Digitizing, Translating, and Standardizing Pr. Albert Lacombe's *Dictionnaire de la langue des Cris*

Daniel Dacanay, Antti Arppe

Special Characters: ā, â, à, ç, ê, è, é, î, œ, ù

A general interest in the lexicographic documentation of the Cree language has existed since at least the beginning of the 18th century. For instance, early translated Cree glossaries, such as those of Henry Kelsey and Thomas Bowrey (1710) and James Isham (1743), generally contained lists of (at most) several hundred words and phrases, with a strong focus on terms relating to the then-profitable fur trade (Tremblay 2005). Relatively simplistic wordlists such as these remained the chief written European sources on the Cree language until the mid-19th century, when, within the span of 10 years, two major bilingual dictionaries were published, seemingly independently of one another. The first of these was the *Dictionary of the Cree Language as Spoken by the Indians of the Hudson's Bay Company's Territories*, compiled by the Anglican missionary Revd. Edwin Arthur Watkins in 1865, containing ~13500 entries spanning several dialects of Cree concentrated in the vicinity of James Bay (chiefly corresponding to the modern dialectal divisions of Swampy Cree, Moose Cree, and East Cree (or James Bay Cree)), as well as occasional entries in Saulteaux and Ojibwe which Watkins considered "mingled with those in common use" (Watkins 1865). The focus of this paper, however, is to be the second of these two early dictionaries, namely, Pr. Albert Lacombe's *Dictionnaire de la langue des Cris*.

The Lacombe Dictionary

Lacombe's dictionary, compiled over the course of "more than twenty years" of missionary work among various indigenous groups in the "missions of Saskatchiwan [sic]" (Lacombe 1874), is a bilingual document, consisting of ~11400 Cree entries with French translations and slightly fewer than 9000 French entries with Cree translations, with considerable lexical overlap between the two. In addition to this, it contains an extensive grammar, and is interspersed with translated example sentences and ethnographic notes throughout. Much like Watkins' dictionary before it, the Lacombe dictionary was designed with the specific intention of training Christian missionaries to "communicate with [the] savages" so as to "spread among the errant tribes the lights of civilisation" (Lacombe 1874). In this capacity, much of the dictionary's content, particularly its example sentences, carry definite proselytic tones (see the Dɪsᴄᴜssɪᴏɴ section), and the entries contain a disproportionate number of terms relating to theology and liturgy. Nonetheless, the Lacombe dictionary constitutes the first major French dictionary of Cree, as well as the first major dictionary (and still the only French dictionary) of the dialect of Plains Cree, or *nêhiyawêwin*.

The contents of the dictionary are formatted in much the same way as most dictionaries of the time, with two columns of entries per page, with each column headed by the first three letters of the entries within it. Each entry consists minimally of a headword, a part-of-speech tag, and a definition, in that order. Concerning the part-of-speech tags, the exact means by which Lacombe determined the part-of-speech for each entry is unclear; however, it appears that, for most entries, the part-of-speech of the Cree headword is made to correspond with the part-of-speech of the translated definition word in French. Additionally, although Lacombe uses over twenty different

acronyms and abbreviations as part-of-speech tags, he only defines thirteen of them in the dictionary preface, making it unclear what many of these tags indicate.

Immediately following the headword, most entries also have their relevant plural suffix(es) listed; for most verbs, this plural suffix is the suffix used for third person plural agreement. For transitive verbs (particularly VTAs, transitive verbs with an animate subject and animate object), Lacombe instead includes a list of other potential derivational endings for the relevant verb stem (the most common being TCHIKEW (*-cikêw*, with 1075 occurrences), TAM or TTAM (*-tam*, 718), HIWEW (*-hiwêw*, 326), and TTAW (*-htâw*, 322)). These suffixes, as with the headword itself, are rendered in majuscule. The definition, by contrast, is rendered in minuscule, as are any example sentences included within the definition. The definitions themselves demonstrate little inter-entry formatting consistency, with some resembling truncated bullet points and others being near page-long paragraphs. However, nominal Cree headwords are almost always given in the singular, and verbal headwords are virtually always presented in the third person singular Independent form. Finally, Lacombe makes liberal use of the shorthand *idem.* to indicate that the meaning of an entry is identical to the meaning of the previous entry (with the exception of any grammatical information communicated using the part-of-speech tags). In total, 1175 entries (or roughly 10.3% of all entries in the dictionary) are glossed as *idem*. An illustrative sample of several entries, taken from page 353 of the Lacombe dictionary, is presented below:

« ISKANAPIHEW, (v. a.) TTAW, HIWEW, TCHIKEW, il arrête sa marche pour se reposer un jour ou plusieurs
« ISKITCHEW, ok, (a. a.) le cœur lui fait défaut, le cœur lui tombe, il a une douleur au cœur

« ISKIKEPIW, ok, (a. a.) il est lassé dans tous ses membres

« ISKAKUTTEW, a, (a. in.) ça finit par tomber, parceque ça y a été trop longtemps attaché

« ISKAKUTTIN, wa, (a. in.) idem

« ISKIPEW, ok, a, (v. im.) l'eau monte. On dirait: n't'iskipân, l'eau monte là où je suis; e.g., kakiyaw ayisiyiniwok ki iskipewok, tous les hommes furent submergés; ekuspi misiwe ki iskipew askiy, alors toute la Terre fut couverte d'eau

Given its foundational status in Cree lexicography, the Lacombe dictionary consists almost entirely of novelly recorded vocabulary; that is to say, virtually every entry in the Lacombe dictionary was recorded directly from first-language consultants by Lacombe himself, with minimal reliance on secondary data. Not only this, but owing to its being written during a period of time in which the vast majority of Plains Cree people spoke their language natively and as their predominant means of communication, the Lacombe dictionary provides a picture of the Cree language during its demographic 'golden age', in contrast to dictionaries compiled in more recent years, which have largely reflected Plains Cree in its modern state as a minority language which is seldom used in most social settings and domains, and whose lexical richness among most contemporary speakers has suffered as a result. In comparison to the entirely novel lexical stock of the Lacombe, most later Plains Cree dictionaries have relied heavily on the contents of existing lexicographic works to inform their choice of entries, with the Lacombe dictionary itself naturally being chief among these sources. In this way, the contents of the Lacombe dictionary serve as a common lexico-genetic link between much of the extant lexicographic documentation of Plains Cree; for instance, Arok Wolvengrey's *nêhiyawêwin*: *itwêwina* (Cree: Words) (2011),

LeClaire and Cardinal's *Alberta Elders Cree Dictionary* (2002), and Pr. Gérard Beaudet's

Cree-English, English-Cree Dictionary (1995) all list the Lacombe dictionary as a lexical source.

However, despite this foundational status, the Lacombe dictionary remains broadly inaccessible

to a large majority of the modern Plains Cree-speaking population on account of two primary

factors, namely, the lack of accurately-transcribed digitizations of its content and the fact that its

translated definitions are written entirely in French, a language spoken by only ~6.2% of the total

population in Alberta and Saskatchewan (2% natively and 4.2% simply reporting "Knowledge of

French") (Statistics Canada 2019a; Statistics Canada 2019b)), the provinces in which the

majority of the Plains Cree-speaking population resides. In fact, the Lacombe dictionary was

initially brought to our attention for precisely this reason; in summer of 2016, one of the

Cree-speaking consultants participating in the audio recordings for the Spoken Dictionary of

Maskwacîs Cree (https://speech-db.altlab.app/maskwacis/entries), named *kîsikâw*, requested that

the dictionary, of which he had acquired and demonstrated to us a physical copy, be digitized and

translated so as to be better available to potential contemporary users. As such, this paper

discusses the initial steps in the creation of a digitized translation of the *Dictionnaire de la*

langue des Cris, as well as provisional attempts to adapt its spelling conventions to those of the

modern Standard Roman Orthography (Okimāsis and Wolvengrey 2008), so as to make it more

accessible to contemporary Cree speakers familiar with this orthography, to allow its contents to

be parseable by computational models trained using the Standard Roman Orthography, as well as

to allow its integration into existing online Cree lexical databases. In doing this, we also discuss

the challenges posed in adapting the dictionary's content for integration into these contemporary

lexical resources, as well as the necessary avenues of development for the dictionary's use as both a practical tool and a historical document by modern Plains Cree speaker communities.

Although no full digitization of the Lacombe dictionary exists, it would be inaccurate to state that the dictionary's contents are entirely unavailable in a digital setting. There exist high-quality, freely available online image scans of the Lacombe dictionary (such as those of the Bibliotheque et Archives Nationales du Québec (BANQ), available using the BANQ website at the following link: https://numerique.banq.qc.ca/patrimoine/details/52327/3994014), as well as digitized plain text versions produced using optical character recognition (OCR). However, useful though they may be, these resources are still inadequate for the use of the dictionary as a large-scale online lexical resource for a variety of reasons; the image scanned versions, despite their quality, are non-searchable, and the OCR digitizations, although complete in scope, are unreliable in terms of their quality.

Existing OCR Scans

Several existing OCR scans of the Lacombe dictionary exist online; this investigation will discuss those available through the University of Alberta's Bruce Peel Collection, the John Carter Brown Library, Oxford University, and the Library of Congress. Despite all originating from (theoretically) identical source material, these four OCR scans differ widely in their accuracy to the original text, particularly concerning the Cree headwords. The general differences in quality between these four OCR versions may be demonstrated by examining how

each of the four models transcribed the entry for the Cree word *âstwâwin*, glossed by Lacombe (who does not mark the initial long vowel) as follows:

« ASTWÂWIN, a, (n. f.) réserve de provisions.

The Peel Collection OCR (http://peel.library.ualberta.ca/bibliography/708.html) is virtually unreadable, inserting spaces and omitting length markers in the Cree headword, mistranscribing the French gloss, and misinterpreting the part-of-speech tagging as running text:

u ASTWA WIN, a, To, Cl résérue & M provisions.

The Oxford University scan (https://hdl.handle.net/2027/hvd.hx79ui) is a slight improvement, correctly transcribing the vowel length diacritic on the second <A>, but still erroneously inserting a space in the middle of the head word. The French translation, however, is truncated to the point of near illegibility:

«ASTWÂ WIN, a, (n. f.) ré   provisions.

The John Carter Brown scan (https://brown.primo.exlibrisgroup.com/permalink/) is the only one of the four to properly transcribe the Cree headword (at least, to the standards of Lacombe's orthography); however, it still omits a portion of the French translation:

« ASTWÂWIN, a, (n. f.) réserve \ provisions.

The Library of Congress scan (https://lccn.loc.gov/02003610) is the highest quality of the four, and correctly transcribes and formats the French gloss and part-of-speech tags. However, the transcription of the Cree headword is missing its length diacritic:

« ASTWAWIN, a, (n. f.) réserve de provisions.

Given errors such as these, which are consistent (and occur in roughly the same degree) across each scanned version, the use of an existing OCR scan of the Lacombe as the basis for an English translation would likely save little time in the overall translation process, as manual annotation would be required throughout the dictionary's contents to correct for transcription errors in Cree headwords, even with the highest quality existing OCR models. This is likely at least partially the result of the bilingual nature of the text, which freely alternates between Cree and French prose with little overt textual separation; any OCR model intended to transcribe this content would need therefore not only be able to accurately transcribe both French and (an idiosyncratic spelling system for) Cree, but would also be able to distinguish between the two on a word-to-word basis. Initial attempts at producing an original OCR version of the Lacombe using Tesseract and the BANQ image scans were unpromising, and showed much the same errors as the above-described scans. However, further research remains to be performed to determine the ideal means to approach OCR on bilingual Cree texts such as the Lacombe. This will principally concern determining whether higher quality transcriptions may be obtained by training a novel OCR model specifically for the text, using an existing French model and then

applying a novel Cree model on the result to identify Cree words, or by running an existing French model and a novel Cree model in parallel. Such investigations are left to future research.

Manual Transcription and Translation

Given the lack of success of previous and current attempts at OCR, the decision was made to transcribe and translate the dictionary's contents manually. This decision was also influenced by concerns over quality of translation, which (based on preliminary trials) could not be guaranteed on a large scale using freely available machine translation software such as Google Translate or DeepL (https://www.deepl.com/). Additionally, the previously-made decision to manually transcribe the dictionary mandated that each individual entry be reviewed by an annotator regardless, mitigating the additional labor necessary for a manual translation. For the sake of both saving time and avoiding redundancy, it was decided that only the 384-page Cree-French section of the dictionary's contents was to be translated, with the 274-page French-Cree section being, for the moment, ignored. This decision was also motivated by lexicographic factors, as the Cree-French section, which provides descriptions and definitions of Cree words themselves, better fits contemporary, descriptive standards for linguistic practice than the French-Cree section, which seeks to rationalize French words in a Cree context. This being said, the translation and transcription of the French-Cree section may still prove useful, and is to be the subject of potential future work.

The full process of transcription and translation took ~18 days of work by a single annotator (the first author), averaging ~110 entries per hour. In total, this translation yielded 11411 entries, a

plurality of which (4543) are already present in contemporary dictionary sources. This transcribed, translated version is available online at https://lacombe.altlab.dev/)

Orthographic Revisions

The aforementioned transcription and translation process resulted in a full version of the Lacombe dictionary's Cree-French section with the Cree words written in Lacombe's orthography and the French glosses for those words translated into English. Although still useful as a historical text, such a digitization remains limited from a practical lexicographic perspective on account of its outdated orthography. This is largely a concern for reasons of searchability, as, without a consistent spelling system, reliably searching for any given entry inevitably becomes a matter of trial and error, particularly when searching databases consisting of several distinct lexicographic sources. Although Lacombe's spelling system bears a strong resemblance to the modern Standard Roman Orthography (SRO), it remains divergent in various respects, such as its inclusion of voiced consonants, its use of <U> to represent some (but not all) unstressed vowels, its use of French-like digraphs and trigraphs such as <TJ> and <TCH> to represent the SRO character <c> (compare, for example, Lacombe's entries WATJIY ("mountain") and WATCHISTUN ("nest") (Lacombe 1874) with the SRO-rendered *waciy* ("hill, mountain") and *wacistwan* ("nest") (Wolvengrey 2011)) and its occasional (but inconsistent) use of double consonants to represent pre-aspiration, as well as other individual idiosyncrasies. Not only this, but the spelling of individual Cree words in the Lacombe dictionary, particularly with respect to diacritics, is subject to occasional error, as Lacombe himself acknowledges the introductory pages of the work:

"Il est très probable qu'on rencontrera dans ce second Dictionnaire certains mots écrits avec une telle quantité, et dans le premier avec une autre quantité. Cela peut venir de la part des imprimeurs, qui ignorant entièrement cette langue, ont pu prendre une voyelle pour une autre, et malgré tout le soin, qui a été apporté, pour la correction des *épreuves*, on ne devra pas s'étonner que dans ce premier travail, il se soit glissé bien des fautes." (Lacombe 1874)

"It is very likely that one shall encounter in this second dictionary certain words written in one way, and in the first [dictionary] in a different way. This may come on behalf of the printers, who, entirely ignorant of the language, may have taken one vowel for another, and in spite of the greatest possible care to ensure the correctness of the proof sheets, one should not be surprised in this first edition to see the inclusion of some errors." (Lacombe 1874)

Given the broadly inconsistent nature of the orthographic irregularities throughout the Lacombe, it was quickly determined that it would be infeasible to regularize the dictionary's content using only find-and-replace methods using regular expressions or a finite state based transcriptor; rather, the anthropogenic idiosyncrasies of the Lacombe were again regularized by manual annotation. Additionally, part-of-speech markings for entries, which in the original Lacombe dictionary were seemingly based largely on French parts-of-speech, were altered to fit modern Cree inflectional subclasses (identical to those used in the existing Cree dictionary *nêhiyawêwin*: *itwêwina* (Wolvengrey 2011), with these tags being assigned based on the judgement of the first author. The resultant revised text was formatted to represent a three-way distinction between

entries; namely, those which are present in other, contemporary digital dictionaries such as *nêhiyawêwin*: *itwêwina*, those which are derivationally analyzable, but not present in any current dictionary, and those which are derivationally unanalyzable (that is, the first author was not able to confidently determine the morphological makeup of the word using available Cree lexical resources). Entries present in other dictionaries were rendered in SRO and given modern part-of-speech tagging. Entries which were derivationally analyzable, but not present in other dictionaries were rendered in SRO and given modern part-of-speech tagging, but were marked with an asterisk. These entries were marked distinctly from entries present in other, modern dictionaries for reasons of attestation; while entries included both in Lacombe and in other, modern dictionaries can be presumed to still be in use by contemporary Cree speakers (given that they were also recorded by modern lexicographers), entries solely attested in Lacombe cannot necessarily be guaranteed to still be in use, given that they were recorded nearly 150 years in the past and have not been recorded since. This distinct marking is therefore intended to aid in the later process of reviewing the dictionary's contents with contemporary native speakers (see the DISCUSSION section); in the eventual fully standardised version, these asterisk markers therefore will be redundant, and will be made invisible to the end user. Entries which were derivationally unanalyzable could not be standardized (as the SRO overtly represents morphological information such as morpheme boundaries), and were thus left in their original orthography and with their original part-of-speech tags:

« SÂKIKUTEW, ok, (v. n.) he shows his nose
« SÂKISKWEW, ok, (v. n.) he shows his head
« SÂKITAWEHAM, wok, (v. n.) he arrives at the mouth of the river or the lake in a canoe

« *sâkikotêw, ok, (VAI-1) he shows his nose

« sâkiskwêw, ok, (VAI-1) he shows his head

« SÂKITAWEHAM, wok, (v. n.) he arrives at the mouth of the river or the lake in a canoe

Similarly to the non-standardized version of the text, the orthographically standardized Lacombe digitization will be available online using an identical interface to the aforementioned https://lacome.altlab.dev/.

DISCUSSION

The end results of this investigation are two-fold; firstly, it has yielded both the first fully-digitized and manually verified text version of the Cree lexicon of the Lacombe dictionary and the first English translation of the contents of that dictionary. Secondly, it has provided an initial, albeit incomplete attempt at the standardization of the Lacombe dictionary to the modern SRO. The immediate practical uses for these resources are manifold, ranging from increasing the size of existing lexical repositories for Plains Cree (such as the University of Alberta-run online dictionary *itwêwina* (itwewina.altlab.app)) by at least 4917 entries (possibly more, given that the example sentences have yet to be accounted for in this count), to providing easy access to historical Cree vocabulary not recorded in contemporary dictionaries, in addition to the general value of the Lacombe dictionary as a historical document. Not only this, but the translated Lacombe is also a valuable source of parallel running text, given the large number of translated example sentences present in the glosses of its entries. Although an exact figure is not known (on

account of these example sentences not being explicitly marked in the text), these example sentences likely constitute at least several thousand tokens of Plains Cree text.

Limitations of the Lacombe dictionary

Despite its many uses, the currently translated and transcribed Lacombe dictionary is far from complete, and represents only an initial benchmark in the full digitization process. Perhaps its most obvious deficiency at present is its omission of the entirety of the French-Cree section, which, although highly repetitious of the Cree-French section, nonetheless contains thousands of tokens and translations of Cree words, many of which are doubtless novel to currently digitized lexical sources. Furthermore, Lacombe's extensive grammatical notes and field observations, including many which document grammatical features now lost in Plains Cree (such as the dubitative verb paradigm, which is largely undescribed in English language sources) remain untranscribed and untranslated, as do his lengthy introductory segments. Additionally, although the example sentences found throughout the entries of the Cree-French section have been transcribed, they have still yet to be orthographically standardized, reducing their searchability (particularly in cases wherein the headword of the entry was also not yet standardizable). Finally, before making use of its contents alongside those of more contemporary lexical resources, it remains necessary (due to the substantial diachronic gap between the early 21st and late 19th century) to review the entries in the Lacombe dictionary with contemporary native speakers (particularly the entries not present in any other, modern dictionaries, as previously discussed). This serves two purposes, being both to identify and explicitly mark archaic or obsolete vocabulary, allowing us to display such vocabulary in online lexical resources with obvious

tagging to indicate that it is 'historical' or 'archaic', as well as to confirm that the definitions that Lacombe provides for certain entries are synchronically valid.

The consultation of native speakers concerning the contents of the Lacombe is not, however, to be limited to the process of semantic validation. Rather, community engagement is also necessary to determine which aspects of the dictionary's contents are congruous with the modern sociocultural sensibilities of Cree speakers and community members. This is of particular relevance given the dictionary's origins as a tool intended to ease the proselytization and conversion of the "savage" inhabitants of the Western Prairies (Lacombe 1874). Although relatively sympathetic to Indigenous perspectives for its time, the Lacombe dictionary nonetheless frequently lapses into moralistic tirades against the "superstitions of the savages", employing example sentences such as *kwatakittâwin kitchi iskutek namawiya kita kisipan*, ("the suffering (in Hell) will have no end"), and using terms such as 'jongleur' (a French term literally meaning 'juggler', but carrying definite charlatanical connotations) to refer to Indigenous medicine men. Rather than adopting the revisionist and historically dismissive approach of simply deleting such entries from the dictionary entirely, it is the opinion of the authors that decisions of 'suitable' and 'unsuitable' content should be left to native speakers and community members. From a practical perspective, this would entail a parallel process of determining which entries, example sentences, and notes should not be presented in generally available online versions of the dictionary at the same time as these entries and example sentences are reviewed by native speakers for their linguistic contents. This would result in two online versions of the translated Lacombe dictionary, the first being a 'revisionist' (or 'modernized') version with standardized modern orthography, archaic vocabulary either overtly tagged or omitted by default,

and definitions modified with community oversight, and the second being a 'non-revisionist' version, with Lacombe's original orthography, definitions, and notes (for practical purposes , essentially identical to the online version linked throughout the paper). As such, in proceeding with the digitization and standardization of the Lacombe dictionary, direct community engagement becomes a matter of practical necessity and priority; indeed, it should be recalled that this project was initiated following a request by a Cree speaker and community member in the first place.

CONCLUSION

Owing to the incomplete and impractical nature of existing digital versions of Pr. Albert Lacombe's *Dictionnaire de la langue des Cris*, the following investigation has produced a digitized, plain text version of the ~11400 Cree-French entries present therein, with the French glosses translated into English. Additionally, preliminary orthographic standardizations have been performed on the Cree headwords of the dictionary to modernize their spelling to the contemporary SRO, although a large portion of these headwords have yet to be derivationally analyzed, and thus cannot be confidently orthographically revised. Similar revisions have been performed on the part-of-speech tags present for each entry. Although usable and freely available in its present state, the current digitized version of the Lacombe is strictly provisional, with plans to fully translate the French-Cree section, the grammatical notes and field observations, and to complete the currently partial orthographic standardizations being prioritized. Additionally, prior to its incorporation into existing digital repositories of Cree vocabulary, we plan to run review sessions of the dictionaries entries with native speakers and community members, both to

identify and appropriately tag archaic vocabulary as well as to moderate and mitigate some of the more intrusive influences of the dictionary's initial intended purpose as a missionary tool.

REFERENCES

Beaudet, Gérard O.M.I. 1995. *Cree-English English-Cree Dictionary*. Winnipeg, MB: Wuerz Publishing Ltd.

Kelsey, Henry and Thomas Bowery. 1710. *A dictionary of the Hudson's-Bay Indian language*. London: s.n.

Isham, James. 1743. *Observations on Hudsons Bay*. Unpublished manuscript.

Lacombe, Albert O.M.I. 1874. *Dictionnaire de la langue des Cris*. Montréal, QC: C. O. Beauchemin et Valois

LeClaire, Nancy and George Cardinal. 2002. *Alberta Elders' Cree Dictionary/alberta ohci kehtehayak nehiyaw otwestamâkewasinahikan* (ed. Waugh, Earle H.). Edmonton, AB: University of Alberta Press.

Okimāsis, Jean and Arok Wolvengrey. 2008. *How to Spell it in Cree : the Standard Roman Orthography*. Regina, SK: Miywâsin Ink.

Statistics Canada, The French Language in Alberta, 2001 to 2016: Facts and Figures. 2019a. https://www150.statcan.gc.ca/n1/pub/89-657-x/89-657-x2019016-eng.htm. Accessed 26 Aug. 2022

Statistics Canada, The French Language in Saskatchewan, 2001 to 2016: Facts and Figures. 2019b. https://www150.statcan.gc.ca/n1/pub/89-657-x/89-657-x2019015-eng.htm. Accessed 26 Aug. 2022

Tremblay, Manon. 2005. Les dictionnaires de la langue crie : histoire et regard critique. MA thesis, Université de Montréal.

Watkins, Edwin Arthur. 1865. *Dictionary of the Cree Language, As Spoken by the Indians of the Hudson's Bay Company's Territories,* London, UK: Society for Promoting Christian Knowledge.

Wolvengrey, Arok. 2011. *Cree: Words*. Regina, SK: University of Regina Press.