Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-7), *Association for Computational Linguistics*, March 22, 2024.

Data Mining and Extraction: the gold rush of AI on Indigenous Languages

Marie-Odile Junker
Carleton University

Abstract

The goal of this paper is to start a discussion on the topic of Data mining and Extraction of Indigenous Language data, describing recent events that took place within the Algonquian Dictionaries and Language Resources common infrastructure. We raise questions about ethics, social context, vulnerability, responsibility, and societal benefits and concerns in the age of generative AI.

1 Introduction

In 2023 Artificial Intelligence (AI) became the buzz word, the cool word that could be attached to anything, and so why not to Indigenous languages, their endangerment, and the savior role that generative AI could play to rescue those languages from disappearance. In practical term, though, AI needs data. And data about Indigenous languages is rare, and rarely of a quality that can be of use, due to lack of standardization of orthography, dialectal variation, attrition, generational differences, and the various ills of colonialism and imperialism that have affected Indigenous groups stability. In this chaotic mix, the work of long engaged groups into the health of their language has become the new gold for the AI machine. Such gold is exemplified by the Algonquian Dictionaries and Language Resources project. We describe three recent events of data mining and extraction of Indigenous language data, that took place within our common digital infrastructure. Our goal is to raise awareness towards the development of a code of conduct or best practices for generative AI development (or not) in the service of Indigenous languages.

2 The Algonquian Dictionaries and Language Resources Project

The Algonquian Dictionaries and Language Resources project has been building a common infrastructure with language websites and tools for more than 20 years in collaboration with various language groups of the Algonquian language family. This family of languages, structurally similar -- yet very different from Indo-European languages -- stretches from the Atlantic Ocean to the Rocky Mountains (see the linguistic atlas: atlas-ling.ca). All dictionaries supported by this common infrastructure were built directly with their original creators, consisting of linguists and community linguists, Indigenous content editors, etc. The project continues to focus on the creation or enhancement of existing dictionaries in order to make them viable for the long term in the digital economy, and aims to address the following needs:

- thematic multimedia dictionaries: research on categorization methods, knowledge graphs
- morpheme dictionaries with a historical perspective, dialectal relationships; applications for the creation of terminology, development of text parsers
- bilingual dictionaries: standardization of English and French keywords, issues of synonymy, homonymy and polysemy
- unilingual dictionaries: definitions in an Indigenous language, sound files
- connections between dictionaries, grammars and text corpora (oral, transcribed and written)

- pedagogical tools: online lessons for first- and second-language, resources for language teachers, training of Indigenous lexicographers
- search engines, syllabic convertors, and verb conjugation apps (with extensive documentation of inflectional morphology)
- database structure with multimedia integration and exports in multiple formats: books, apps, online
- documentation and conversational resources; addition of Algonquian dialects to the Linguistic Atlas: www.atlas-ling.ca

The atlas, started in 2005, currently contains 68 speakers, 25,108 sound files, 58 communities (20 languages, 47 dialects) on 21 topics of conversation (plus one of songs and stories).

The 12 different dictionaries currently supported are variously active. Some represent earlier stages of a particular language now severely endangered, or even reconstructed historical ones like the Proto-Algonquian dictionary, some (for whom the language is still spoken) have very active editorial teams who meet weekly to address users' questions and constantly improve the content. Most also contain a series of related language resources, like oral stories, book catalogue, lessons, terminology forum, modelled after the eastcree.org website started in 2000 in collaboration with the Cree School Board in Quebec, Canada. The data and the software developed reside on protected (in Canada) secure servers, with separate access (by language and by level) for each group of users.

The maintenance of these resources has always been problematic. It is because there was no capacity at the local, and no support at the governmental, level for the Algonquian languages, that the idea of a common infrastructure was developed. By pulling resources together, each group could take a turn at sustaining the whole; each funded project could pitch in to help the other ones. While these resources have always generated some interest outside the circle of the communities they were intended to serve (current average of over one million words searched annually in the dictionaries), there has been a recent surge of interest from Artificial Intelligence players and we have observed more and more data-mining events that have forced us to halt some work we were doing on search optimization, APIs availability, open-source practices and free open-access. In this paper, we disclose (anonymously[1]) three of these events that illustrate well the vulnerability of Indigenous groups in the face of AI. Our goal is to raise awareness towards the development of a code of conduct or best practices for AI development (or not) in the service of Indigenous languages.

3 The Good: Data mining with permission

This event took place between July and October 2023. The request came from a team of linguists in a European university, conducting research on morphological change, using AI for processing large data sets of as great a variety of languages as possible. Most of their research to date had been conducted on Indo European languages, and they wanted to include Algonquian languages to their dataset and research program, so as to avoid biases. For that, they needed access to conjugation guides[2] and the Proto-Algonquian dictionary. The questions they raised in their initial request were genuine and clear. For example: "Do you have the data in a .csv file? If not, are you OK with our data science consultant "scraping" the site? If we want to archive the database we build during this project, can we include this data (in a potentially open-source format, with appropriate citation of data sources)?" After several written exchanges and a zoom meeting they came up with an agreement proposal which we further edited together. It was quite restrictive and respectful of OCAP[3].

These European linguists submitted their data usage agreement to the Indigenous group. In light

[1] This paper is single authored to preserve anonymity of the stakeholders. It is the result of discussions with the Indigenous partners involved. All errors are mine.

[2] See for example: East Cree Verb Conjugation Guide (Southern dialect): southern.verbs.eastcree.org/; (Northern dialect): northern.verbs.eastcree.org ; Innu Verb Conjugation Guide: verbe.innu-aimun.ca ; Guide de conjugaison atikamekw: verbes.atikamekw.ca

[3] OCAP refers to the principles of (Canadian) First Nations ownership, control, access and possession (https://fnigc.ca/) - which assert that First Nations have control over the data collection processes, and that they own and control the way in which this information can be used.

of the next event (below), that group declined permission. The European linguists then announced they would respect that decision and leave Algonquian languages out of their research.

4 The Bad: Data extraction without permission

This event took place over more than a year, but was brought to our attention in March 2023, when a couple of students contacted the infrastructure administration / dictionary editors, announcing that their thesis director (Let's call him/her professor X) had a project about one of the Algonquian languages (Let's call it language Y) for machine learning and AI driven translation. These academics are not linguists, but computer scientists in the field of Natural Language Processing. Earlier, professor X had made several unsuccessful attempts to enlist into their research program organizations or Indigenous scholars active in language Y. Professor X was told that they were not interested in AI and machine translation, at the moment.

Like the previous group described above, the students of Professor X asked for .csv files of the data, and full access to language Y databases. When told they needed permission from the language group to get such data, the team forged ahead anyway, mining the entire dataset of trilingual examples of language Y dictionary; a feature not accessible as such to the public. They then ran their experiments on machine learning with the mined data. They presented a paper at a conference in October 2023, claiming collaboration with Indigenous group Y, and citing the source they had mined. This is where one of the Indigenous scholars and co-editors of dictionary Y, who was attending the conference, noticed. The Indigenous organizations concerned then co-signed a letter with the four editors of the dictionary, asking for deletion of the mined data from all repositories, corpora, backups, etc., as soon as possible and the removal of the term "collaboration" from their research program, along with a reminder of the principles of OCAP.

5 The Ugly: extracting their data and selling it back to the Indigenous people

The last example happened in the Winter and Spring of 2023. A private web designing company offered to an Indigenous institution to improve the look of an Oral Stories database that had been built in collaboration with the Algonquian language resources project in 2010-13. The person who signed the contract within the Indigenous organization was unaware of what the project consisted of and based their judgement on the looks of the site, not its structure (no awareness of the existence of back end and careful data organization). The private company scraped off the trilingual data (audio, text and video) from the public interface, and rebuilt the database incorrectly, with many mistakes, but a lovely look. They put the data on a commercial server, and nobody knows the exact terms of this server provider. They were paid a large sum of money, 10% of which would have probably been sufficient to simply update the site properly without messing up the data. This last example points to the vulnerability of the people and the organizations themselves in the face of convincing commercial entities.

6 Misunderstanding tools and potential further misuse

This somewhat anecdotal last case illustrates the degree to which the general public misunderstands Indigenous language tools and what their expectations have become. In late 2023, I was contacted by a social service in the prison system in Quebec to help them adapt a short Cri text into roman orthography, because "not all prisoners can read syllabics". They wanted to be inclusive with the 5 languages spoken in that prison. They had obtained the text from our syllabic convertors[4], mistaking it for an equivalent of Google Translate. It turned out that the text was purely a French text, converted into syllabics (Figure 1).

[4] Cree syllabics convertors (Jancewicz et al, 2014): https://syllabics.atlas-ling.ca/

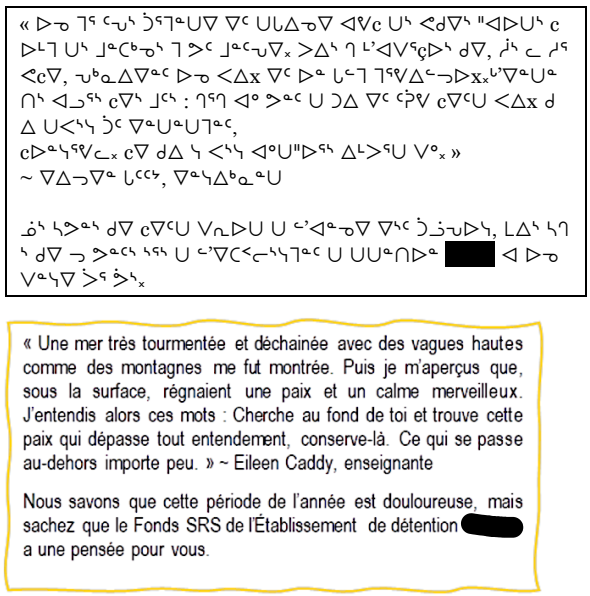

Figure 1: Cree Syllabic script on French text.

When I replied with explanations, they were truly incredulous; I do not know for how long they had been mistakenly using this tool to produce such "multilingual" messages. No wonder the Cree or Naskapi prisoners could not "read" or make sense of that syllabic text! Imagine what could happen when such data is scraped off the social media site of this organization and fed into the AI machine for further learning, thus compounding the confusion.

7 Discussion and Conclusion

The three events presented above are probably not the only ones that took place in 2023, as far as data mining and extraction is concerned on the Algonquian Dictionaries and Language Resources or the Linguistic Atlas. These are the ones for which we have solid proofs[5].

There is a huge economic imbalance at play here: the resources discussed, although freely available, are not supported, except by sporadic research grants to a few stable full-time academics[6], or by Indigenous organizations whose funding is usually unpredictable. Staff is hired on short-term or part-time contracts. Projects like ours wonder to whom we must "sell" ourselves next, in order for the language resources that we have worked so hard to develop, to survive. AI on the other hand, comes with lots of money, but none of the two above research cases discussed, offered any compensation[7]. Data available or extractible on the web is considered a free natural resource. But is it? Even if compensation was offered, how do we prevent downstream computational uses like models propagating mistakes in data? How are the models of generative AI going to feed back into the language and affect its (d)evolution? In a context of fragile language transmission, linguistic insecurity of younger speakers, and often severe endangerment, who can truly verify and validate the machine learning production? Who wants to? What kind of community-based important work is not going to be done if this is done instead? Whose agenda is controlling the field?

Many of these issues look like a new form of colonialism we could call "digital colonialism". While there has been some work by GIDA (the Global Indigenous Data Alliance) to address the ethics of Indigenous data and complement the FAIR (Findable - Accessible - Interoperable - Reusable) principles with the CARE (Collective Benefits - Authority to Control - Responsibility - Ethics) principles (Carrol et al., 2020, 2022) and Figure 2, we have not found (at the time of writing) any update on the GIDA site to help us deal specifically with the reality of Indigenous language data extraction.

[5] Meta released a paper about pre-trained Text-To-Speech systems including Canadian Indigenous languages, without stating that they obtained any permission and without evaluating the models with speakers. Since our databases contain lots of sound files, it is likely that they have been or will be mined like that for TTS research and development. See Pratap V. et.al (2023).

[6] These academics are usually in Linguistics or Language departments, an area way less funded, both in terms of academic positions and grants, than say, Computer Science and Engineering. While Inuktitut has some support due to its official status in Nunavut (Canada), there is no long-term governmental support for any Algonquian languages.

[7] The linguists of case one confirmed compensation for data had not been included in the research grant budget.

Figure 2: FAIR and CARE principles.

So, what might be some practical steps to address this current situation?

Protect the integrity of the data

While up to know the Algonquian project team has only been interested in anonymously tracking some human use of the Algonquian language resources to improve them, we recently asked our system administrator to implement analytics in order to inform our policies regarding bots. We have also started to post the following messages on top of all the dictionaries and conjugation apps credit pages: *Data-mining and scraping strictly prohibited.*

Our first attempt said "without permission" but we quickly realized we lacked time and human resources to handle requests for permissions and liaison with all the individual stakeholders.

Alert and inform all Indigenous and non-Indigenous stakeholders about the new reality of data-mining and generative AI

Most people have no idea what is happening with their own personal data, let alone what could happen to their language data. When contacted by people for projects some possible internal questions for Indigenous groups are:

- Who does this data belong to? Do I (as an individual) have a right to grant a permission to use it to someone? / Do I have the right to give it away? Who has this right?[8]
- Am I sufficiently skeptical about what is being offered for my language: are they using me for their own purpose, or to claim partnership?
- Do I have any control in this project? Did I define its goal, process? Do I/we need this? Or are they using me/us, paying me/us and I/we need the money?

When contacted by commercial entities to "update" or "modernize" their web resources, ask:

- How are you planning to retrieve the data? Can I provide it to you in proper form?
- Are you preserving all the functionalities of the site (front end, back end) for content updates?
- Where are you going to keep, store, distribute this data? Are you going to sell something to third parties?
- What programming tools, (open-?) source code are you going to use? And (since AI is increasingly used for code writing as well) how do you guarantee clean code?
- Do we get to keep the code?
- Who is going to provide updates and for how long?

When discovering data has been used without permission, do as the group in the second case did: ask for deletion of the mined data from all repositories, corpora, backups, etc. Demand a retraction and compensation? Explore legal avenues?

Some possible questions that genuinely respectful researchers should ask themselves:

- What can we offer in return?
- Do people need this or are we trying to convince them they do?
- What is my true goal/ purpose?
- Do I respect a "do no harm" principle?

The last three questions would also apply to the genuinely respectful commercial providers.

The ComputEL community of scholars should be engaging in this reflection and looking to contribute creative solutions. Let's not let the history of colonialism repeat itself! After the

[8] The Canadian government just completed (January 2024) a public consultation on generative AI and copyright, that will hopefully take into account Indigenous languages and cultures.

forests, the rivers, the soil, now the language? The challenge is upon us!

Acknowledgments

I am thankful to three anonymous reviewers for their suggestions on how improve my initial submission, to Te Taka Keegan for an earlier discussion of these problems with our team, to Delasie Torkornoo, and to my other colleagues and Indigenous partners who wish, at this point, to remain anonymous.

References

Jérémie Ambroise, Anne-Marie Baraby, Marie-Odile Junker, and Yvette Mollen (eds). 2023. Conjugaison des verbes innus (6th ed.). https://verbe.innu-aimun.ca

Marie-Odile Junker and Nicole Petiquay (eds). 2020. Conjugaison des verbes atikamekw (2nd ed.). https://verbes.atikamekw.atlas-ling.ca/

Marie-Odile Junker and Marguerite MacKenzie (eds). 2016. East Cree (Southern Dialect) Verb Conjugation (4th ed.). https://www.southern.verbs.eastcree.org/

Marie-Odile Junker and Marguerite MacKenzie (eds). 2020. East Cree (Northern Dialect) Verb Conjugation (5th ed.). https://www.northern.verbs.eastcree.org/

Stephanie Carroll, et al. 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19: XX, pp. 1–12. https://static1.squarespace.com/static/5d3799de845604000199cd24/t/6397b1aff7a6fb54defdf687/1670885815820/dsj-1158_carroll.pdf

Stephanie Carroll, Jewel Cummins, and Andrew Martinez. 2022. Indigenous Data Sovereignty and Governance. Global Indigenous Data Alliance. https://static1.squarespace.com/static/5d3799de845604000199cd24/t/640792a43ba5c11a1073bbc8/1678217895508/TheCAREPrinciples.pdf

Bill Jancewicz, Marie-Odile Junker and Delasie Torkornoo. *Cree Syllabics Convertor.* 2014-present https://syllabics.atlas-ling.ca/

ISED Citizen Services Centre (Innovation, Science and Economic Development Canada). 2024. Consultation on Copyright in the Age of Generative Artificial Intelligence. https://ised-isde.canada.ca/site/strategic-policy-sector/en/marketplace-framework-policy/consultation-paper-consultation-copyright-age-generative-artificial-intelligence

Marie-Odile Junker, Marguerite MacKenzie, Nicole Rosen, J. Randolph Valentine and Arok Wolvengrey. (2005-present) *Algonquian Linguistic Atlas*. https://www.atlas-ling.ca/

Marie-Odile Junker (dir.) (2005-present) *Algonquian Dictionaries and Languages Resources Project*. https://www.algonquianlanguages.ca/

OCAP: Ownership, Control, Access and Possession. OCAP® is a registered trademark of the First Nations Information Governance Centre (FNIGC) in Canada. https://fnigc.ca/ocap-training/.

Vineel Pratap et al. (2023) Scaling Speech Technology to 1,000+ Languages. Preprint https://arxiv.org/pdf/2305.13516.pdf