Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism

Anonymous ACL submission

Abstract

In this paper, we challenge the ACL community to reckon with historical and ongoing colonialism by adopting a set of ethical obligations and best practices drawn from the Indigenous studies literature. While the vast majority of NLP research focuses on a very small number of very high resource languages (English, Chinese, etc), some work has begun to engage with Indigenous languages. No research involving Indigenous language data can be considered ethical without first acknowledging that Indigenous languages are not merely very low resource languages. The toxic legacy of colonialism permeates every aspect of interaction between Indigenous communities and outside academic researchers. Ethical research must actively challenge this colonial legacy by explicitly acknowledging and centering Indigenous community goals and Indigenous ways of knowing. To this end, we propose that the ACL draft and adopt an ethical framework for NLP researchers and computational linguists wishing to engage in research involving Indigenous languages.

1 Introduction

Beginning with our community's first academic conference in 1952 (see Reifler, 1954) and continuing with the establishment of the Association for Computational Linguistics (ACL)[1] in 1962 (MT Journal, 1962), the members of our research community have examined a huge range of topics, ranging from linguistic and computational linguistic models and theories to engineering-focused problems in natural language processing.[2]

While great progress has been made in recent years across many NLP tasks, the overwhelming majority of NLP and CL research focuses on a very small number of languages. Over the 70 years from 1952 to 2022, the vast majority of CL and NLP research has focused on a small number of widely-spoken languages, nearly all of which represent politically- and economically-dominant nation-states and the languages of those nation-states' historical and current adversaries: English, the Germanic and Romance languages of western Europe, Russian and the Slavic languages of eastern Europe, Hebrew, Arabic, Chinese, Japanese, and Korean. Bender (2009) surveyed papers from ACL 2008 and found that English dominated (63% of papers), with 20 other languages distributed along a long Zipfian tail (Chinese and German shared the number 2 slot at just under 4% of papers each); across all ACL 2008 long papers, only three languages (Hindi, Turkish, and Wambaya) were represented outside of the language families listed previously. This lack of diversity directly impacts both the quality and ethical status of our research, as nearly every successful NLP technique in widespread current use was designed around the linguistic characteristics of English.[3]

A special theme designed to address this shortcoming has been selected for the 60th Annual Meeting of the ACL in 2022: *"Language Diversity: from Low Resource to Endangered Languages."* This theme is to be commended as a step towards a more linguistically diverse research agenda. Yet as we expand our research to a broader and more inclusive set of languages, we must take great care to do so ethically. The endangered Indigenous languages of the world are not merely very low resource languages. The toxic legacy of colonialism permeates every aspect of interaction between Indigenous communities and outside academic re-

[1] Originally founded as the Association for Machine Translation and Computational Linguistics, the current name stems from 1968 after the publication of the 1966 ALPAC report.

[2] See Linguistic Issues in Language Technology (2011) and Eisner (2016) for excellent discussions on the distinction between computational linguistics (CL) and natural language processing (NLP).

[3] A small minority of successful NLP techniques were designed taking into account the characteristics of a few other languages, nearly all from the Indo-European and Sino-Tibetan language families.

searchers (Smith, 2012). Ethical research must actively challenge this colonial legacy by explicitly acknowledging and centering Indigenous community goals and Indigenous ways of knowing.

To this end, we propose an ethical framework for NLP researchers and computational linguists wishing to engage in research involving Indigenous languages. We begin in §2 by examining the abstracts of *ACL papers from the past several years, replicating the results of Bender (2009), confirming that recent *ACL papers still lack significant language diversity. In §3 we address research practices and ongoing colonialism in Indigenous communities. Finally, we examine decolonial practices appropriate for a draft framework of ethical obligations (§4) for the ACL research community.

2 Recent *ACL papers lack significant language diversity

We begin by examining the abstracts of *ACL papers from the past several years to confirm the results of Bender (2009), namely that recent *ACL papers still lack significant language diversity. We collect a corpus of 9602 recent *ACL abstracts from the ACL Anthology;[4] more than 80% fail to mention any language. Essentially all such papers that fail the #BenderRule assume English as the language of study (Bender, 2019). Vanishingly few abstracts mention any Indigenous language. While 66 abstracts mention Arabic, fewer than 20 abstracts mention any other African language. Only 11 abstracts mention any Indigenous language of North America. Only 2 abstracts mention an Indigenous language of Australia. Only 1 abstract mentioned an Indigenous language of Zealandia. No abstracts mentioned any Indigenous language of South America.

Table 1 shows a Zipfian distribution predominated by four language families: Indo-European (dominated by English), Sino-Tibetan (dominated by Mandarin Chinese), Japonic (essentially all Japanese), and Afro-Asiatic (dominated by Arabic and Hebrew). Indo-European languages are assumed (English) or explicitly mentioned in 97% of abstracts. The next three most mentioned language families account for another 1% of abstracts.[5] Combined, only 165 out of 9602 abstracts (1.7%) mention any language from any other language family.

83.26%	7995	Implictly assume English
13.70%	1315	Indo-European (incl. English)
4.50%	432	Sino-Tibetan
1.12%	108	Japonic
0.85%	82	Afro-Asiatic
0.41%	39	Turkic
0.26%	25	Koreanic
0.25%	24	Austroasiatic
0.24%	23	Dravidian
0.22%	21	Uralic
0.21%	20	Austronesian
0.09%	9	Basque
0.09%	9	Atlantic-Congo
0.07%	7	Na-Dene
0.05%	5	Kra-Dai
0.02%	2	Arnhem
0.02%	2	Iroquoian
0.02%	2	Inuit-Yupik-Unangan
0.01%	1	Sumerian

Table 1: Of 9602 *ACL abstracts (2013–Nov. 2021),[4] percentage and number of abstracts that explicitly mention at least one language from the language family.

3 Research and Ongoing Colonialism in Indigenous Communities

Endangered Indigenous languages are not merely very low-resource languages. Each Indigenous community represents a sovereign political entity. Each Indigenous language represents a crucial component of the shared cultural heritage of its people. The rate of intergenerational transmission of Indigenous language from parent to child in many Indigeneous communities has declined and is continuing to decline (Norris, 2006), resulting in a deep sense of loss felt by older generations who grew up speaking the Indigenous language as well as by younger generations who do not speak the language who experience a diminished sense of cultural inclusion (Tulloch, 2008). Language is an integral part of culture, and declines in robust Indigenous language usage have been correlated with serious negative health and wellness outcomes (Chandler and Lalonde, 2008; Reid et al., 2019).

[4] Since 2013, the ACL Anthology has included abstracts for TACL papers. Since 2017, the ACL Anthology has included abstracts for papers published at ACL, EACL, AACL, NAACL, EMNLP, and the Comptuational Linguistics journal. See Appendix A for details.

[5] Note that this is less than the percentages for these three language families listed in Table 1. This is because some abstracts mention multiple languages. This additional 1% represents abstracts that mentioned a language from the Sino-Tibetan, Japonic, or Afro-Asiatic language families and did not also mention an Indo-European language such as English.

At the same time, Indigenous individuals and Indigenous communities have suffered greatly from colonial practices that separated children from communities, actively suppressed Indigenous language and culture, misappropriated land and natural resources, and treated Indigenous people, cultures, and languages as dehumanized data to study (Whitt, 2009; NTRC, 2015; Leonard, 2018; Bull, 2019; Dei, 2019; Guematcha, 2019; Bahnke et al., 2020; Kawerak, 2020). As Smith (2012) notes, "*research* is probably one of the dirtiest words in the indigenous world's vocabulary;" it is "implicated in the worst excesses of colonialism" and "told [Indigenous people] things already known, suggested things that would not work, and made careers for people who already had jobs." It is then, hardly surprising that "After generations of exploitation, Indigenous people often respond negatively to the idea that their languages are data ready for the taking" (Bird, 2020).

Outside perceptions of Indigenous peoples are inextricably linked to corresponding histories of colonization, and are typically accompanied by (usually outdated and incorrect) assumptions about the "proper" roles of Indigeneous peoples today that correspond with neither reality nor Indigenous people's views of themselves (Deloria, 2004; Leonard, 2011). When a linguist (or a computer scientist) begins the process of interacting with an Indigenous community and working with that community's Indigenous language, the starting "lens through which others view [the linguist's] professional activities will at least partly reflect what 'linguist' has come to mean, and that this in some cases will occur regardless of whether [the linguist] personally exhibit a trait that has come to be associated with this named position" (Leonard, 2021).

Given the distinct value systems and distinct views of reality of research scientists and Indigenous communities, it is not surprising that even good-faith efforts of well-meaning outside researchers are often viewed by Indigenous communities as irrelevant at best and exploitative at worst. Research scientists rarely consider the philosophy of science (Popper, 1959) on which our research is predicated; as Wilson (2001) notes, this is defined by an ontology, epistimology, methodologies, and axiology that are seldom acknowledged. In our field, these often surface as unacknowledged positivist (Comte, 1853) assumptions that science is value-neutral and empirical observations and logical reasoning fully and completely define the nature of science and reality (Egan, 1997). By failing to acknowledge and critically examine these philosophical foundations, we implicitly and unconsciously elevate our ideas of research and language work above those of Indigenous communities (Leonard, 2017). Indigenous communities are rightly taking up the slogan "Nothing about us without us" (see, for example, Pearson, 2015). Even when we consider the "lived experiences and issues that underlie [the] needs" of Indigenous communities, these community priorities are typically treated as subordinate to research questions deemed valuable by members of academe (Leonard, 2018; Wilson, 2008; Simonds and Christopher, 2013). Credulous evangelical claims of technology as savior[6,7] only exacerbate these tensions.

4 Prerequisite Obligations for Ethical Research involving Indigenous Languages and Indigenous Peoples

When CL and NLP researchers begin to work with Indigenous language data without first critically examining the toxic legacy of colonialism and the self-identified priority needs and epistemology of the Indigenous community, the risk of unwittingly perpetuating dehumanizing colonial practices is extremely high. It is therefore critically urgent that the ACL, perhaps through the recently-formed Special Interest Group on Endangered Languages (SIGEL), should begin a process of drafting and adopting a formal ethics policy with respect to Indigenous communities, Indigenous languages, and Indigenous data. In doing so, we should draw upon the recent Linguistics Society of America (2019) ethics statement, the foundational principles of medical ethics (autonomy, non-maleficence, beneficence, and justice; Beauchamp and Childress, 2001), the recommendations of Bird (2020), and the wisdom of Indigenous scholars such as Deloria, Wilson, Smith, and Leonard.

As a beginning, we have identified four key ethical obligations that should at a minimum be included in such an ethics policy: cognizance, beneficence, accountability, and non-maleficence.

[6] "The number of endangered languages is so large that their comprehensive documentation by the community of documentary linguists will only be possible if supported by NLP technology." (Vetter et al., 2016)

[7] "Languages that miss the opportunity to adopt Language Technologies will be less and less used, while languages that benefit from cross-lingual technologies such as Machine Translation will be more and more used." (ELRA, 2019)

4.1 Obligation of cognizance

The colonial political and racial ideas and behaviors that support and enable colonization and oppression are intentionally invented historical creations (Allen, 2012; Kendi, 2017). Before we engage with Indigenous peoples, let alone work with Indigenous data, we must intentionally make ourselves cognizant of this history. As outside academic researchers, we stand in a privileged position, and as such have an urgent obligation to educate ourselves about this history and about current practices that perpetuate these systems of oppression in the present day (Kendi, 2019; Smith, 2012).

Before we are capable of ethically engaging with Indigenous data, we must learn the ways in which Indigenous communities approach reality and science, and accept that these are fully formed and fully valid worldviews with which we have an obligation to fully engage. Our research is premised on a particular philosophy of science which is nearly always left unstated. We must make ourselves cognizant of our own ontology, epistemology, methodology, and axiology, and the fact that there are alternative philosophies of science that are equally valid. We must educate ourselves about Indigenous ontologies, epistemologies, methodologies, and axiologies that are centered around relationality (Wilson, 2008).

The *obligation of cognizance* therefore mandates that we as researchers intentionally and thoroughly educate ourselves about colonization of Indigenous communities; about the role that academic researchers have had and continue to play in the exploitation of Indigenous communities, Indigenous languages, Indigenous culture, and Indigenous data; and about Indigenous expectations and ways of being centered on relationality that differ from those we typically encounter in our research.

4.2 Obligation of beneficence

Indigenous communities are sovereign political entities with inherent political and human rights which are enumerated in the Declaration on the Rights of Indigenous Peoples (United Nations, 2007). This includes the right to protect and develop their culture (Article 11), the right to dignity (Article 15), the right to develop and elect their own decision-making institutions (Article 18), and the right to "maintain, control, protect, and develop their intellectual property over [their] cultural heritage, traditional knowledge, and traditional cultural expressions" (Article 31).

The *obligation of beneficence* therefore mandates that we as researchers ensure that our work benefits the Indigenous communities with which we work in ways that those communities recognize as beneficial; we as outside researchers seeking to engage with Indigenous communities have an obligation to learn about and to meaningfully support priority areas identified by Indigenous governing bodies and decision-making institutions that fall within our respective scopes of expertise.

4.3 Obligation of accountability

As outside researchers seeking to work with Indigenous data, we have a responsibility to seek out respectful and meaningful relationships with the Indigenous communities whose data we seek to use. We have a responsibility to develop these relationships in ways that are appropriate and meaningful to the Indigenous communities with which we seek to work. We must intentionally acknowledge and accept the rightful authority of Indigenous communities' governing and decision-making bodies over those communities' own respective languages, cultures, and data.

The *obligation of accountability* therefore mandates that we as researchers develop meaningful relations with the sovereign governing bodies of the Indigenous communities with which we seek to engage, and that we be meaningfully accountable to such bodies in our work involving their data.

4.4 Obligation of non-maleficence

Colonization and colonial practices have inflicted substantial and often genocide-scale harm on Indigenous communities over the past five centuries (Smith, 2017), harm that is ongoing and is often perpetuated by modern research practices. We must intentionally adopt the prime ethical directive of the medical community, often stated in the Latin aphorism *Primum Non Nocere* "Above all, do no harm" (Smith, 2005). There are many good and laudable reasons why we should choose to engage in research with Indigenous communities; but none of these reasons is powerful enough to justify harm to these communities caused by our research.

The *obligation of non-maleficence* therefore mandates that above all else, we do no harm to Indigenous people and Indigenous communities; if we can do good through our research without doing harm, that is well, but it is better to not engage than to cause harm.

References

Theodore W. Allen. 2012. *The Invention of the White Race*. Verso Books. Two volumes.

ALPAC. 1966. Language and machines — computers in translation and linguistics. A Report by the Automatic Language Processing Advisory Committee.

Melanie Bahnke, Vivian Korthuis, Amos Philemonoff, and Mellisa Johnson. 2020. Navigating the New Arctic NSF Comment Letter.

Tom L. Beauchamp and James F. Childress. 2001. *Principles of Biomedical Ethics*. Oxford University Press, New York.

Emily Bender. 2019. The #BenderRule: On naming the languages we study and why it matters. *The Gradient*.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Julie Bull. 2019. Nothing about us without us: An Inuk reply to exploitive research | Impact Ethics. Memorial University Centre for Bioethics.

Michael J. Chandler and Christopher E. Lalonde. 2008. Cultural continuity as a protective factor against suicide in first nations youth. *Horizons*, 10(1):68–72. Special Issue — Hope or Heartbreak: Aboriginal Youth and Canada's Future.

Auguste Comte. 1853. *The Positive Philosophy of Auguste Comte*. John Chapman, London. Condensed and translated from the original *Cours de Philosophie Positive* (1830–1842) by Harriet Martineau.

George J. Sefa Dei. 2019. Foreword. In *Decolonization and Anti-colonial Praxis*, volume 8 of *Anti-colonial Educational Perspectives for Transformative Change*, pages vii – x. Brill, Leiden, The Netherlands.

Philip J. Deloria. 2004. *Indians in Unexpected Places*. University Press of Kansas, Lawrence.

Kieran Egan. 1997. *The Educated Mind: How Cognitive Tools Shape Our Understanding*. University of Chicago Press.

Jason Eisner. 2016. How is computational linguistics different from natural language processing? Quora.

ELRA. 2019. Lt4all: Language technologies for all – call for participation.

Emmanuel Guematcha. 2019. Genocide against indigenous peoples: The experiences of the truth commissions of Canada and Guatemala. *The International Indigenous Policy Journal*, 10(2).

Kawerak. 2020. Knowledge Sovereignty and the Indigenization of Knowledge.

Ibram X. Kendi. 2017. *Stamped From The Beginning: The Definitive History of Racist Ideas in America*. Bold Type Books.

Ibram X. Kendi. 2019. *How to Be an Antiracist*. One World.

Wesley Y. Leonard. 2011. Challenging "extinction" through modern miami language practices. *American Indian Culture and Research Journal*, 35(2):135–160.

Wesley Y. Leonard. 2017. Producing language reclamation by decolonising 'language'. *Language Documentation and Description*, 14:15–36.

Wesley Y. Leonard. 2018. Reflections on (de)colonialism in language documentation. In Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, editors, *Reflections on Language Documentation 20 Years after Himmelmann 1998*, Special Publication 15, chapter 6, pages 55–65. Language Documentation & Conservation.

Wesley Y. Leonard. 2021. Centering Indigenous ways of knowing in collaborative language work. In Lisa Crowshoe, Inge Genee, Mahaliah Peddle, Joslin Smith, and Conor Snoek, editors, *Sustaining Indigenous Languages: Connecting Communities, Teachers, and Scholars*, pages 21–34. Northern Arizona University.

Linguistic Issues in Language Technology. 2011. Interaction of linguistics and computational linguistics. Volume 6.

Linguistics Society of America. 2019. LSA revised ethics statement.

Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, editors. 2008. *Proceedings of ACL-08: HLT*. Association for Computational Linguistics, Columbus, Ohio.

MT Journal. 1962. Professional society formed. *Mechanical Translation*, 7(1):1.

Mary Jane Norris. 2006. Aboriginal languages in Canada: Trends and perspectives on maintenance and revitalization. In Jerry P. White, Susan Wingert, Dan Beavon, and Paul Maxim, editors, *Aboriginal Policy Research*, volume 3: Moving Forward, Making a Difference, pages 197–226. Thompson Educational Publishing, Toronto.

NTRC. 2015. Honouring the truth, reconciling for the future: Summary of the final report of the Truth and Reconciliation of Canada.

Luke Pearson. 2015. Nothing about us, without us. that's why we need Indigenous-owned media. *The Guardian*.

Karl Popper. 1959. *The Logic of Scientific Discovery*. Hutchinson & Co.

Papaarangi M. Reid, Donna M. Cormack, and Sarah-Jane Paine. 2019. Colonial histories, racism and health — the experience of Māori and Indigenous peoples. *Public Health*, 172:119–124. Special issue on Migration, Ethnicity, Race and Health.

Erwin Reifler. 1954. The first conference on mechanical translation. *Mechanical Translation*, 1(2):23–32.

Vanessa W. Simonds and Suzanne Christopher. 2013. Adapting western research methods to indigenous ways of knowing. *American Journal of Public Health*, 103(12):2185–2192.

Cedric M. Smith. 2005. Origin and uses of primum non nocere — above all, do no harm! *The Journal of Clinical Pharmacology*, 45(4):371–377.

David Michael Smith. 2017. Counting the dead: Estimating the loss of life in the Indigenous Holocaust, 1492 – present. In *Proceedings of the Twelfth Native American Symposium*.

Linda Tuhiwai Smith. 2012. *Decolonizing Methodologies: Research and Indigenous Peoples*, 2nd edition. Zed Books.

Shelley Tulloch. 2008. Uqausirtinnik annirusunniq — Longing for our language. *Horizons*, 10(1):73–76. Special Issue — Hope or Heartbreak: Aboriginal Youth and Canada's Future.

United Nations. 2007. United Nations declaration on the rights of Indigenous peoples.

Marco Vetter, Markus Müller, Fatima Hamlaoui, Graham Neubig, Satoshi Nakamura, Sebastian Stüker, and Alex Waibel. 2016. Unsupervised phoneme segmentation of previously unseen languages. In *17th Annual Conference of the International Speech Communication Association (InterSpeech 2016)*, San Francisco, California, USA.

Laurelyn Whitt. 2009. *Science, Colonialism, and Indigenous Peoples: The Cultural Politics of Law and Knowledge*. Cambridge University Press.

Shawn Wilson. 2001. What is an Indigenous Research Methodology? *Canadian Journal of Native Education*, 25(2):175–179.

Shawn Wilson. 2008. *Research Is Ceremony: Indigenous Research Methods*. Fernwood Publishing.

A *ACL abstract corpus 2013–Nov. 2021

The *ACL XML files (2013–2021) from the ACL Anthology GitHub repository were downloaded on 6 November 2021.

2013.tacl.xml
2014.tacl.xml
2015.tacl.xml
2016.tacl.xml
2017.acl.xml
2017.cl.xml
2017.eacl.xml
2017.emnlp.xml
2017.tacl.xml
2018.acl.xml
2018.cl.xml
2018.emnlp.xml
2018.naacl.xml
2018.tacl.xml
2019.acl.xml
2019.cl.xml
2019.emnlp.xml
2019.naacl.xml
2019.tacl.xml
2020.aacl.xml
2020.acl.xml
2020.cl.xml
2020.emnlp.xml
2020.tacl.xml
2021.acl.xml
2021.eacl.xml
2021.emnlp.xml
2021.naacl.xml
2021.tacl.xml

The abstracts were extracted from the XML files. From the resulting abstracts all words that begin with an uppercase letter were examined manually to identify all explicitly mentioned language names. All processing steps are described, with specific shell commands used, in the `data` annex that accompanies this paper.