Strengthening Relationships Between Indigenous Communities, Documentary Linguists, and Computational Linguists in the Era of NLP-Assisted Language Revitalization

Darren Flavelle
CILLDI
University of Alberta
dflavell@ualberta.ca

Jordan Lachler
CILLDI
University of Alberta
lachler@ualberta.ca

Abstract

As the global crisis of language endangerment deepens, Indigenous communities have continued to seek new means of preserving, promoting and passing on their languages to future generations. For many communities, modern language technology holds the promise of accelerating that process. However, the cultural and disciplinary divides between documentary linguists, computational linguists and Indigenous communities have posed an on-going challenge for the development and deployment of NLP applications that can support the documentation and revitalization of Indigenous languages. In this paper, we discuss the main barriers to collaboration that these groups have encountered, as well as some notable initiatives in recent years to bring the groups closer together. We follow this with specific recommendations to build upon those efforts, calling for increased opportunities for awareness-building and skills-training in computational linguistics, tailored to the specific needs of both documentary linguists and Indigenous community members. We see this as an essential step as we move forward into an era of NLP-assisted language revitalization.

1 Introduction

The creation of NLP applications for Indigenous languages[1] has been an area of increasing interest (Arikpo and Dickson, 2018; Cadotte et al., 2022; Ortiz-Rogriguez 2022; Mohanty et al., 2023), even as the development of such tools lags behind those for majority languages (Littell et. al. 2018). Many have recognized (Liu et al., 2022; Schwartz, 2022) that one of the key challenges is that developing such applications for Indigenous languages requires the close collaboration of three disparate groups – computational linguists, documentary linguists, and members of Indigenous language communities.

In his paper on decolonising language work, Bird (2020) describes the steps which he believes are necessary in deepening engagement with language communities, decrying the ‘moralistic tropes’, the ‘nostalgia and sentimentalism’, and calling out the ‘professional narrowness of the focus on linguistic structures’; all of these contribute to the divide between Indigenous language communities and linguists of all stripes.

Nonetheless, the value of this relationship is widely recognized, as noted by Liu et al. (2022): "In the development of language technology, providing the speech communities a central role in the design and implementation of language tools may improve the likelihood of the tools’ success.”

This paper will discuss the challenges that these three groups face, certain steps that have already been taken to address the issue, and further recommendations that we have to improve the situation.

Section 2 will give an overview of what we perceive to be the main challenges to effective collaboration between these three groups. Section 3 will highlight some of the responses that the academic community has already taken to address these issues. Section 4 discusses the successes and limitations of those responses, and provides suggestions to resolve those issues and overcome future challenges. Section 5 provides a conclusion.

2 Articulating the Challenges

The overall challenges to collaboration among the three groups can perhaps best be understood by examining the challenges present in the relationships between each pair of groups.

[1] We have decided against providing a definition for "Indigenous" as no official definition has been agreed upon by any UN-system body; according to the UN the most fruitful approach is to identify, rather than define indigenous peoples. This is based on the fundamental criterion of self-identification as underlined in a number of human rights documents.

Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), pages 25–34
May 5, 2023 ©2023 Association for Computational Linguistics

2.1 Documentary Linguists and Indigenous Communities

The key challenge that these two groups have faced over the years stems from the different motivations they have had for engaging in the work language documentation.

For the majority of the history of linguistics involving Indigenous communities, documentary practices have centered academic concerns (Czaykowska-Higgins, 2009). This history did little to engender trust between language communities and documentary linguists, and stories of communities feeling exploited by extractive research practices are all too common. In recent decades, however, there has been a significant shift in practice towards more community-based approaches, placing the needs and interests of the Indigenous community closer to the forefront.

Documenting any language is a lengthy and complex process. This work requires the development and maintenance of long-term relationships between the linguists and their language consultants, and in the context of Indigenous language work, it is also necessary to develop and maintain that relationship with the Indigenous community more broadly. Not only is it important to understand that the process is not swift, but the speakers most often worked with are Elders, meaning that time is of the essence. (Siefart et al., 2018; Fitzgerald, 2021; Khawaja, 2021).

Negotiating between the needs of the researcher (e.g. meeting grant deadlines, getting publications, finding and keeping a steady academic position) and the needs of the community (e.g. documenting traditional knowledge, developing pedagogical materials, creating new speakers) can be an ongoing source of tension (Leonard, 2018; Paksi and Kivinen, 2021). Building relationships and maintaining them are of paramount importance to the ongoing work of documentary linguists; these are exemplified by the 5 R's of Research in Indigenous Research Contexts: respect, reciprocity, relevance, responsibility, and relationship (Restoule, 2008; Tessaro et al., 2018).

2.2 Documentary Linguists and Computational Linguists

While documentary linguists and computational linguists both come from and typically operate within an academic context, those similarities have not guaranteed successful working relationships.

To begin with, documentary linguists and computational linguists typically have little direct experience in each other's areas of specialization. Coursework in computational linguistics is rarely required (or even available) to students training to be documentary linguists, and vice-versa, and there are few if any linguistics departments that can be said to traditionally have strong programs in both areas.

This means that not only do that these linguists-in-training miss out on the opportunity to learn even the basic concepts of each other's fields, they also miss out on the opportunity to build connections with others who may go on to specialize in those areas. This has the effect of siloing these two groups off from one another even from their earliest stages of training.

Even when documentary and computational linguists do manage to come together to discuss possible collaborations, there are several ways in which Indigenous language can seem like a "poor fit" for traditional approaches to NLP development.

First, even relatively well-documented Indigenous languages lack the large-scale corpora that much of modern NLP development relies upon. The creation of such corpora is simply not feasible in situations where there are small numbers of speakers, and often just a single linguist working on the language. This places constraints on the computational methods that are available for use with these languages, and may also limit the types of applications that can be developed.

Second, NLP development often assumes the existence of a standardized version of the language in question, including both a standardized orthography, as well as a standardized and thoroughly documented set of grammatical rules. This is lacking for nearly all Indigenous languages, which often show significant dialectal and communalectal variation at all levels of the grammar. In many cases, speakers and communities place a high value on their specific, local ways of speaking, subverting the prevailing ideology of language standardization. Traditional NLP methods do not always handle such variation easily, and it may be seen as an unnecessary burden to need to account for it. For a more fulsome discussion of the usual needs of NLP for under-resourced languages, see Besacier et al. (2014).

Third, Indigenous languages are often typologically quite distinct from languages with existing NLP applications. Phenomena such as noun incorporation, complex agreement systems, and non-configurationality can present significant (though quite interesting) computational challenges (Sag et al., 2002; though for a counter to this, see Van Gysel et al., 2021). While many computational linguists have been eager to tackle such challenges, their presence means that using "out-of-the-box" computational approaches developed for majority languages is often not effective.

These factors, among others, may make some computational linguists hesitant to engage with documentary linguists on projects for Indigenous languages. The production of NLP applications for these languages will likely be slower, more complex and more labor-intensive than for majority languages. As a result, projects such as these run counter to the typical incentive structures found in academia, making it riskier for early-career computational linguists to devote their time and expertise to projects when there is no guarantee of tangible short-term results that can be reported on in journals and conference proceedings.

2.3 Indigenous Communities and Computational Linguists

While documentary linguists have the opportunity (and obligation) to spend significant time in the language community they are working with, computational linguists typically do not. Although this often makes sense from an efficiency perspective – the computational linguist's time is better spent developing the applications rather than traveling to the community to engage with speakers and learners – the lack of personal connections between the computational linguists and the language communities can make it more difficult for the computational linguists to be aware of, or to fully understand, the needs of those communities, and the challenges they face.

By the same token, even community members who work closely with documentary linguists may be completely unaware that computational linguists exist, let alone what type of work they do or how that work may be of benefit to the community's efforts at revitalization.

As such, it often falls to the documentary linguist to bridge this gap between the other two groups. They frequently work to make the computational linguists more aware of the priorities of the community, while at the same time trying to make the community more aware of the potential benefits of various NLP applications. They do this work not because their training in language documentation makes them particularly well-suited for the task, but because they are the ones who are in actual direct contact with the other two groups.

One key area where lack of familiarity with each other has been known to lead to conflict is around data sovereignty. Issues of data access, use, ownership and monetization are of great importance to Indigenous communities, who have suffered from the misappropriation and exploitation of their languages and cultures. The work of organizations such as the First Nations Information Governance Centre (https://fnigc.ca/) highlights both the importance and the complexity of these issues, including the need to develop culturally-appropriate and community-specific approaches to data sovereignty.

Computational linguists are typically unfamiliar with such concerns (for many of the reasons discussed above), and may feel that they represent further barriers to the timely production of the tools they are working to develop.

2.4 Summary

As we have seen, there are complex and often long-standing challenges to effective collaboration present in the relationships between any two of the three groups under discussion. When we seek to bring all three groups together to support the continued vitality of Indigenous languages, these challenges can be compounded, taking a task that was already difficult and making it appear daunting.

3 Academic Responses

Being aware of both these complexities as well as the urgency to overcome them, the academic community has taken a variety of concrete steps to begin addressing this challenge over the last several years. Several important initiatives can be highlighted here.

ComputEL began in 2014 as a two-day workshop that was part of the 52nd annual meeting of the Association for Computational Linguistics. It was billed as "The use of computational methods in the study of endangered languages". ComputEL-2 took place in 2017, this time as a two-day event co-located with the International Conference on Language Documentation and Conservation (ICLDC) (http://ling.lll.hawaii.edu/sites/icldc/) at the University of Hawaii, one of the largest and most prestigious conferences in its field.

The ComputEL workshops focus on "the use of computational methods in the study, support, and revitalization of endangered languages. The primary aim of the workshop is to continue narrowing the gap between computational linguists interested in working on methods for endangered languages, field linguists working on documenting these languages, and the language communities who are striving to maintain their languages." (https://altlab.ualberta.ca/computel-2/)

Subsequent gatherings have continued over the past six years, developing into a largely annual event co-located with either ICLDC or an ACL conference: 2019 ComputEL-3 @ ICLDC; 2021 ComputEL-4 online (w/ ICLDC); 2022 ComputEL-5 in Dublin @ ACL; 2023 Comput-EL-6 online (w/ ICLDC).

The development of the one-time workshop into an annual conference speaks to the recognition of the importance and timeliness of the work in this area.

Building on the development of ComputEL, the ACL Special Interest Group in Endangered Languages (SIGEL) was founded in 2019. The purpose of that group is to "foster computationally grounded research in all useful aspects in documenting, processing, revitalizing and supporting endangered languages, as well as minority, Indigenous and low-resource languages."

SIGEL has just over 150 members currently (March 2023) and has taken over the responsibility for organizing the ComputEL conferences. SIGEL has begun to organize an online speaker series focused on sharing best practices in this area. The first event was held in October 2021 with the theme of Automatic Speech Recognition in Native American Languages.

Relatedly, a separate ELRA/ISCA SIG, the Special Interest Group in Under-resourced Languages (SIGUL) was founded in 2017, and had its first meeting co-located with INTERSPEECH that same year. SIGUL positions its gatherings as "a forum for the presentation and discussion of cutting-edge research in text and speech processing for under-resourced languages by academic and industry researchers." (https://sigul-2022.ilc.cnr.it/)

"Under-resourced" is a very broad category when it comes to text and speech processing, but it certainly includes all Indigenous and/or endangered languages, in addition to others.

SIGUL further mentions: "It is also very important that these occasions leave space for communities and representatives of under-resourced and endangered languages, in order to ensure that the research and development of technological solutions are in line with the needs and demands of those communities, with a view to open and inclusive research with strong social impact."

The creation of these groups – as well as others such as Americas NLP (https://turing.iimas.unam.mx/americasnlp/) – the continuation of these conferences, and the publications that result from them, show clearly that much important work is being done in this area. However, these gatherings have so far struggled to attract a balanced mix of their target demographics – computational linguists, documentary linguists, and, most importantly, community members working to revitalize their languages.

While all of the organizers recognize the importance of "leaving space" for community voices in such gatherings, their very nature as academic gatherings (typically co-located with other, larger academic gatherings), with abstract deadlines, scientific committees and published proceedings, make it challenging to meaningfully include such voices. This is perhaps unsurprising, as we are still in the early days of organizing gatherings of this type. Much can likely be learned from the history of ICLDC and other gatherings such as CoLang (https://www.colanginstitute.org/), both of which have evolved over the past decade to be more inclusive of community voices in their presentations and courses, and have placed community needs closer to the centre of their remit.

While each of these organizations seeks to foster collaboration quite broadly across the three groups, there has been some notable success at the level of individual projects, such as those described in Kuhn et al. (2020). It is noteworthy that this effort, specifically, was quite amply funded, had the backing of the National Research Council of Canada, and was able to enlist experts from all three groups. This shows that given enough time, funding, and expertise, significant progress can be made in developing language technology for Indigenous languages, and as such it makes a strong "business case" for increased support to projects of this type. Clearly, though, this model of mass collaboration is not so easily extended to other contexts, especially in countries lacking a robust and well-funded research infrastructure. As such, the challenge of developing more flexible and sustainable models of collaboration in this area remains.

4 Recommendations

Building on the good work that has already been done to bridge the divide that exists between the three groups, we can provide several specific recommendations to further strengthen these relationships.

4.1 Documentary Linguists and Indigenous Communities

The issues of trust and access have been an ongoing theme in the literature on endangered language documentation (Burnette and Sanders, 2014; Meissner, 2018), and a variety of best practices have been developed to promote successful collaborations between documentary linguists and communities (Penfield et al., 2008; Thieberger, 2012; Austin, 2014; Austin and Sallabank, 2018). As such, we will focus our recommendations on the pairings involving computational linguists.

4.2 Documentary Linguists and Computational Linguists

The disciplinary divide between these two groups is as wide as perhaps any other within linguistics, broadly conceived. As we seek to move forward into an era of NLP-assisted language documentation and revitalization, it has become necessary for those who are working as, or training to become, documentary linguists to develop greater familiarity with computational linguistics.

While this remains difficult to achieve within one's graduate training, as noted above, gatherings such as ComputEL and the annual SIGUL meetings, as well as their respective proceedings, can be quite helpful, providing a forum for connecting with and learning from computational linguists who are already engaged in work with other endangered and/or under-resourced languages, and who are thus familiar with at least some of the concerns that are front of mind for documentary linguists and Indigenous communities.

However, it must be pointed out that the learning curve for documentary linguists moving into the realm of computational linguistics can be quite steep, especially when they have had no coursework in the area. Many (though by no means all) of the articles in those proceedings are not easily understood by those who are in the early stages of trying to learn how computational linguistics may be helpful to their work in documentation and revitalization. (We choose not to cite any specific papers here, not wishing to unduly single out any particular contributions.)

This type of impenetrability to outsiders, of course, is in no way unique to the literature on computational linguistics, but is rather a systematic and deeply-ingrained cultural practice within academia more broadly. In this particular instance, however, it does represent a missed opportunity to make the work of computational linguists more legible to documentary linguists (and, thereby, hopefully, Indigenous community members as well), especially when that is clearly in line with the stated goals of the groups organizing the conferences and publishing the proceedings.

One can imagine ways to make this research more easily interpretable. For instance, it might be possible to have an editorial committee composed of documentary linguists who can review submissions and highlight areas that need further exposition for non-specialists. These could then be addressed by edits to the paper made by the authors themselves, or perhaps by the inclusion of expository endnotes provided by the editors. From this, a set of authorial best practices for writing within this particular subfield may develop, helping to maximize the value of the research for its intended audiences.

There are clear logistical challenges to implementing such a system, aside from the extra workload it would impose on already overstretched academics. For instance, to make a complex 8 page article more understandable to non-specialists, it may be necessary to lengthen it to 10 or 12 pages, at which point it may exceed the page limits set by the conference organizers or publishers. Likewise, extra steps in editing will require a longer timeline to get from submission to publication.

In the end, it is a matter of the priorities of the conference organizers, the scientific committees and the proceedings' editors as to how they see their work best contributing to narrowing the gap between their target demographics.

More immediately helpful may be opportunities for documentary linguists to receive direct, hands-on training in the basics of computational linguistics and NLP development. This training should have three tangible benefits:

First, it should help documentary linguists to understand the benefits that computational approaches may hold for them in their own work, e.g. addressing the transcription bottleneck through the development of ASR applications (Amith et al., 2021), as well as the potential limitations of such approaches (Prud'hommeaux, 2021).

Second, they should develop greater familiarity with how pedagogically-oriented language technology (e.g. Spaced-Repetition vocabulary learning systems, automated quizzes, I-CALL (Intelligent Computer-Assisted Language Learning) applications) are developed (Zhang et al., 2022), and may be incorporated in revitalization efforts (Lewis 2023).

Third, this training should allow the documentary linguists to prioritize the areas of NLP they wish to learn about, and which areas they wish to leave for collaborations with computational linguists with a specialization in that area.

While some training opportunities in this area exist – such as some of the courses at ESSLII (European Summer School in Logic, Language and Information) or at the Linguistic Society of America's Summer Institutes – they are not normally targeted specifically to documentary linguists, and do not take into account their particular needs. This type of customized training is an area where some of these newer organizations such as SIGEL and SIGUL could take the lead, building on their existing networks in order to facilitate collaboration between linguists of different stripes. Indeed, initial planning is now underway for a series of SIGEL-sponsored online training workshops in various aspects of NLP aimed specifically at documentary linguists, providing an additional forum where these two groups can come together. Opportunities such as these should help to broaden the impact of groups such as SIGEL and SIGUL beyond conferences and publications.

Lastly, the challenge of data paucity remains relatively intractable, although some efforts at faster, larger-scale language documentation are being developed (e.g. Boerger and Stutzman, 2018; Moe, 2023). Here, the challenge may lie with the computational linguists to sharpen their skills and be able to do more with less data, including finding ways to use data from majority languages to support the development of tools for Indigenous languages. Progress is being made in this area on a number of fronts (Harrigan et al., 2021, Yadav et al., 2022), giving hope that the smaller-sized corpora of Indigenous languages may not always be such a disadvantage when it comes to NLP application development.

4.3 Indigenous Communities and Computational Linguists

The proceedings of ComputEL and SIGUL, among other venues, have provided computational linguists the opportunity to learn more about the needs of language communities, as well as some of the challenges they face in their efforts to document and revitalize their languages. Since most computational linguists have little opportunity for in-community work, this burgeoning literature serves an important function of making the concerns of the language communities more apparent for computational linguists.

Unfortunately, the reverse is not true – there is not currently a readily accessible way for Indigenous language communities to become more educated on language technology, NLP development, and the potential value of computational linguistics to language revitalization efforts.

This leaves communities at a (further) disadvantage, in essence removing the option of developing such tools as part of their revitalization strategy. While the benefit of various NLP applications to community-based revitalization is an open question worthy of continued investigation (Liu et al., 2020), it is clearly problematic that most communities do not presently even have the option to consider how their on-going work could feed into the development of such applications, or how such tools might support their longer-term aspirations.

This lack of awareness and access can have further consequences as communities attempt to navigate through the language technological landscape. By now, it is a familiar story to hear about communities who have invested large sums of time and money (neither of which they have in abundance) into working with an outside company to develop a language app. While the value of seeing your language in digital form and being able to access information about it on your phone should not be underestimated, it is also clear that many of these apps have limited pedagogical value, and frequently leave the community with on-going maintenance costs. (This can be contrasted with the approaches from organizations such as 7000 Languages (https://7000.org), which seek longer-term and more collaborative approaches to community language app development.)

As such, training community members to be discerning developers and consumers of language technology is an important step in the process of providing communities the "central role in the design and implementation of language tools" that Liu et al. (2022) call for.

One potential model for such training can be found at CILLDI, the Canadian Indigenous Languages and Literacy Development Institute at the University of Alberta (https://uab.ca/cilldi). They offer a technology-focused course as part of the Community Linguist Certificate program, a six-course sequence designed to equip Indigenous students with the tools necessary to guide revitalization efforts in their own communities.

In past years, this course focused on the use of recording equipment, basic audio and video editing, and best practices in metadata and archiving, as these were essential technological skills needed by community members seeking to carry out documentation on their own languages. Over time, with the further spread of technology into Indigenous communities, more and more community members (typically though not exclusively from the younger generations) have learned many of these skills already, making it less useful to have a course that focuses solely on those basic activities.

This has allowed CILLDI to broaden the scope of the course to address key questions related to language technology. These include: What is the relationship between language documentation and NLP? What types of NLP applications are available for endangered languages? Which of them are relatively simple and can be developed from existing resources in the community, and which require more time and effort to create and maintain? What is the revitalization value of such applications (either in streamlining the documentary process, or in supporting language teaching and learning)? How can communities balance the costs (time, money, speaker availability) with the perceived benefits as part of their language revitalization plan?

While CILLDI offers this training in the context of a certificate program through a university, it is not hard to imagine more flexible models of delivering the same training that would have lower costs and potentially reach a wider audience, whether that be through community-based workshops, webinars, or open-access learning modules hosted on a website. This will be key in order to make such information more accessible to communities in other regions of the world.

Whatever the format, though, providing training opportunities of this kind for community members is essential to enabling communities to take the lead in decisions on the types of language technology that are appropriate for them, regardless of the priorities of any non-Indigenous companies or institutions they may be working with.

4.4 Summary

It is important that documentary linguists be able to learn about the development of NLP applications, and how they can aid the documentation and revitalization efforts in Indigenous communities. In addition, community members themselves need to become more aware of the options available to them in NLP-assisted efforts at revitalization. Through these opportunities to share and learn together, computational linguists will gain a better understanding of the concerns and priorities of the Indigenous communities with respect to the work being carried out on their languages. All of this supports the overall goal of bringing these three groups closer together, and strengthening the relationships that serve as the foundation to this work.

5 Conclusion

In this paper, we have looked at the relationship between three groups: computational linguists, documentary linguists, and Indigenous communities. These groups have distinct yet overlapping interests when it comes to the development and deployment of language technology. The challenge over the years has been to find ways for these three groups to work together better.

As in all relationships, communication and respect are the keys to understanding and trust. This can be clearly seen in the improvements in the working relationships between Indigenous communities and documentary linguists over the past several decades. By making the effort to better understand each other's needs and perspectives, the two groups have been able to make progress toward more respectful and equitable relationships, thus better enabling the documentary work that provides the basis for any computational applications.

A greater challenge has remained in building similarly productive relationships with computational linguists. Initiatives created by organizations such as CILLDI, CoLang, ComputEL, SIGEL, SIGUL, and others, have begun to bridge the gap in understanding between documentary linguists and Indigenous communities on the one hand, and computational linguists on the other. However it is clear that there is still a long way to go in strengthening these relationships.

Expanding opportunities for documentary linguists and Indigenous community members to learn more about computational linguistics, the diversity of NLP applications, and the potential value of such technology in supporting language revitalization is an urgent concern if much progress is going to be made in the coming years, before even more languages fall silent. As we make our way through the International Decade of Indigenous Languages (https://www.unesco.org/en/decades/indigenous-languages), it is imperative that more individuals and organizations step up to create these types of opportunities for awareness-building and skills-training.

In the long run, it is clear that training Indigenous people to be linguists, programmers and developers who can create applications for their own languages is the ideal solution. Indeed, recent years have seen more Indigenous people pursuing these career paths, to the great benefit of each of these fields (e.g.https://natives4linguistics.wordpress.com/). For too many Indigenous students, though, these options remain out of reach, and the immediate needs of their communities and their languages often put these pursuits on the backburner.

Language revitalization will always be a multi-generational societal project, but the process can be accelerated by the thoughtful development and deployment of NLP applications. As such, we are collectively obliged to do the critical work to strengthen the relationships between these three groups, for the benefit of current and future generations.

6 Limitations

This position paper is limited by the available resources in the scholarly discourse of this topic, and the professional experience the authors have had in working with members of all three groups highlighted in this paper.

References

Amith, Jonathan D., and Shi, Jiatong, and Castillo García, Rey. 2021. End-to-end automatic speech recognition: Its impact on the workflow in documenting Yoloxóchitl Mixtec. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 64–80, Online. Association for Computational Linguistics.

Arikpo, Iwara and Dickson, Iniobong. 2018. Development of an automated English-to-local-language translator using natural language processing. *International Journal of Scientific and Engineering Research*, 9:378-383.

Austin, Peter K. 2014. Language documentation in the 21st century. *JournaLIPP*, 3:57-71.

Austin, Peter K. and Sallabank, Julia. 2018. Language documentation and language revitalization: Some methodological considerations. *The Routledge handbook of language revitalization*, pages 207-215. Routledge.

Besacier, Laurent, and Barnard, Etienne, and Karpov, Alexey, and Schultz, Tanja. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85-100.

Bird, Steven. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, 3504-3519. International Committee on Computational Linguistics.

Boerger, B. H. and Stutzman, V. 2018. Single-event rapid word collection workshops: Efficient, effective, empowering. *Language Documentation and Conservation*, 12:147-193.

Burnette, Catherine and Sanders, Sara. 2014. Trust development with Indigenous communities in the United States. *The Qualitative Report*, 19:1-19.

Cadotte, Antoine, and Ngoc, Tan and, Boivin, Mathieu and, Sadat, Fatiha. 2022. Challenges and perspectives for Innu-Aimun within indigenous language technologies. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 99-108.

Czaykowska-Higgins, Eva. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian Indigenous communities. *Language Documentation and Conservation* 3(1):15-50.

Fitzgerald, Colleen. 2021. A framework for language revitalization and documentation. *Language*, 97(1):e1-e11.

Harrigan, Atticus G., and Antti Arppe 2021. Leveraging English word embeddings for semi-automatic semantic classification in nêhiyawêwin (Plains Cree). In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas (NAACL-HLT 2021)*, 1:113-121. doi: https://aclanthology.org/2021.americasnlp-1.12/

Kuhn, Roland, and Davis, Fineen, and Désilets, Alain, and Joanis, Eric, and Kazantseva, Anna, and Knowles, Rebecca, and Littell, Patrick, and Lothian, Delaney, and Pine, Aidan, and Wolf, Caroline, and Santos, Eddie, and Stewart, Darlene, and Boulianne, Gilles, and Gupta, Vishwa, and Owennatékha, Brian, and Martin, Akwiratékha', and Cox, Christopher, and Junker, Marie-Odile, and Sammons, Olivia, and Souter, Heather. 2020. The Indigenous languages technology project at NRC Canada: An empowerment-oriented approach to developing language software. In *Proceedings of the 28th International Conference on Computational Linguistics*, 5866-5878.

Khawaja, Masud. 2021. Consequences and remedies of Indigenous language loss in Canada. *Societies*, 11(89) https://doi.org/10.3390/soc11030089

Leonard, Wesley Y. 2018. Reflections on (de) colonialism in language documentation. *Reflections on language documentation 20 years after Himmelmann 1998*, pages 55-65.

Lewis, Robert. 2023. "A Survey of computational infrastructure to help preserve and revitalize Bodwéwadmimwen". Forthcoming in *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Littell, Patrick, and Kazantseva, Anna, and Kuhn, Roland, and Pine, Aidan, and Arppe, Antti, and Cox, Christopher, and Junker, Marie-Odile. 2018. Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, 2620-2632. Santa Fe, New Mexico: Association of Computational Linguistics. Retrieved from: https://www.aclweb.org/anthology/C18-1222

Liu, Zoey, and Richardson, Crystal, and Hatcher Jr, Richard, and Prud'hommeaux, Emily. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 1:3933 - 3944.

Meissner, Shelbi. 2018. The moral fabric of linguicide: un-weaving trauma narratives and dependency relationships in Indigenous language reclamation. *Journal of Global Ethics*, 14:266-276.

Moe, Ronald. 2007. Dictionary development program. *SIL Forum for Language Fieldwork*, 3:55-65.

Mohanty, Sushree Sangita, and Parida, Shantipriya, and Dash, Satya Ranjan. 2023. Role of NLP for corpus development of endangered languages. *Grenze International Journal of Engineering and Technology*, Jan Issue, 1318-1323.

Ortiz-Rodríguez, Fernando, and Mishra Tiwari, Sanju, and Panchal, Ronak, and Medina-Quintero, Jose-Melchor, and Barrera, Ruben. 2022. MEXIN: Multidialectal ontology supporting NLP approach to improve government electronic communication with the Mexican Ethnic Groups. *The 23rd Annual International Conference on Digital Government Research* pages 461-463.

Paksi, Attila, and Kivinen, Ilona. 2021. Reflections on power relations and reciprocity in the field while conducting research with Indigenous peoples. In *Indigenous Research Methodologies in Sámi and Global Contexts* pages 201-228. Brill.

Penfield, Susan D., and Serratos, Angelina, and Tucker, Benjamin V., and Flores, Amelia, and Harper, Gilford, and Hill Jr, Johnny, and Vasquez, Nora. 2008. Community collaborations: Best practices for North American Indigenous language documentation. *International Journal of the Sociology of Language*. 191:187-202.

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

Restoule, Jean-Paul. 2008. The values carry on: Aboriginal identity formation of the urban-raised generation. *The Canadian Journal of Native Education*, 31:15-33.

Sag, Ivan A., and Baldwin, Timothy, and Bond, Francis, and Copestake, Ann, and Flickinger, Dan. 2002. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico,* *February 17–23, 2002 Proceedings 3*, pages 1-15. Springer Berlin Heidelberg.

Schwartz, Lane. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), pages 724-731.

Seifart, Frank, and Evans, Nicholas, and Hammarström, Harald, and Levinson, Stephen. 2018. Language documentation twenty-five years on. *Language*, 94:e324-e345.

Tessaro, Danielle, and Restoule, Jean-Paul, and Gaviria, Patricia, and Flessa, Joseph, and Lindeman, Carlana, and Scully-Stewart, Coleen. 2018. The five R's for indigenizing online learning: A case study of the First Nations schools' principals course. *Canadian Journal of Native Education*, 40(1):125-143.

Thieberger, Nicholas (Ed.). 2012. *The Oxford Handbook of Linguistic Fieldwork.* Oxford University Press.

Yadav, Hemant and Sitaram, Sunayana. 2022. A survey of multilingual models for automatic speech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5071–5079, Marseille, France. European Language Resources Association.

Van Gysel, Jens E., and Vigus, Meagan, and Chun, Jayeol, and Lai, Kenneth, and Moeller, Sarah, and Yao, Jiarui, and O'Gorman, Tim, and Cowell, Andrew, and Croft, William, and Huang, Chu-Ren, and Hajič, Jan, and Martin, James H., and Oepen, Stephan, and Palmer, Martha, and Pustejovsky, James, and Vallejos, Rosa, and Xue, Nianwen. 2021. Designing a uniform meaning representation for natural language processing. *KI-Künstliche Intelligenz*, 35(3-4):343-360.

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