a variety of formats including spreadsheets, XML, and JSON files. Compared to tools such as FLEx (Beier and Michael, 2022), it supports a very restricted set of lexicographical data, but no such data exists in the original dictionary in any case. This is a common situation for community-developed resources, and the relatively lightweight nature of MotherTongues allows for the creation of dictionaries with a minimum of technical expertise. That said, the absence of grammatical information in the Michif Talking Dictionary limits its usefulness for language learners, and we hope to address this in a subsequent revision.

As detailed in sections 2 and 3, there were four separate sources of information used to produce the talking dictionary:


Type here to search 15678 Mich...

nariyal

Pronunciation Guide

Approximate Matches by Similar Letter Combinations

moose

aen nariyael

Figure 2: Approximate Search

1. The corrected OCR dictionary text.
2. The original recordings.
3. The metadata spreadsheet identifying the speaker, date and location of each recording along with the pages of the dictionary covered and any comments on audio quality.
4. The ELAN files containing speech segments and aligned lexical entries and examples for each recording.

Unfortunately, the need to rapidly organize a distributed annotation effort, turnover of key personnel, and other difficulties arising from the COVID-19 pandemic led to widespread inconsistencies within and between these data sources. The initial version of the talking dictionary reflected these inconsistencies; the audio was widely misattributed, mismatched with the text, and of poor quality as it was mistakenly taken from the low-quality files used for visualization rather than the original master recordings. In the absence of a content management system adapted to this task, it is imperative that the project manager work in close collaboration with technical resources to identify and correct these problems. It would be useful to continuously build and deploy the electronic version of the dictionary, and to track any integration problems, from the beginning of the annotation process.

The first priority when building the dictionary was thus the reconstruction of the metadata and retrieval of the original audio files. As well, while the post-correction of the OCR output resulted in a fairly consistently formatted text faithful to the original print version, the organization of the entries in this text created numerous problems when converting them to a structured format for presentation.

Among other things, this required the development of a language identification system, detailed in Section 4.1.

Finally, after extracting structured text from the dictionary entries, a subsequent matching was performed against the ELAN annotations to identify and extract the corresponding audio segments. Because of the divergence between original and post-corrected text, as well as the fact that annotators frequently (but inconsistently) corrected the text in the annotations, this required a multi-stage heuristic strategy in order to maximise the audio coverage, detailed in Section 4.2.

Because of the extensive recording and annotation efforts detailed in Section 2, there are often recordings of multiple speakers for both Michif definitions and example sentences. This level of complexity in the dictionary entries was not supported by the current version of MotherTongues at the time. We therefore extended both the dictionary builder and the Web user interface to support it, using a more flexible JSON-based input format.[5]

In order to quantify progress in improving the conversion workflow, 100 random entries were sampled and manually converted to this format, and the performance of the system evaluated using precision and recall over definitions and examples. Along with the audio coverage, this F1 score was also recorded and tracked for each weekly build of the dictionary during the development process.

4.1 Entry Extraction

The text of the Turtle Mountain Dictionary consists of 350 pages of Michif lexical entries and example sentences, organized into 9,181 English headwords followed by one or more Michif definitions and associated example sentences in English and Michif. We use "definitions" to describe these because they are not necessarily Michif lexical entries; in many cases, they give a *description* of the English word rather than the actual term used in Michif. For example, the definition for *zucchini*, shown in Figure 3, literally means 'a type of pumpkin', while the example sentence simply uses *zucchini*. [6] Likewise, the defintion of *zinnia* literally means 'flowers of all sorts of colours'. No lexical information such as part of speech, verb class, order, or gender is provided in the original text.

`zucchini en sort di sitroouy; I like zucchini cooked any way. Niweehkishpwow zucchini pikou ishi ay-ishikeeshishoust.`

`zinnia lee flueur tout sort di koulueur.`

Figure 3: Examples of descriptive definitions

reflect `wawshaynikayw, wawshayshkoutayw, nanawkatawayistamihk, kanaw katawayhtem;` *The mirror reflects the light.* `Wawshaynikayw le meerway. Wawshayshkoutayw li meerway.` *He'll reflect on his past actions.* `Kananawkatawayistam tawnshi aykitahkamikishit. Kanawkatawayhtem kawpaytootahk.`

Figure 4: Entry structure *(English in italic)*

Though the text of the dictionary entries have a relatively consistent structure, the English example sentences and their Michif translations are not attached to the corresponding Michif definitions or consistently ordered. In general, they are organized in pairs of English and Michif texts. However, these pairs may contain varying numbers of sentences, which in turn may correspond to one or more examples. For example, in Figure 4, there are four Michif definitions and two English example sentences, each of which has two different corresponding Michif examples. The extraction process must therefore:

1. Identify and separate the headword and the individual definitions.
2. Separate English and Michif example texts.
3. Create pairs of English and corresponding Michif examples.
4. Match Michif example texts to the corresponding definition words.

In the majority of cases the dictionary text follows one of two straightforward patterns; either

[5] Our modifications will be included in the next release of MotherTongues but are also available at https://github.com/p2wilrc/mothertongues/

[6] *zucchini* is also commonly used in Québec French instead of the standard *courgette*.

English and Michif examples alternate, or a single English example is followed by multiple Michif example sentences, one for each definition. In some cases, the individual examples also consist of multiple sentences.

To split the text into sentences, we used the PySBD library (Sadvilkar and Neumann, 2020), which required some post-processing to compensate for inconsistencies in how punctuation and abbreviations were used in the original dictionary. The initial version of the dictionary used the off-the-shelf langid.py library (Lui and Baldwin, 2012) to identify "not English" sentences as presumably Michif. This performed poorly, because obviously, Michif is not present in the langid.py model, but also because the orthography used in the Turtle Mountain Dictionary was specifically designed to resemble English (Laverdure et al., 1983).

Instead, we created a binary classifier for English versus Michif, using fastText (Bojanowski et al., 2017) with 5-gram subword features, making the assumption that the English headwords are valid English and the Michif definitions are valid Michif. We manually created a development set consisting of 1250 Michif and and 1239 English example sentences to evaluate the performance of these models, obtaining 99.4% accuracy, compared to 84.3% for the original langid.py based approach. Because any error is unacceptable in the final dictionary, we maintain a separate list of "overrides" to correct any errors found in testing. Likewise, we keep a list of "uncorrectable" dictionary entries with manually extracted definitions and examples where the original text cannot be parsed.

Once the English and Michif sentences have been identified and pairs of examples created, they are scored against the Michif definitions using the minimum Levenshtein distance between the definition and any subsequence of the example, with whitespace and punctuation removed. In some rare cases, this leads to incorrect matches due to the fact that the definitions are fully-inflected forms rather than lemmas and may not match the forms used in the examples. It may be useful to implement and evaluate a lemmatizer to improve the example matching.

4.2 Annotation Matching

As mentioned in Section 2, the original recordings contain 181 hours of audio. Of these, there are 105 hours of speech, which were annotated to identify the 18 hours of speech corresponding to the Michif dictionary entries and examples. This number is considerably smaller than the total amount of speech, as all entries and examples were read multiple times, with the best reading selected for the dictionary. There are also numerous discussions between the speaker and the linguist regarding the text. After extracting the structure of the dictionary entries, we process the annotation files using pympi-ling (Lubbers and Torreira, 2013-2021), collecting all the tiers for an aligned annotation in a single "Span" and matching these spans to entries in the dictionary.

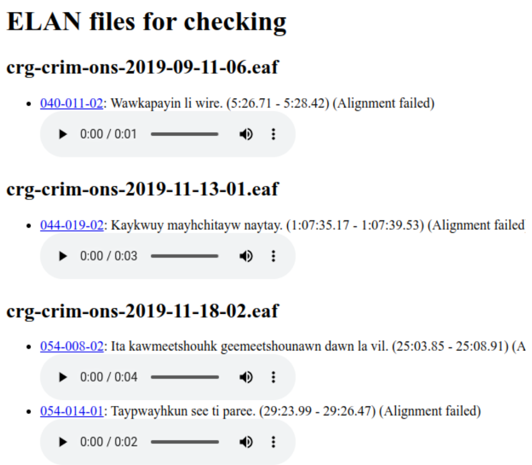
To compensate for the variable correction of OCR errors in the ELAN files, we perform a severe normalization of the text before matching annotations, collapsing various ambiguous characters or sequences (for example, w/vv, t/f, as well as the ones noted previously). In addition, we neutralize common spelling variations in the Michif text such as ou/oo. In some cases, the text is reduplicated in the annotations, so we check and repair this as well.

Finally, although we used a controlled vocabulary for the type of annotations, the difference between definitions and examples is not at all clear in the original dictionary, so they are often misannotated. In the case where this misannotation is unambiguous, we were able to repair these automatically with a Python script, but in some cases this was not possible. For this reason, the matching algorithm collects as many annotations as possible, matching on both the English and Michif text, then ordering by match and annotation type as well as normalized Levenshtein distance.

A significant challenge for the audio matching is reassembling the multiple fragments of an example which were split by voice activity detection. Annotators were instructed to select only one instance of any definition or example for a given speaker, and to use annotation types for the subsequent fragments, but this is not done consistently. In the case where an audio clip is to be spliced together from multiple instances, the original fragments are sometimes out of order in the recording, and while this is indicated by annotator notes, it is done in free text rather than with a controlled vocabulary, requiring heuristics and in many cases manual corrections to the annotator notes in order to get the correct ordering in the output. We discuss the detection and correction of these errors in the next section.

5 Verification and Re-Annotation

In testing the talking dictionary, it became obvious that many audio entries were incomplete or mismatched to the text. Given the scale of the recordings and annotations and the limited resources available, we attempted to use force-alignment to detect these problems, similarly to how the Festvox system (Anumanchipalli et al., 2011) excludes incorrectly labeled prompts to avoid egregious errors in unit-selection synthesis. Of course, no pretrained acoustic models exist for Michif. Using the "universal" grapheme-to-phoneme technique from Pine et al. (2022), we create an approximate phonetic transcription of the Michif text, then use the same alignment technique as Littell et al. (2022) with a narrow beam search, flagging examples that fail to align for review. To streamline the workflow, we collect the audio clips on an HTML page, shown in Figure 5, which we package with the relevant ELAN annotation files and preference (`.pfsx`) files which direct ELAN to open directly on the annotation to be reviewed.



ELAN files for checking

crg-crim-ons-2019-09-11-06.eaf

- 040-011-02: Wawkapayin li wire. (5:26.71 - 5:28.42) (Alignment failed)

crg-crim-ons-2019-11-13-01.eaf

- 044-019-02: Kaykwuy mayhchitayw naytay. (1:07:35.17 - 1:07:39.53) (Alignment failed

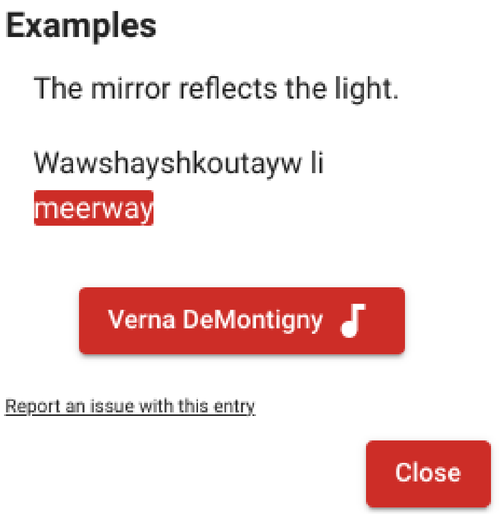
crg-crim-ons-2019-11-18-02.eaf

- 054-008-02: Ita kawmeetshouhk geemeetshounawn dawn la vil. (25:03.85 - 25:08.91) (A

- 054-014-01: Taypwayhkun see ti paree. (29:23.99 - 29:26.47) (Alignment failed)

Figure 5: HTML page for reannotation

Since false positives are not problematic (we can simply listen to them to determine that they are correct), a weak alignment model of this sort is quite effective, allowing us to detect and correct several hundred annotations which could not be fixed by the automated processes described in Section 4.2, generally in cases where one segment of an example that was split by VAD was not properly labeled by the annotators. An unintended side benefit of this verification is that it gives us word-level time alignments for the example sentences. We therefore extended the MotherTongues system to include a "read-along" style highlighting of each word when listening to the examples in the talking dictionary, as shown in Figure 6.



Examples

The mirror reflects the light.

Wawshayshkoutayw li meerway

Figure 6: Read-along highlighting

6 Conclusions

Some of the technical difficulties we had to overcome in creating this resource stem from organizational difficulties exacerbated by the COVID-19 pandemic. Others may simply be inherent to a large-scale, widely distributed and heterogeneous data collection and annotation effort. For future projects of this scale, it is crucial to endure that metadata is continuously validated and to avoid, at all costs, duplicating it across multiple unsynchronized data sources. It is equally important to involve a variety of perspectives in the design of the data collection and processing workflows, including members of the speech community, documentary and computational linguists, and to allow for iterative improvements to these processes.

When structured data is created as part of the data collection and annotation process, this data should be considered authoritative and maintained as such. If created from an unstructured data source (such as the OCR output of the paper dictionary), there should either be a robust process to pull changes and corrections from this original data source into the structured data, or the original unstructured data should be archived and left alone. This may require careful consideration of the dependencies between different steps in the process to avoid duplicate or conflicting efforts.

Some of these issues could be avoided with sufficient and appropriate tooling. In particular, while

ELAN is a robust and highly useful tool for annotation, it is difficult to integrate with external sources of metadata, distributed filesystems, or version control systems. While ELAN is highly extensible, with numerous third-party plug-ins and add-ons, it inherently operates at a single-file level, making it cumbersome to perform tasks involving individual annotations across a large number of EAF files. This could potentially be achieved by adding an API to ELAN which would allow it to be controlled by an external content management system.

Overall, this project has resulted in a resource that will be of long-term use in Michif language teaching, revitalization, and study. The dictionary application is now not only accessible to a wide range of users, but is also searchable, and the recorded Michif pronunciations of the headwords and example sentences will be extremely valuable for learners. Moreover, a total of 16 Métis team members were trained in language documentation and Indigenous language technologies, developing local capacity. In particular, the annotators involved in this project developed technical skills while also gaining valuable exposure to the Michif language. They will be able to carry this experience and knowledge with them as they continue their language journeys and contribute to future language revitalization initiatives.

Acknowledgements

This work would not have been possible without the participation of Verna DeMontigny, who not only recorded every single page of the dictionary, but also provided invaluable expertise as a native Michif speaker, educator, and Elder. Thank you also to the other Michif speakers who lent their voices to the dictionary: Connie Henry, Albert Parisien, Sr., and the late Sandra Houle. This work would equally have been impossible without the contribution of Turtle Mountain Community College, who graciously gave permission to use the original dictionary text. Samantha Cornelius and Janelle Zazalak coordinated the annotation effort, while Jacob Collard and Fineen Davis implemented the initial processing of the OCR text and annotations for MotherTongues. Tiara Opissinow provided essential assistance with project management and reannotation. David Delorme-Forsman also assisted the reannotation effort. Christopher Cox and students in the School of Linguistics and Language Studies at Carleton University contributed to the digitization and OCR correction processes, while Gilles Boulianne of the CRIM aided greatly with the ELAN-VESTA annotation process. We are extremely grateful to the National Research Council for supporting the project, and especially to Roland Kuhn for his support, patience, and enthusiasm. Kihchi-marsii!

References

Gopala Krishna Anumanchipalli, Kishore Prahallad, and Alan W Black. 2011. Festvox: Tools for creation and analyses of large speech corpora. In *Workshop on Very Large Scale Phonetics Research, UPenn, Philadelphia*, volume 70.

Peter Bakker. 1997. *A Language of Our Own : The Genesis of Michif, the Mixed Cree-French Language of the Canadian Metis*. Oxford University Press, Oxford & New York.

Christine Beier and Lev Michael. 2022. Managing Lexicography Data: A Practical, Principled Approach Using FLEx (FieldWorks Language Explorer). In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister, editors, *The Open Handbook of Linguistic Data Management*, page 301–314. The MIT Press.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Kari A.B. Chew, Sara Child, Jackie Dormer, Alexa Little, Olivia Sammons, and Heather Souter. 2023. Creating Online Indigenous Language Courses as Decolonizing Praxis. *The Canadian Modern Language Review*, 79(3):181–203.

Gabriel Dumont Institute. 2012. Michif Dictionary and Phrase Primer. https://www.metismuseum.ca/michif_dictionary.php. Accessed: 2024-01-31.

Vishwa Gupta and Gilles Boulianne. 2022. CRIM's Speech Recognition System for OpenASR21 Evaluation with Conformer and Voice Activity Detector Embeddings. In *International Conference on Speech and Computer*, pages 238–251. Springer.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 4, pages 19–24. IEEE.

Patline Laverdure, Ida Rose Allard, and John C. Crawford. 1983. *The Michif Dictionary: Turtle Mountain Chippewa Cree*. Pemmican Publications, Winnipeg.

Patrick Littell, Eric Joanis, Aidan Pine, Marc Tessier, David Huggins Daines, and Delasie Torkornoo. 2022. ReadAlong studio: Practical zero-shot text-speech alignment for indigenous language audiobooks. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 23–32, Marseille, France. European Language Resources Association.

Patrick Littell, Aidan Pine, and Henry Davis. 2017. Waldayu and waldayu mobile: Modern digital dictionary interfaces for endangered languages. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 141–150, Honolulu. Association for Computational Linguistics.

Mart Lubbers and Francisco Torreira. 2013-2021. pympi-ling: a Python module for processing ELANs EAF and Praats TextGrid annotation files. `https://pypi.python.org/pypi/pympi-ling`. Version 1.70.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Aidan Pine, Patrick William Littell, Eric Joanis, David Huggins-Daines, Christopher Cox, Fineen Davis, Eddie Antonio Santos, Shankhalika Srikanth, Delasie Torkornoo, and Sabrina Yu. 2022. G$_i$2P$_i$ rule-based, index-preserving grapheme-to-phoneme transformations. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 52–60, Dublin, Ireland. Association for Computational Linguistics.

Nicole Rosen, Marie-Odile Junker, Delasie Torkornoo, and Andrei Belcin. 2016. Michif Online Dictionary. `https://dictionary.michif.atlas-ling.ca/`. Accessed: 2024-01-31.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Olivia Sammons. 2019. *Nominal Classification in Michif*. Ph.D. thesis, University of Alberta.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).