Acta Borealia

A Nordic Journal of Circumpolar Societies

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/sabo20

Indigenous language technology in the age of machine learning

Sjur Nørstebø Moshagen, Lene Antonsen, Linda Wiechetek & Trond Trosterud

To cite this article: Sjur Nørstebø Moshagen, Lene Antonsen, Linda Wiechetek & Trond Trosterud (2024) Indigenous language technology in the age of machine learning, Acta Borealia, 41:2, 102-116, DOI: 10.1080/08003831.2024.2410124

To link to this article: https://doi.org/10.1080/08003831.2024.2410124

ACTA BOREALIA
2024, VOL. 41, NO. 2, 102–116
https://doi.org/10.1080/08003831.2024.2410124

Routledge
Taylor & Francis Group

OPEN ACCESS

Indigenous language technology in the age of machine learning

Sjur Nørstebø Moshagen, Lene Antonsen, Linda Wiechetek and Trond Trosterud

Department of Language and Culture, Faculty of Humanities, Social Sciences and Education, UiT – The Arctic University of Norway, Tromsø, Norway

ABSTRACT
Most modern language technology for proofing tools, machine translation and other applications is based on machine learning. However, very few Indigenous languages have the necessary amount of texts for making tools based on this technology. When most language technology is based on large language models (LLMs), it bears the risk of most of Indigenous language online text being produced by neural text generation. The result would be that online texts cannot be trusted as a source for authentic Indigenous languages anymore. An alternative is the work done at UiT – The Arctic University of Norway during the last 20 years, based on linguistics. Sámi language tools have been made available for both industry and language communities, with open licenses. These have been widely used by translators, teachers and various software companies. The article analyzes the following four parts of language technology development: language data, language tool development, making the tools available to users, and ethical use of available language technology tools. We make extensive use of the CARE principles, and discuss the shortcomings of existing software and data licensing schemes. Finally, we introduce a 3D table to help classify language technology projects with respect to their suitability for Indigenous languages.

ARTICLE HISTORY
Received 28 June 2024
Accepted 22 August 2024

KEYWORDS
Sámi languages; indigenous languages; language technology; community ownership; technology ownership; open access; data governance

Eamiálbmot giellateknologiija dihtoroahppan áigodagas

ABSTRÁKTA
Eanaš ođđaáigásaš giellateknologiija mii lea geavahusas nugo divvunprográmmain ja dihtorjorgaleamis, lea vuođđoduvvon dihtoroahppamii mii lea dahkkon hirbmat teakstačoakkáldagaiguin. Muhto hui hárve eamiálbmotgielain leat doarvái teavsttat hukset reaidduid dáinna teknologiijain. Molssaevttolaš čoavddus lea teknologiija mii lea vuođđoduvvon lingvistihkkii. Dađi bahábut software-buvttadeaddjit dađistaga leat dahkan váddáseabbon ja maiddái veadjemeahttumin oažžut dákkár reaidduid olámuddui giellageavaheddjiide go máŋgga oktavuođas buvttadeaddjit hehttejit goalmmátoasehasaid software leat olámuttos dahje vejolaš sajáiduhttit sin software bokte. Go eanaš giellateknologiija lea vuođđuduvvon stuorra giellamodeallaide, de mii sáhttit boahtit dán dillái ahte eanaš online eamiálbmotgielaid teavsttat leat ráhkaduvvon nevrála teakstagenereremiin, ja dáid gielaid divvunprográmmat leat fas vuođđuduvvon teakstačoakkáldagaide main leat generejuvvon teavsttat. Dalle ii sáhtáše šat luohttit eamiálbmotgielaid online-teavsttaide autenttalažžan. Oktan molssaeaktun teknologiijii vuođđuduvvon stuorra giellamodeallaide livččii bargu maid UiT leat dahkan sámegielaide maŋimuš 20 jagi. Sámi lingvisttalaš resurssat leat leamaš olámuttos sihke industriijii ja giellaservošiidda, dain leat rabas liseanssat, maiddái gávppálaš geavahussii. Resurssaid atnet máŋggalágan geavaheaddjit, nugo jorgaleaddjit, oahpaheaddjit ja software-fitnodagat. Artihkkalis mii muitalit giellateknologalaš ovdáneamis, ja fáttát leat gielladáhtat, giellareaidduid ovddideapmi, mo dahkat reaidduid olámuddui geavaheddjiide, ja maiddái giellateknologalaš reaidduid ehtihkalaš geavahus. Analysas mii čujuhit CARE-prinsihpaide, ja mii digaštallat dálá software ja dáhtáid liseanssaid váilevašvuođaid. Loahpas mii evttohit 3D-tabealla veahkkin klassifiseret giellateknologalaš prošeavttaid dan mielde mo dat heivejit eamiálbmotgielaide.

CONTACT Sjur Nørstebø Moshagen sjur.n.moshagen@uit.no Department of Language and Culture, Faculty of Humanities, Social Sciences and Education, UiT – The Arctic University of Norway, PO Box 6050 Stakkevollan, N-9037 Tromsø, Norway

Introduction

The recent development of large language models (LLMs) promises computer applications to produce fluent text for a wide range of tasks. Whereas this promise is increasingly met for English and other majority languages, for Indigenous and minority languages these applications produce seemingly fluent text, which upon closer inspection ranges from misleading output to utter gibberish. From our experience working with Sámi language technology during the past 20 years, we have seen that Indigenous languages have managed to produce necessary language tools using rule-based technologies *without* large amounts of data. The new scenario, however, uses methods that require large amounts of data, and opens the horizon for more powerful tools, on the one hand, but also produces bad output when data is sparse or of poor quality, on the other hand.

By and large, the main algorithmic code behind LLMs based upon neural networks is known. What is left for competition is the data used for training the actual models, the training configuration, and the hardware to train it on. For the commercial LLM providers (OpenAI, Google and others) the data is closed, inaccessible and not available for academic evaluation, comparison or investigation. At the same time the data is never produced by the entities building the models, it is taken from wherever it can be found, usually without asking.

The problem when it comes to Indigenous languages is not that the developers of LLMs do not have enough of the existing data. The problem is that there is not enough data in the first place. And the data scarcity issue is not going to vanish by itself. For most languages in the world there are too few speakers to produce enough text to build robust models for their language.

The data-driven approach is becoming increasingly problematic. On the one hand, the question of ownership poses itself. The relation between those who provide the data and the ones that use the data to make products is not in balance, neither economically nor ideologically. On the other hand, deep learning creates both machine translation systems and proofing tools based upon text collections rather than upon linguistic (i.e. grammatical and semantic) insight. The worst-case scenario is one where the major share of online text in Indigenous languages is produced by generative neural models, or neural machine translation and proofing tools are based upon noisy collections of text with varying quality. In addition, due to the frequent lack of evaluators that are able to correct the output, there is no reliable assessment available to the uninformed – meaning non-bilingual – users (Wiechetek, Pirinen, and Kummervold 2023).

The result will be that online text for Indigenous languages cannot be trusted anymore, it may as well be generated from a mixture of text written by inexperienced writers and even non-speakers, and in the worst case on machine-generated text based upon such input. Language standardization will be impossible when spellcheckers are created from text with many errors, thereby promoting erroneous forms to norm. The smaller the language is, the higher the chance that a substantial part of all available text will be error-prone and hard to understand.

In large parts of the world the digital infrastructure has become so deeply integrated with the whole society that one cannot do without – one has to use a computer or a mobile phone to get in touch with tax authorities, health care and so on. The digital infrastructure is like electric power or water pipes, it must be available to everybody on equal terms. In e.g. the Nordic countries this explicitly entails all or most Indigenous languages. Nevertheless, none of the de facto or de jure minority languages are served on an equal footing by the technology providers, compared to the services for the majority languages. In most cases, they are not served at all.

This article will discuss the dilemmas both LLMs and language technology development in general present minority language communities with. We will analyze LLMs and language technology development with respect to: (1) language data, (2) tool development, (3) making the tools accessible, and (4) use of language technology (LT) tools. Part of the analysis will be done by applying the CARE principles to the topic in question.

The second section discusses the background to the present situation. The third section is the main text, looking at language data, language technology tool development, access to computer systems for writing and reading text and discusses the ethics of Indigenous language technology use. At the end of that text we suggest a table of yes-no questions to help classify language technology projects with respect to their suitability for Indigenous languages. Finally comes our conclusion.

Background

Although Indigenous languages cover all continents, they are invisible in maps that indicate only state borders. This has an effect on our perception of existing languages in the world and often marginalizes the huge amount of language communities that are not equivalent to a nation state. In this article, we discuss the case of Indigenous languages in particular as the situation of language technology development for Indigenous languages differs from other languages. The multi-lingual infrastructure *GiellaLT*, used for many of the tools presented in this article, and discussed later, covers many of the Indigenous languages of the Arctic, cf. Figure 1, and most of the examples referred to – directly or via references – are taken from these languages.

Central to the content of this article is the assumption that a language community is or wants to be present in the digital world. While this is true for the Indigenous and minority languages in the Nordic countries, this does not need to be true or even a relevant question for many other language communities. As argued by Schwartz (2022), which route to take for the future of a language needs to be decided by the language community, no-one else. Thus, the issues and analyses presented here are principally only valid for the languages we have worked with, but should be more or less generalizable to all minority and Indigenous language communities who want a digital presence for their language.

Sámi language technology as an example case

In this article we will use the Sámi languages as an example case. They are not representative for Indigenous languages in general, they will rather give a picture of what can be achieved for relatively well-resourced Indigenous languages.

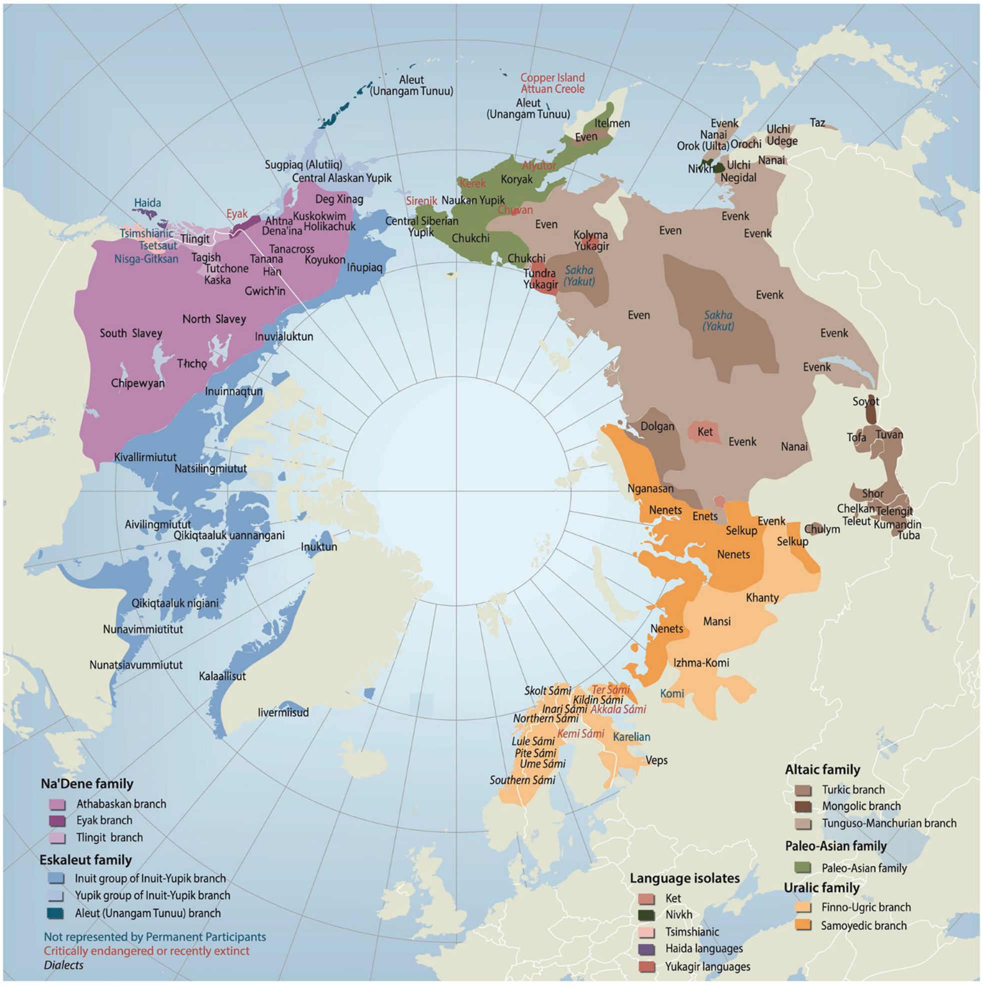

Figure 1. Arctic Indigenous Languages Map (2019). License: CC-BY-SA.

Sámi language technology in Norway started as an institutional initiative around the turn of the century. UiT – The Arctic University of Norway established the research group *Giellatekno* for Sámi language technology, and at the same time *Sámediggi* (The Sámi Parliament) in Norway decided to use resources for creating proofing tools for North and Lule Sámi and established the *Divvun* development group. These groups have worked on research and development of Sámi language technology ever since and now constitute a research and development community of between 10 and 15 persons, in addition to an international network of cooperation partners. This initial work eventually led to the release of the first Sámi spellcheckers in 2007.

The groups have, partly in cooperation with others, since then produced different language technology tools for seven Sámi and approximately 20 other minority languages. All the tools are based upon a computational model of the lexicon and morphology (modelled as finite state transducers). Not all tools are available for all languages, but they include keyboards, spellcheckers, grammar checkers, morphology-enriched dictionaries, machine translation, speech synthesis, annotated corpora and tree-banks. The language technology tools developed as well as the effect of their (quite extensive) use is discussed in Antonsen and Trosterud (2020).

All of these tools are freely available and open source and have been well received by the respective user communities. User feedback has been essential for the maintenance and improvement of the tools. This work is included in a multi-lingual infrastructure named GiellaLT (The GiellaLT infrastructure n.d.), at present containing approximately 140 languages, most of them low-resource languages with less than 10k speakers (Moshagen et al. 2023).

It is important to note that the main initiative for developing these tools came from the Sámi people, as represented by the Sámi Parliament. Our work – done by both native speakers and community outsiders – is thus legitimized by the Sámi language communities and is perceived as work done for the best of the communities.

Earlier research

In the field of natural language processing there is a growing concern that both research focus and methods are biased towards English. The key publication (Bender 2019) presents the *Bender Rule*, stating “Always name the language you’re working on”. The point is that “English is neither synonymous with nor representative of natural language” (Bender 2019). In the article she goes through a range of phenomena that make English stand out as anomalous and thereby poorly fit as a yardstick for language processing: it has a well-established orthography (represented by ASCII only), massive amounts of training data, very limited morphology, a relatively fixed word order, and a standardized concept of “word”. Bender has been widely cited, and her critique was taken up by e.g. Lane Schwartz (Schwartz 2022). Working on Yupik, Schwartz demands the Association for Computational Linguistics (ACL) must confront ongoing colonialism. He proposes a set of ethical prerequisite obligations for ACL when working with Indigenous languages, the most central point being an adaptation of the Hippocratic oath, formulated as *Primum non nocere / (above all,) do no harm*, in this case: do not harm the language community. We believe this to be an important point. The lesson to be drawn from Schwartz’ paper is one of responsibility. However, he is less specific as to what harm natural language processing may do, and even less so when it comes to alternatives. We will address both questions in section 3.

In an earlier paper, Trosterud (2013) shows that only a few languages at that time were included in the major operating systems, by means of keyboard, proofing tools and localization support. At that point, it was still quite easy to include language processing programmes as third-party solutions for computers. As shown by Moshagen, Trosterud, and Antonsen (2019), this possibility has, with a few exceptions, been severely restricted during the last decade.

To our knowledge there is no earlier research on actual access to language technology tools for minority language speakers on various platforms besides Moshagen, Trosterud, and Antonsen (2019). One exception could be Mahelona et al. (2023), but overall, most research focuses on various technical aspects of data accessibility, governance, etc, and data only. This leaves out the core point of scientific research and the original idea behind developing language tools – its value for human society and the language users. An algorithm that produces technically acceptable output that is completely useless or even worse – harmful – for a language community, misses its point in being a benefit for humanity.

In an early article, Cuckier and Mayer-Schoenberger (2013) introduce the term *datafication* for the process of rendering into (digital) data aspects of the world that have not been quantified before, in order to draw conclusions from them. They claim that the unprecedented amount of data changes the approach to them in three ways. By going from collecting *some data* to (in principle) *all data*, questions of sampling become irrelevant. This automatically implies a change from *clean data* to *messy data*, but given the amount of data, clean data will outnumber the messy ones. With "all data" available, connections are too numerous to establish what causes the system's behaviour. Instead, we witness *correlations*. One of the examples used by the authors is machine translation based upon far larger data sets than ever before. These observations are relevant for the following discussion on Indigenous language technology.

Campolo and Schwerzmann (2023) discuss the conceptual change from a situation where "rules are applied to data to produce outputs" to one where the order is reversed to "an exemplary type of authority [from] machine learning". Their topic is programming languages rather than natural languages, but the important point is the one related to authority, or control: this new and reversed order prevents the language community from having control over their own language. A similar point was made by Trosterud (2022), in an article on European language standardization bodies that discusses the shift from spellcheckers based on lexicon and grammar to spellcheckers based on large collections of text.

In a similar way, the question of data control has come up in an Indigenous language setting. The Maori people have opted against open source to protect their language, cf. a recent article in *Wired* (Coffey 2021). The author presents the case of Maori speech recognition, where work done by the language community was taken over by a commercial company, who paid informants for reading in sound files and thereafter made the resulting speech recognition model into private property.

Within artificial intelligence (AI), there is a research direction that investigates how corpora for languages with less resources form part of multilingual language models. One case in point is Lin et al. (2022), presenting a model for 20 different languages. Sixteen of these languages are among the world's top 100 languages, and the remaining four (Finnish, Swahili, Estonian and Basque) have more than 1 million speakers and are official nation state or regional languages. Seen from the perspective of Indigenous languages, sets like these are not representative.

ACL now has a special interest group on under-resourced languages (SIGUL), with so far one conference in the ACL anthology. In this conference, Doğruöz and Sitaram (2022) discuss the insights sociolinguistics can give to planning of language technology projects, especially concerning whether they serve the needs and preferences of the language communities.

Methodological and ethical considerations

Before we dive into the control or ownership discussion, it is necessary to introduce two sets of principles, abbreviated the FAIR and CARE principles. The FAIR principles (n.d.) stem from the open data movement, where the goal is to improve the "Findability, Accessibility, Interoperability, and Reuse of digital assets" (Wilkinson et al. 2016). FAIR is especially relevant in a minority language context. The availability of the most important digital assets, high-quality native language material, is proportional to the size of the language community and all available language material is thus important.

Building Indigenous and minority language technology on the FAIR principles alone is problematic. Although positively received, the FAIR principles have not been deemed enough for work with Indigenous languages and communities, and thus Carroll *et al*. (2020) proposed the CARE principles as a complement, to ensure "Collective Benefit, Authority to Control, Responsibility and Ethics" (Research Data Alliance International Indigenous Data Sovereignty Interest Group 2019) in research involving Indigenous communities and data. The data must be used for the benefit of the language community, hence following responsible and ethical principles, and the language community must be in control of their own data and their own language.

While the work on Sámi language technology in practice has followed the FAIR principles since the beginning long before these principles where formulated, the topics of this article go to the core of the CARE principles, as will be discussed in the following sections.

Language technology and CARE

Indigenous and minority languages are in a two-way vulnerable position. Due to modernization and nation-building, they are under strong pressure of language shift, whereas the same modernization process also forces the languages into new domains. For language communities going through a revitalization process, many of the language users will have the language in question as a second language. In the Nordic countries, enabling the use of Indigenous languages in all aspects of life has been an explicit political goal during the last decades. None of the Nordic national minority languages were used as written languages during the twentieth century modernization process, and in order for them to function in a modern society there is a need for explicit standardization work. In this respect, the Nordic countries are representative of most countries that want to encourage the use of minority languages.

In order to cope with this, the Indigenous language communities in the Nordic countries are presently engaged in language planning work, including a wide range of questions relating to both orthography, grammar and terminology. All this makes it imperative for the language communities to be able to explicitly govern the content of language-aware computer programmes, and to take an active part in the development of language technology for their languages.

This section is split into four parts: first we discuss the ethics and the CARE principles related to language data, and the collection and sharing principles of them, as well as data as a prerequisite for building language technology tools. Secondly, we discuss procedures for building language technology tools. The third section discusses platform access, the question of who controls the human language aspects of digital systems, and the ethics of this control. Finally, we discuss the ethical considerations related to actual use of language technology.

Language data

For the majority languages, the question of control over language data is seen as a question of the authors' individual control over their own texts. A case in point is the US Authors Guild, organizing almost 14,000 members, which in 2023 filed a lawsuit against OpenAI for illegal use of texts written by their members (Guild 2023). For Indigenous language communities, text ownership is not always clear, and perhaps not always relevant either. Instead, ownership of many texts could be seen as belonging to the community as a whole.

Just as in the case of land ownership, large parts of existing texts in Indigenous languages do not have a specific author. The most valuable texts, from both a linguistic and a cultural view, are typically traditional stories, legends and poetry, often written down or recorded by visiting linguists or social anthropologists. Another source of indigenous texts in e.g. the Nordic countries is translations of governmental and other official documents in the majority languages. These are important sources when wanting to document Indigenous languages' lexicon and grammar, but often both author and translator are unknown, or the translator has no copyright. Even though these two text types – traditional and bureaucratic texts – often do not have personal authors like the plaintiffs behind the Authors Guild lawsuit, the text corpus is just as crucial for language technology development, and as such may have profound consequences for the language community in question.

In terms of language technology work, the Divvun and Giellatekno groups have collected speech and text data, and developed language resources such as dictionaries, morphological descriptions including large lexicons, and syntactic models. These resources are available under various open-source or open-access licenses, except for the copyrighted texts. The copyrighted texts have been collected on behalf of the Norwegian Sámi Parliament to ensure Sámi ownership of the corpus collection. Language technology resources like morphological and lexical descriptions have a shared ownership between UiT – The Arctic University of Norway and the Norwegian Sámi Parliament. This ensures, both symbolically and legally that the Sámi people, represented by the Norwegian Sámi Parliament, own and control the use of these resources.

Although well-intended, there are several problems with the present ownership system. First of all, the Sámi parliaments of Sweden and Finland have not been part of the agreements. Including them is being discussed now with respect to texts and text collections. Another issue is connected with open access to language data. Although openness is at the core of the FAIR principles, licenses and procedures are not in place for the Sámi language resources, ensuring that users of open data are also following the CARE principles – an attempt at doing so is the Māori Kaitiakitanga license (TeHikuMedia n.d.) and the Principles of Māori Data Sovereignty (Te Mana Raraunga 2018).[1] The consequences of the lack of enforcement of the CARE principles imply that a number of LLMs have been developed using available Sámi corpus data, but with little to no quality assurance of the generated output. The result is that these language models generate Sámi text that will look Sámi to an unknowing person, but in reality often is just gibberish (cf. Wiechetek et al. 2024).[2] It is obvious that undesirable output is not generating *Collective Benefit* (cf. CARE). The question is: how do we ensure there will be enough critical analysis of language technology use for Indigenous languages that we detect and stop use that is detrimental to the collective, instead of beneficial?

In our experience the most serious drawback of Indigenous language communities cooperating with closed software companies is that the communities risk handing over linguistic data to software companies as the companies' closed source. The bottleneck for minority language technology is always the availability of fluent writers and in some cases even speakers. For many language communities, there may be few literate speakers or a small number of fluent speakers of the traditional language, and the output of their work should not be given to one company or even one purpose only: when the resulting software has become outdated or if the company goes bankrupt, the linguistic resources are lost to the community. Resources produced by public funding ending up as closed source resources for private companies going bankrupt is a total waste of both funding and time, and the risk of this happening far outweighs any short-term benefits.

The usual way to protect data is by license agreements saying "resource X can only be used in accordance with license Y". There are two problems with this model. Firstly, as shown by the case raised by the Authors Guild, even strict licenses have so far not been able to prevent large companies from using license-protected data as raw material for machine learning. Secondly, and more severely, data licenses only regulate author attribution (the creator must be mentioned) and commercial aspects (the data may or may not be used for commercial purposes). No license has so far touched upon the crucial aspect of minority language data: responsible use and collective ownership. Consequently, what is needed is a license saying that the data should be used only according to the Hippocratic oath of Schwartz (2022). The *Kaitiakitanga* license mentioned above is a first attempt but only addresses the usage aspect indirectly by requiring explicit permission from the language community before using the data.

It is no coincidence that the focus of the strict licenses is upon commercial use. But language is not a part of the capitalist model, it is part of human society, even constituting it. Thus, linguistic consequences of any given license have never been given a thought. Rather than being concerned about the (unfortunately marginal) commercial potential of minority language resources, what is needed is a discussion on the consequences of the technology for the language societies. Even more so, we need a discussion of linguistic human rights (cf. Skutnabb-Kangas and Phillipson 2022), and what that entails in the digital domain.

Building language technology tools

Building language technology tools for Indigenous languages is not without its pitfalls. Standard evaluation methods within the field are made for automatic evaluation, so that the outcome of different parameter settings may be evaluated in an efficient way. Unfortunately, there are few Indigenous language translations and even fewer translators available. A recent example of how to get past this problem is described in Paul et al. (2024, 12), evaluating their North Sámi model against a translation from an English Wikipedia-based QuAC dataset into North Sámi made by another NMT-based machine translation programme. By doing that they overcome the inaccessibility of North Sámi evaluators, in essence by making a system where the neural machines controlled each other, and no native speakers are involved.

The Sámi language standardization institution *Giellagáldu* (Giellagáldu n.d.) decides upon correct Sámi grammar and spelling. However, the process of standardizing even central aspects of the literary language is far from being finalized, and published text is likely to contain language use not acceptable to native speakers, or orthography outside the standard, be it deliberately or due to lack of knowledge of the standard. Much of the text found on social media is written even more freely and to a lesser extent following any standard. In other words, when making neural language models based on an algorithm learning from written texts, one cannot assume that the text corpus represents the standard (cf. also the discussion in Trosterud

2022). As an illustration, the staple corpus of any machine learning process is Wikipedia, but alas, almost all of the Wikipedia editions for Indigenous languages are written by people not knowing the language in question. As an example, consider the *circumpolar* Wikipedia editions (Trosterud 2021), where the language is characterized by articles built upon the same model ("X is a region in Y") with a narrow vocabulary and non-native and often wrong grammar. Any model built upon such language will have the same characteristics. If such data is taken as the basis for AI-based normative tools, then the potential catastrophe poses a threat to the literacy of Indigenous languages.

At the same time, many members of Indigenous language communities are hoping that the technological advancement will help support their language. They argue that it is crucial that their language takes part in the latest technological development, exemplified by ChatGPT, out of fear of being left behind. It becomes an Indigenous language technology version of the FOMO (Fear Of Missing Out) phenomenon. There is some foundation in this, as documented by Lindgren (2010). Their point is that delayed development leads to domain loss, which of course is one of the forces behind language shifts. This is also the line of thinking behind the initiative on Sámi language technology development, and the reason the work was started in the early 2000s. But as argued above, LMMs have so far had little to offer to Indigenous language communities and can even be harmful to the language communities. This is not to say that LLMs could not become better even for Indigenous Languages, or that technological advancement cannot improve the situation for them. However, care must be taken to avoid harm.

In terms of the CARE principles, one could analyze the process of building language technology tools as follows. In terms of *Collective Benefits*, one should ensure inclusive development and innovation (the *C1* of the CARE principles). This entails including language group members in the actual work, building knowledge and know-how in the process. It also helps building an understanding of what can and cannot be done with language technology, which again helps prioritizing what benefits the community most. Do we need spellcheckers and keyboards, or machine translation first? Which direction of machine translation? Speech synthesis? It is rarely obvious what path to take and what to prioritize, but community insiders that are able to write their language, perhaps even with knowledge in language technology, will help a lot in the process. Much can also be learned from other language communities with similar needs and desires, as well as from those that have already developed some tools.

The core of the CARE principles is that a language community can have *Authority to Control* about what is being developed. As argued above, not all tools are necessarily good for the language community, and the development targets must be aligned with the needs of the language community. As an example: one possible overarching goal could be to make it both possible and supported to become linguistically independent of the outside world, in terms of digital communication, and to build a sense of putting your own language first, always. That entails machine translation (MT), but not necessarily bidirectional MT. Since MT rarely is perfect, and less so the fewer text resources there are in the language, it is possibly a good choice to avoid MT *into* the Indigenous language, especially since most speakers usually are bilingual in the majority language as well. There is no need to automatically create a lot of bad machine translated text in the native language. At the same time, it would be very useful to have even relatively bad MT *from* the Indigenous language, so that speakers can write their own language, and the text is still understandable for outsiders using MT. The idea is to move the burden of understanding from the Indigenous people to the majority language community, and in this way free the Indigenous community from having to always communicate in the majority language. Whether this is a desirable goal must be decided by the language community, not by outsiders, and for this the community needs *Authority to Control*.

The same goes for data ownership and data use. As argued in *Principles of Māori Data Sovereignty* (Te Mana Raraunga 2018), language data belongs to the language community, and use of such data must be controllable and governed by the language community to ensure that the use is beneficial to the community.

The *R* of the CARE principles stands for *Responsibility*. In the context of building language technology tools, responsibility implies direct involvement with the language community or representatives of it, and ensuring that the development also strengthens the community, as well as to ensure that the tools built are actually useful and the ones needed by the community.

The last element of the CARE principles is *Ethics*. Its first sub-point emphasizes what has been described as the Hippocratic oath earlier in this article; in the CARE version it says: minimize harm and maximize benefit. When building language technology tools, this implies inter alia building tools that support Indigenous language use and avoid making tools that cause language shift. It is not always obvious how tools could cause language shift. One option to prevent this from happening could be to include follow-up studies on usage and longer-term consequences in future project plans.

Another ethical consideration is one of justice, especially related to imbalance in power relations. It is striking how large the distance is between language communities and the large international companies developing the systems and platforms on which language technology tools are going to be mostly used. Both from our own experience and from what we have learned from other initiatives working on language technology solutions for Indigenous languages, it is very hard to make actual tools available to the user communities (see the next section). Almost always when something is actually happening, it is the tech company taking the initiative, not the other way around. It is of course good that tech companies are taking initiatives, but it very clearly illustrates the uni-directional way of communication and power imbalance.

The last ethical consideration mentioned by the CARE principles concerns future use. Regarding language technology tools, one aspect would be to ensure long-term support for the maintenance of the linguistic resources, for the infrastructure needed to build the tools, and updating the integration with external systems. Another aspect is to plan resource building and development such that they can be reused or reapplied to new domains or environments with as little effort as possible. As mentioned elsewhere in this article, the most limited resource in work with Indigenous languages is human resources, people knowledgeable of the language in question. Therefore, securing long-term support and development of the tools in question requires ensuring that the language experts focus only on the kind of work where their knowledge is indispensable. The GiellaLT infrastructure described earlier in the article, supports these goals by emphasizing reuse and separating language-independent code from language-specific code, as well as being backed by governmental institutions with long-term commitment to the development of Indigenous language technology tools.

Getting language technology into the hands of users

As society has become more and more digital over the last 10–20 years, particularly the last few years with the emergence of generative artificial intelligence and LLMs, the access to services when using our own language has dramatically changed, in a very digitally divided way. For the languages that are supported by LLMs, the possibility of interacting with the world through digital devices has increased many times, whereas chances of digital use have been drastically reduced if one's language happens to fall outside of these models. But also in other areas actual access to existing language technology can be severely limited, usually for no apparent reason.

Examples abound from low-level tools like keyboards to high-level speech assistants. Moshagen (2022) documents several cases, including how physical keyboards for iPads and Android tablets are outside the reach of third parties, creating major issues for pupils in Sámi schools and classes. Another example documented in Moshagen (2022) is the lack of support for third party proofing tools in web-based office software suites, and also the lack of support for all of the language codes in ISO 639 (Wikipedia n.d.). Lack of support for ISO 639 makes most languages unknown both to the host system and to individual software packages and creates problems for the users when they want to use services for their language, as well as for developers making tools to support these languages.

A further example is software localization, which is dependent on the goodwill of the developer(s) of the software in question. A more serious concern for society is related to health care. As the population grows older (see e.g. Naumann and Hess 2021), there is a need to enable more people to live in their homes longer as they grow old. This requires that basic health care and monitoring can be provided in the living rooms of elderly people. The systems that provide this service must be able to communicate with their users in their own language, nothing else is going to work. But presently no such system is able to speak and understand e.g. North Sámi, or any other Indigenous or minority language. And worse, there is no system in place to make that possible independently of the producers of these systems.

As shown above, society has allowed control over the digital infrastructure to be in the hands of large, non-transparent, international technology companies. We as a society presently do not have the tools or means to make sure the digital infrastructure is actually open to everyone and to every single language. It is as if we asked a plumber to provide water pipes and water to our community, then discover that some villages did not receive water, and realize there are no means for a plumber to provide pipes to them. And worse: we realize that we are not allowed to let other plumbers do the remaining job, so the villages will forever be without water, unless the original plumber changes their mind.

Language works in similar ways on computer systems. Language as such is intrinsic to every computer system. One can hardly conceive of a computer system without any trace of human language. You need language – written, spoken, symbolic – to interact with a computer: to read menus, to give commands, to tell it to do something. Some features can be encapsulated in nice graphics, but you still need to communicate the functionality of a feature via natural language.

English has become the de facto computer as well as research subject and object language, to the extent that many do not even think about it (see Bender 2019 for an analysis of the consequences of this). But the software industry has fooled itself into thinking that English equals *the* software and that all other languages are "localizations", instead of seeing language as a necessary but independent part of every software system, and all languages as equal peers. Being independent, it would have been possible for others to replace a language or add another one. Being independent, all linguistic aspects of the system would be accessible for third party developers.

The plumber problem has become more serious over the last decade and there are particularly three development paths that have created these problems. The first one is the shift to the "cloud", to servers owned and controlled by the major technology platforms. By moving away from personal devices to the cloud, one removes at the same time the ability to install third party software like spelling and grammar checkers. The user only has access to what is provided by the platform owners. That does not include support for minority and Indigenous languages.

The second development trait is the increasing focus on advanced language technology and monetization of that technology. This covers services like machine translation, speech-to-text, advanced search, and speech assistants. Because of the role of these services for generating income, they are closed off and not available for third parties to adapt to new languages. It is thus not possible to build on top of existing services to provide similar services for Indigenous and minority languages, and these languages are therefore shut off from the places where people are used to finding these services for the majority languages. In several cases there is no way to provide services for Indigenous and minority languages, e.g. speech assistants, text to speech and dictation, because the whole platform that the service is built on is closed off.

Several of the services mentioned in the previous paragraph tie in with the third development: the recent advances in machine learning, artificial intelligence and LLMs. These models require large amounts of texts to be trained on, and even if there has been progress in making working models with somewhat less amounts of texts by transfer learning and similar methods, everything we know about example-based learning implies that most languages of the world will not be part of this technology.

The situation is a parallel one to the plumber metaphor: we as a society accept that the control of the core infrastructure for our society is in the hands of someone with no obligations to the society and also in the hands of someone the society has no control over. From both a democratic and an equality point of view this can hardly be seen as anything but highly problematic. When the mayors from 108 Norwegian municipalities beg Google to provide tools in the language of their school children (New Norwegian), and nothing happens, the situation is very far from acceptable (Aksnes 2021).

The logical follow-up question is how society can regain control. So far this has happened piecemeal and in small steps, one app, one functionality and even one or a few languages at a time. Given there are more than 7,000 languages in the world, that is clearly not a fruitful way to go.

A first step is to make the technology providers accept language as an expression of culture and identity, that language diversity is all about function, not form, and that a language belongs to its speakers, not to a commercial company. Principle D of the ABCD principles presented by Schwartz (2022) is "above all, *do no harm*", which in this context means: do not hinder a language community from using, seeing and speaking their language, in any parts of life. Hindering that does a lot of harm.

The CARE principles provide a strong guideline on how to ensure best practices: give the language communities *Authority to Control* where, when and how to use their language, even in a digital world. Such control will also give *Collective Benefits* to the language community, as their language becomes an unquestionable part of the digital infrastructure. It is also the *Responsible* thing to do: only the language community itself can determine how to *minimize harm* and *maximize benefits*, to cite E1 of the CARE principles.

Ethics of language technology use

In the previous sections we covered CAREful use of data to build language technology tools, and making the tools and languages available to the language communities on digital platforms. The remaining aspect of language technology for Indigenous languages to be discussed regards the ethics of using the available tools.

The tools we build cannot be better than the data or work we put into making them. Over time there have been multiple examples of tools that have had serious quality issues, but still they have been made available for public use. Also the opposite has been true: a lot of work has gone into building quality tools that in the end are not being used by the language community. In this section we will analyze these cases and their relation to the CARE principles.

The probably most well known and most frequently used language technology tool is our machine translation (MT) application, a rule-based system translating from North Sámi to Norwegian, developed 10 years ago (Giellatekno Apertium n.d.). The quality of the output varies regarding syntax and morphology, but it preserves the semantics of the original text quite well. For a user, there is no doubt about this being machine-generated. The main goal of the system has been to making texts written in North Sámi available to people that do not speak the language, and thereby liberate North Sámi writers from having to write in Norwegian just so everybody will understand. In this way, North Sámi writers can use their native language in more circumstances than before the MT system was introduced, which gives a collective benefit to the North Sámi population.

In the last couple of years a number of LLM-based MT systems have appeared, with translation to and from numerous Sámi languages. While there are clearly use cases for these systems, the output produced requires a different type of attention from the users. Often the output looks good, at least for a language like North Sámi that has several millions of words of parallel text. Still, due to the nature of LLMs, there is no guarantee that the semantics of the input text is preserved in the output text. That is, to avoid serious harm when using such tools, the user needs to be fluent in both the source and target languages, and make sure that the semantics of the source language is preserved in the output.

In cases where we know that the training material for the LLMs does not exist, or only exists in small quantities with grammatical errors or a lot of spelling errors, the output is in general flat-out wrong (see Wiechetek et al. 2024). Still, on the surface the output text looks good and in the example from Wiechetek et al. (2024) it looks "very South Sámi" for people not knowing the language. Speakers of the target language that have the time and knowledge required to post-edit such output may perhaps find it useful, but if non-speakers take strings of non-existing words at face value and publish it as if it were South Sámi, the result contributes to damaging the literacy of the language. The ethics of publishing the tools and using them includes making sure the users and thus also the "developers" understand the limitations and known issues of the tools so that one easily can avoid producing text that in essence is detrimental to the language community.

Even with regard to more traditional tools like spellcheckers it is very much possible to make and release tools with problematic properties. For spellcheckers released early one needs to make clear to the users that it really *is* an early release, and that they cannot trust it. In reality many people are so in need for tools helping them write their language that they will use whatever they get their hands on, while at the same time not having the necessary training to evaluate the speller actions: are the words being flagged actually wrong? Do the correction suggestions make sense? The end result is that the final text may contain a number of errors, while people believe it is correct, and in the worst-case cause people to establish misconceptions about how to write their language.

We have also seen cases where tools have been developed, resources have been used to build high-quality tools, only to realize that the tools are not being used by the language community. This implies both a lot of wasted time, time that could have been spent on other things when resources are scarce, and that the effort was a missed opportunity for language revitalization. Both concerns raise ethical considerations – how should we ensure that we spend our time best, and how do we ensure that what we do is well perceived by the language communities? One lesson learned so far is to make sure that a language community is involved from the very beginning, and to ensure that at least some representatives are taking an active part throughout the project. Such involvement is also important to ensure sustainable development after the initial project period is over, as no language tool is ever finished.

Applying the CARE principles to language technology

The CARE principles[3] (Carroll et al. 2020) were written to augment data management with a rule set appropriate for Indigenous communities. But as shown above, it is not only data that needs to be managed in an appropriate way, even the development, distribution and use need the same kind of oversight and consideration regarding Indigenous languages.

In Table 1 we have tried to formulate the CARE principles for three areas of LT development from data to final tools in the hands of users, all starting with a D (hence the *3D* moniker): Data, Development, and Distribution and use. The core of each principle is formulated as a yes-no question appropriate for the specific LT area, and if one can answer yes on all questions, the most important things should be in order. The idea is that this table or adaptions thereof can be applied to any LT project for an Indigenous language and be used as a guiding tool in building a sustainable project that will support language use, revitalization, and thus support future survival of the language community.

Some comments on specific cells in the table are in order. First of all, not all cells apply in all cases: if a tool is developed and delivered as a web service, most questions related to distribution become irrelevant. For the cell D3-E1 ("Do you ensure tools are actually used?"), the underlying issue is that resources are scarce for Indigenous languages, and developing LT tools is a large investment for anyone undertaking such a project. It is thus important in an ethical evaluation of the project to ensure that the tools are actually used. The difference between D3-A2 ("Can misuse be detected?") and D3-A3 ("Can usage detrimental to the language community be restricted?") is that the first is concerned about getting the data needed to identify an issue, whereas the second relates to the action of limiting damage, if needed. Cell D3-R3 is only relevant if deemed so by the language community members, and this question is there to check whether such restrictions are needed.

Table 1. "3D" analysis of Language Technology core areas using the CARE principles.

CARE	Principles	D1 –Data	D2 –Development	D3 – Distribution and use
Collective Benefit	C1 (Inclusive development and innovation)	Is public data shared with the community?	Is development done in cooperation with community members or entities?	Are tools supported on all major platforms?
	C2 (Governance and citizen engagement)	Is data use transparent and auditable?	Is development open and auditable?	Is distribution without restrictions on all platforms?
	C3 (Equitable outcomes)	Is value generated from data brought back to the community?	Is outcome shared with community members or entities?	Is distribution designed to reach all community members?
Authority to Control	A1 (Recognizing rights and interests)	Is the language community informed about data use?	Is the language community informed about development goals?	Is distribution transparent and auditable?
	A2 (Data for governance)	Is data governable by and accessible to the community?	Can the language community influence development goals?	Can misuse be detected?
	A3 (Governance of data)	Is the license appropriate?	Are development goals aligned with community needs?	Can usage detrimental to the language community be restricted?
Responsibility	R1 (Positive relationships)	Does the language community trust you?	Are community members working in the project?	Does usage strengthen the language community?
	R2 (Expanding capability and capacity)	Are community members working with the data?	Do you have a feedback channel for the language community?	
	R3 (Indigenous languages and worldviews)	Does the data cover the full or known language?	Is development aligned with relevant community values?	Can distribution and use be limited to community members?
Ethics	E1 (Minimal harm and maximal benefit)	Has data collection avoided harm, will collected data create benefits?	Are developed tools supporting language use and revitalization?	Do you ensure tools are actually used?
	E2 (Justice)	Is data collection done with community acceptance?	Are development costs reasonable for the community?	Is distribution equally accessible to all community members?
	E3 (Future use)	Does the data have enough metadata for future use?	Is development building knowledge and sustainability?	Is distribution infrastructure adaptable to different environments and scalable to future needs?

Conclusion

We have in this article outlined the digital reality for low and extremely low resource Indigenous languages, using the Sámi languages as an example. We have also analyzed the different steps in language technology development, from data to use, in terms of the CARE principles.

Regarding language data, we conclude that the existing licenses used in the digital economy are not well suited for Indigenous languages. They do not consider harmful use of the licensed data, nor the linguistic human rights of the language communities. The Māori Kaitiakitanga license is a first attempt at covering the collective ownership of Indigenous language data, but it does not cover other issues described in this article. It is not obvious what a good licensing model would look like, so that is left for future work. We then analyzed language technology tool development in terms of the CARE principles, and outlined how such development could be done according to those principles. Although there certainly are areas for improvements, the GiellaLT infrastructure is a good foundation for developing Indigenous language technology along the lines of the CARE principles. Still, such a development infrastructure is just a framework: the actual adherence to the CARE principles must be done by the people using it. Regarding getting the available language technology tools into the hands of the language communities, we used the plumber metaphor to illustrate how the software industry has been allowed to build a core societal infrastructure without any requirement on supporting all the languages of the society, despite that being an explicit political goal in the Nordic countries. Our outlined solution is to systematically let the language-specific parts of software systems be opened, without any need to open the code of the underlying computer system. We then considered ethical use of available language technology and gave examples of various types of problematic use of tools and resources. Finally, we created a table for LT projects, to help classify their suitability for Indigenous languages, applying the CARE principles to each of the core parts of an LT project.

Today, there is a situation of imbalanced openness: most of the little data there is, is openly available online, whereas the possibility for Indigenous language communities to integrate their own solutions in Big Tech applications is more or less blocked. The result is that Indigenous language communities are left with either no applications or very bad applications. This imbalanced openness has a parallel in imbalanced power, and lack of communication: access to core systems is purely in the hands of the system developers, and it is usually very hard to get in touch with the people making decisions that can actually help improve the situation for the Indigenous language communities.

This leads to Big Tech monopolizing access to digital language use. As a consequence, language services emerge as a new form of colonialism. Language use is no longer in the hands of the language community but governed by companies outside any democratic control. Bringing back this control to the language communities is crucial for language survival.

Our vision for an ethical language technology that includes the needs of all language communities, not just the big majority languages, is one where each and every language community on their own or together can decide what they want, where they want it, and make it happen without the burden of convincing a far-away corporate headquarter that it is needed. Our vision is that the software industry recognizes all languages as equal peers and allows each and every language free access to the digital infrastructure of our societies, no questions asked. This does not require opening up the source code of any system, only the language part. That should not be too much to ask.

Acknowledgements

Thanks to all colleagues, reviewers and others giving valuable feedback to the article. All remaining errors are the sole responsibility of the authors. The tools developed are being used in education and language teaching by the language communities. The Sámi text corpus is collected and maintained by the Divvun group on behalf of the Norwegian Sámi Parliament.

Notes

1. See also https://tehiku.nz/te-hiku-tech/te-hiku-dev-korero/25141/data-sovereignty-and-the-kaitiakitanga-license for more pointers.

2. In all fairness it should be mentioned that for some of the Sámi languages the output can be both good and useful, but you need to speak the Sámi language in question to be able to make proper use of the output, and ensure proper quality of the content.

Funding

Giellatekno is funded by UiT – The Arctic University of Norway. Divvun is funded by the Norwegian Ministry of Local Government and Regional Development. Both groups have members from the Sámi language communities.

ORCID

Sjur Nørstebø Moshagen http://orcid.org/0000-0003-3771-9521
Lene Antonsen http://orcid.org/0000-0002-7927-8536
Linda Wiechetek http://orcid.org/0000-0002-5171-0841
Trond Trosterud http://orcid.org/0000-0002-2300-2995

References

Aksnes, Solveig Nyhus. 2021. “Stor ordførarprotest mot Google – vil at elevar får eit betre tilbod på nynorsk [Large Mayor Protest Against Google – Want Pupils To Get Better Support For Norwegian Nynorsk].” Accessed October 13, 2023. https://www.nrk.no/mr/ordforarar-krev-at-google-gir-eit-betre-tilbod-pa-nynorsk-til-skuleelevar-1.15353991.

Antonsen, Lene, and Trond Trosterud. 2020. “Med et tastetrykk. Bruk av digitale ressurser for samiske språk” [With the press of a button. Use Of Digital Resources For Sámi Languages]. In Samiske tall forteller. *Kommentert samisk statistikk*, 12: 43–68. Kautokeino: Sámi allaskuvla.

Arctic Indigenous Languages Map. 2019. *Ságastallamin – Telling the Story of Arctic Indigenous Languages Exhibition*. Indigenous Peoples’ Secretariat and UiT – The Arctic University of Norway. Adapted from “Conservation of Arctic Flora and Fauna, CAFF 2013 – Akureyri. Arctic Biodiversity Assessment. Status and Trends in Arctic Biodiversity. – Linguistic Diversity (Chapter 20), page 656. https://site.uit.no/sagastallamin/language-map/. License: CC-BY-SA.

Bender, Emily. 2019. “The #BenderRule: On Naming the Languages We Study and Why It Matters.” *The Gradient*. Accessed June 20, 2024. https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/.

Campolo, Alexander, and Katia Schwerzmann. 2023. “From Rules to Examples: Machine Learning’s Type of Authority.” *Big Data & Society* 10 (2). https://doi.org/10.1177/20539517231188725.

Carroll, Stephanie Russo, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, et al. 2020. “The CARE Principles for Indigenous Data Governance.” *Data Science Journal* 19 (1): 43. https://doi.org/10.5334/dsj-2020-043.

Coffey, Donavyn. 2021. “Māori are Trying to Save their Language from Big Tech.” *Wired*. Accessed June 20, 2024. https://www.wired.co.uk/article/maori-language-tech.

Cuckier, Kenneth, and Viktor Mayer-Schoenberger. 2013. “The Rise of Big Data: How It’s Changing the Way We Think About the World.” *Foreign Affairs* 92 (3): 28–40.

Doğruöz, A. Seza, and Sunayana Sitaram. 2022. “Language Technologies for Low Resource Languages: Sociolinguistic and Multilingual Insights.” In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, edited by Maite Melero, Sakriani Sakti, and Claudia Soria, 92–97. Marseille: European Language Resources Association.

FAIR Principles. n.d. Accessed June 20, 2024. https://www.go-fair.org/fair-principles/.

Giellagáldu. n.d. Accessed June 20, 2024. https://www.giella.org/.

The GiellaLT infrastructure. n.d. Accessed June 20, 2024. https://giellalt.github.io/.

Giellatekno Apertium. n.d. "En fri/åpen kildekode maskinoversettelsesplattform." Accessed June 20, 2024. http://jorgal.uit.no/.

Guild, Authors. 2023. “The Authors Guild, John Grisham, Jodi Picoult, David Baldacci, George R.R. Martin, and 13 Other Authors File Class-Action Suit Against OpenAI.” Accessed September 20, 2023. https://authorsguild.org/news/ag-and-authors-file-class-action-suit-against-openai/.

Lin, Xi Victoria, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, et al. 2022. “Few-shot Learning with Multilingual Generative Language Models.” In *The 2022 Conference on Empirical Methods in Natural Language Processing. Proceedings of the Conference*, edited by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, 9019–9052. Kerrville, TX: Association for Computational Linguistics.

Lindgren, Anna-Riitta. 2010. “Modernisation and Small Languages – Fatal Language Sociological Delay?” In *Planning a New Standard Language: Finnic Minority Languages Meet the New Millennium*, edited by Helena Sulkala, and Harri Mantila, 74–94. Helsinki: Finnish Literature Society.

Mahelona, Keoni, Gianna Leoni, Suzanne Duncan, and Miles Thompson. 2023. “Open AI’s Whisper is Another Case Study in Colonisation.” Accessed October 12, 2023. https://blog. papareo.nz/whisper-is-another-case-study-in-colonisation/.

Moshagen, Sjur Nørstebø. 2022. “Samisk språkteknologi i 2021 [Sámi language technology in 2021].” *Sprog i Norden* 52 (1): 93–102.

Moshagen, Sjur Nørstebø, Flammie Pirinen, Lene Antonsen, Børre Gaup, Inga Mikkelsen, Trond Trosterud, Linda Wiechetek, and Katri Hiovain-Asikainen. 2023. “The GiellaLT infrastructure – A Multilingual Infrastructure for Rule-based NLP.” In *Rule-Based Language Technology*, edited by Arvi Hurskainen, Kimmo Koskenniemi, and Tommi Pirinen, 70–94. Tartu: Northern European Association for Language Technology. http://hdl.handle.net/10062/89595

Moshagen, Sjur Nørstebø, Trond Trosterud, and Lene Antonsen. 2019. “Language Technology for Indigenous Languages: Achievements and Challenges.” In *Collection of Research Papers of the 1st International Conference on Language Technologies for All (LT4All)*, edited by Gilles Adda, Khalid Choukri, Irmgarda Kasinskaite-Buddeberg, Joseph Mariani, Hélène Mazo, and Sakti Sakriani, 210–222. Paris: UNESCO. European Language Resources Association (ELRA). https://lt4all.elra.info/proceedings/lt4all2019/.

Naumann, Elias, and Moritz Hess. 2021. “Population Ageing, Immigration and the Welfare State: The Political Demography in Western Europe.” In *Global Political Demography: The Politics of Population Change*, edited by Achim Goerres, and Pieter Vanhuysse, 351–371. Cham: Springer International Publishing.

Paul, Ronny, Himanshu Buckchash, Shantipriya Parida, and Dilip K. Prasad. 2024. “Towards a More Inclusive AI: Progress and Perspectives in Large Language Model Training for the Sámi Language.” Accessed June 20, 2024. https://arxiv.org/abs/2405.05777.

Research Data Alliance International Indigenous Data Sovereignty Interest Group. 2019. “CARE Principles for Indigenous Data Governance.” The Global Indigenous Data Alliance. GIDA-global.org.

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)*, edited by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, 724–731. Dublin: Association for Computational Linguistics. . https://aclanthology.org/volumes/2022.acl-short/.

Skutnabb-Kangas, Tove, and Robert Phillipson. 2022. “*The Handbook of Linguistic Human Rights*.” In *Blackwell Handbooks in Linguistics*. Chichester: Wiley Blackwell.

TeHikuMedia. n.d. “Kaitiakitanga-License.” Accessed June 20, 2024. https://github.com/TeHikuMedia/Kaitiakitanga-License.

Te Mana Raraunga – Māori Sovereignty Network. 2018. "Principles of Māori Data Sovereignty." October 2018.

Trosterud, Trond. 2013. “A Restricted Freedom of Choice: Linguistic Diversity in the Digital Landscape.” *Nordlyd* 39 (2): 89–104. https://doi.org/10.7557/12.2474.

Trosterud, Trond. 2021. “The Circumpolar Wikipedia Editions.” Arctic Knot Conference 2021. https://upload.wikimedia.org/wikipedia/commons/5/5f/ArcticWikipedias_Trosterud_English.pdf.

Trosterud, Trond. 2022. “Normative Language Work in the Age of Machine Learning.” In *The Role of National Language Institutions in the Digital Age Contributions to the EFNIL Conference 2021 in Cavtat*, edited by Jozić Željko, and Sabine Kirchmeier, 61–69. Budapest: Nyelvtudományi Kukatóközpont.

Wiechetek, Linda, Flammie A. Pirinen, Børre Gaup, Trond Trosterud, Maja Lisa Kappfjell, and Sjur Moshagen. 2024. “The Ethical Question – Use of Indigenous Corpora for Large Language Models.” In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, edited by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, 15922–15931. Torino: ELRA and ICCL. https://aclanthology.org/volumes/2024.lrec-main/.

Wiechetek, Linda, Flammie Pirinen, and Per Kummervold. 2023. “A Manual Evaluation Method of Neural MT for Indigenous Languages.” In *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*, edited by Anya Belz, Maja Popović, Ehud Reiter, Craig Thomson, and João Sedoc, 1–10. Shoumen: INCOMA Ltd. https://aclanthology.org/volumes/2023.humeval-1/.

Wikipedia. n.d. “ISO 639.” Accessed June 20, 2024. https://en.wikipedia.org/wiki/ISO_639.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3:160018. https://doi.org/10.1038/sdata.2016.18.