The Ethical Question –
Use of Indigenous Corpora for Large Language Models

Linda Wiechetek, Flammie Pirinen, Maja Lisa Kappfjell, Trond Trosterud, Børre Gaup and Sjur Nørstebø Moshagen
UiT Norgga árktalaš universitehta
Tromsø, Norway
giellalt@uit.no

Abstract

Creating language technology based on language data has become very popular with the recent advances of large language models and neural network technologies. This makes language resources very valuable, and in case of Indigenous languages especially, the scarce resources are even more precious. Given the good results of simply fetching everything you can from the small set of languages dominating the internet and training neural networks, there have been several attempts at doing the same for *all* languages. However, online Indigenous language resources are not comparable to the ones for dominating languages. They do not represent a broad range of genres and there is no guarantee that targeted forms outnumber erroneous ones. For most languages the people working on the language models do not speak the language of the models and bad performance is thus hard to notice. Problems related to intellectual property rights and copyrights loom high in contemporary discussions on neural language models. For Indigenous languages this issue is more critical than for dominating languages, since the amount of text available is so small (often barely counting millions of words). The use of artificial intelligence (AI) generated text as next-generation training data becomes even more problematic in cases where the available corpora are so small, since the share of AI-generated text will be larger. In this article we address these problems and describe our alternative, an ethical and sustainable way to work with Indigenous languages in the age of large language models.

Keywords: Indigenous languages, ethical LLMs, language technology

1. Introduction

In this article we discuss how large language model (LLM) driven language technology can be used when making beneficial language tools for Indigenous languages in an ethical way. While these ideals do not apply for Indigenous languages exclusively, their case illustrates most clearly when data driven language technology is neither ethical nor beneficial. Data driven methods make certain unspoken assumptions based on the majority languages they usually work with. When these methods are applied to languages with substantially different starting points without adapting to them, the technology fails.

We are a team of developers and linguists working with all the Sámi languages as well as on a number of other Indigenous and minority languages. The team includes both native and non-native speakers of these languages and we develop language tools on request of and in cooperation with the language communities. Our focus is on making tools that the language communities need to be able to maintain their languages in a digitized society. South Sámi (with 500 speakers and a 2 million token large corpus) is one of the main languages we work with, and between the Sámi languages it is on the lesser speaker/resource side. South Sámi both illustrates our point for extremely low resource languages and is the native language of one of the authors and we thus use that as our example language. It is also one of the least documented and standardized one within the Sámi language family.

Working with Indigenous languages implies that one has to work with a low amount of data (typically ranging between nothing and a couple of million words, cf. Antonsen and Trosterud (2020) for a survey of Sámi corpora). For rule-based methods this is not an obstacle, and in our case such methods have been used to create morphological analyzers, proofing tools and machine translation. However, the recent trends of large language models (LLM) based on data that is acquired mainly by scraping the internet has catapulted the role of data into new dimensions. For language communities that have only had access to the newest developments within natural language processing (NLP)[1] via their second language, either the majority language of their national state or a language learned in school, seeing such a tool in their first language is a huge step towards visibility and recognition. But not everything that glitters is gold, and fascination can obscure reality and impede critical thinking.

In the case of machine translation (MT) between an Indigenous and a majority language, bilingual users can easily disregard content errors and low

---

[1] first Google translate, then ChatGPT

LREC-COLING 2024, pages 15922–15931
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

quality of Indigenous language output. While the tool may be beneficial for a general understanding for monolingual users, the generated text should never be redistributed without quality assurance. The worst case scenario, where the output text turns the meaning into the opposite of what the source text says, does not only fail to be beneficial, but it is harmful for the language and therefore unethical.

If we transform the Indigenous language scenario into English, it would mean a flooding of the internet with generated sentences of the type in ex. (1) (a manual translation of the South Sámi neural MT output in ex. (2) to English preserving the output's lexical, spelling and grammatical errors, as well as other shortcomings, all of which are marked in boldface),[2,3] which then were to be used to train English language models. No native speaker of English would be happy with such a scenario or find it beneficial.

(1) Hundred indigenous people**'s** and environmental **campaigns** have **one mainhaerniem** in Oslo **to wind-blowing**, which **tear** two wind **powers**, **to which** the Norwegian government **has called** ”**to offend to** human rights”.

(2) Tjuetie *aalkoealmetji jïh *byjresekampanjh leah *aktem *åejviehaerniem *Oslosne *biegkemeurhkedh, juktie *rïjvestidh göökte *bïegkefaamoeh, *mejtie nöörjen reerenasse lea *gohtjeme ”*almetjereaktide *mïedtelidh”.

One thing that needs to be understood about Indigenous language communities is that many of them have a very short literary tradition, contrary to languages with huge language communities, a long literary tradition and an extensive standardization process. For Indigenous languages standardization is still an ongoing process. Written texts display a high variation in spelling and grammar and a lot of digitally available written text is not what the language community desires as their language. Instead, language experts have a much more important role within the society and their knowledge is often not available via written text. Additionally, large parts of the digitally available text is written by non-natives and language learners, which means that the corpus harvested from the internet cannot possibly be representative of the current written standard. And cleaning the harvested text is not possible without understanding of the language, or without using independently developed tools that have been quality assured by native speakers. That is, one needs native speakers as part of the process, independently of technology or methods.

For machine learning based NLP to deliver ethical and beneficial tools, it needs to comply with a number of factors. Firstly, its output needs to be evaluated by language experts from the respective language community. Secondly, authorship and language expertise needs to be discerned via curation, verification and annotation of the corpora. If one blindly uses texts that have not been verified and corrected to the current standard to create proofing tools for spelling and grammar as well as language generators, this will be harmful for the language standardization process and the language.

The scarcity of good writers makes original language data highly valuable for the whole language community both economically and ideologically. When for example valuable data and knowledge is taken for free and even without any reference to its origin to produce tools that then again are sold back to the language community instead of making it freely available to anyone in the language community and without them being able to modify the product, this is not ethical and respecting ownership. Ethical use of valuable data could for example be when language technology developers make language tools that are beneficial to the respective language community together with them, gives credit to the owners of the data, and makes the tools freely available to anyone in the language community.

2. Background

Ethical and sustainable use of language resources for Indigenous and minority languages have come in focus with the sudden boom of LLMs for multilingual NLP. Language resources have been created and curated for a long time, but the main focus of LLM builders is generally to automatically harvest text from the web. In this section we study the background of the languages and approaches relevant to this article.

2.1. Language background

We illustrate our point with examples taken from South Sámi, an Indigenous Uralic language spoken in Norway and Sweden (Ylikoski, 2022). According to Blokland and Hasselblatt (2003, p.110), there are about 2,000 ethnic South Sámi, of which approximately 300-500 are South Sámi speakers. South Sámi is an official language in altogether four municipalities in Norway and 10 municipalities in Sweden.The standardized orthography is in princi-

---

[2] Original English text: 'Hundreds of Indigenous and environmental campaigners have blocked a main thoroughfare in Oslo to demand the demolition of two windfarms that have been described by the Norwegian government as a ”violation of human rights”.'

[3] The error analysis of the generated sentence in South Sámi can be found later in this article.

ple the same for South Sámi in Sweden and Norway. In actual use there is still a lot of variation due to orthographic differences between the respective majority languages. This includes Swedish *ä* and *ö* vs Norwegian *æ* and *ø* respectively – the South Sámi standard defines *ö* and *æ* as the correct letters. There are three major South Sámi varieties: northern (or Åsele) South Sámi and southern (or Jämtland, Sweden) South Sámi (Sammallahti, 1998, p.24) and in addition the dialect in Røros (Norway). There are minor phonological and morphological differences between these dialects. The written standard of South Sámi was recommended by *Samisk Språknemnd* in 1976 and adopted in 1978 by the Norwegian Ministry of Church and Education and the Swedish School Board and was codified in a grammar *Sydsamisk grammatikk* (Bergsland, 1982) and dictionary *Åarjelsaemien-daaroen baakoegærja* (Bergsland and Magga, 1993). For a presentation, see Wiechetek and Kappfjell (2023).

However, there is a lack of standardization and clarification when it comes to grammatical variants due to language change and simplification.

For the South Sámi digitally available text collection (SIKOR, 2018), the majority of the writers are those who have learned South Sámi in school using a weak language model and with just a few hours of education a week. South Sámi (translated) text is often heavily influenced by the original Norwegian or Swedish text.

When the existing South Sámi text corpus is used as a basis for LLMs, such expressions will frequently pop up as the most "used and common" in Artificial Intelligence (AI) versions of South Sámi text. To the extent that AI will be used for South Sámi, such forms will be fed back to future models and to an increasing degree govern how the model claims the language "should be". Moreover, generalizations based upon text collections will not be able to reflect decisions made by the standardization body for South Sámi like for example a decision on how to express negation (Mattsson Magga et al., 2009). Pressure from the majority language also leads to the use of non-standardized syntactic constructions that are not desired by the language community. This is the case for a construction equivalent to "to have" in English. Instead of using the idiomatically correct genitive case of the owner and *leah* 'to be' as in ex. (3) (literally 'Mine are red trousers'), people use a literal translation with the verb *utnedh* "to own, possess" as in ex. (4).

(3) Mov **leah** rööpses
I.GEN be.PRES.3PL red
måvhkah.
trouser.NOM.PL
'I have red trousers.'

(4) *Manne **åtnam** rööpses
I.NOM own.PRES.1SG red
måvhkah.
trouser.NOM.PL
'I own red trousers.'

2.2. Previous Publications

Previous publications discussing use of corpora for Indigenous language technology have touched on the following aspects and consequences relevant to this article: polluting the corpora with generated texts (Shumailov et al., 2023), decolonizing language technology (Bird, 2020), the prospect of participation of the language community (Birhane et al., 2022), anglocentricity and lack of research into other languages (Joshi et al., 2020), ownership of the data and legality of the web scrapes and such, among others.

There have also been several popular science articles on problems arising from appropriating corpora published online. Newspapers and journals keep giving these topics increased attention even outside the academic sphere, as for example "AI often mangles African languages [...]", [4] "Indigenous groups fear culture distortion as AI learns their languages".[5] The legal and ownership issues in scraping and furthermore repurposing the scraped content by the AI models has likewise been a hot topic in popular discussion in recent years and months, with professional unions like script writers, authors[6], as well as New York Times themselves has also voiced concern recently[7] and visual artists getting ahead of lagging legislation in terms of ethical use of the AI, at least in the American context at the moment.

It should be noted that many of the referred publications talking about issues e.g. in limited language resources being overwhelmed by LLM-generated data (Shumailov et al., 2023) present the problem from the point of view of English, which is the largest language in the dataset; of course the problem is much worse the smaller the language data we talk about.

The Indigenous research group *Te Hiku* in Aoteoroa has done lot of research in ethics of modern

---

[4] `https://www.science.org/content/article/ai-often-mangles-african-languages-local-scientists-and-volunteers-are-taking-it-back` Retrieved: 2023-10-20

[5] `https://www.japantimes.co.jp/news/2023/04/10/world/Indigenous-language-ai-colonization-worries/` Retrieved: 2023-10-20

[6] `https://www.washingtonpost.com/opinions/2023/10/19/ai-large-language-writers-stealing/` Retrieved: 2023-10-20

[7] `https://theconversation.com/the-new-york-times-ai-copyright-lawsuit-shows-that-forgiveness-might-not-be-better-than-permission-222904` Retrieved 2024-03-13

language technology and Indigenous language resources. Their focus is on legal issues and practical ways of handling data ownership and governing the ethical use of Indigenous corpora via copyright and end-user license agreement approach.[8] The license approach attempts to force ethical use and co-operation with Indigenous people all steps of the way by use of copyright and license deals, this is a good strategy, but it limits a lot of legitimate open source use since it is not practical for all open source users to contact team of few people everytime they use an open source product. On the other hand, the big actors who have already harvested this data from the internet ignoring existing copyright and intellectual property rights are unlikely to care or begin to care about this new license. While this article does touch on the legal issues as well, it is not our overall focus; moreover, we give a best common practice for our work and suggested actions for people who are using the lexical resources.

There exists several corpora that are specific for the Indigenous languages of the world, for example for Sámi languages there is the Sámi international corpus SIKOR (2018).

3. Neural 'assumptions' about corpora

Data driven natural language processing often has the starting point that corpora encode the language knowledge in itself; if you just get large enough data you will get a representative view of the language that can be learned from or encoded in statistics or neural networks. This has seemed to work well for several of the biggest majority languages but often faces problems when the same methodology is applied to smaller languages. Nevertheless, the expectation seems to be that it is solely a question of data quantity: when we get more data the problems start to disappear. However, there are several limitations to this both on quantity and on quality: while the quantity of data required by neural approaches is getting smaller over time, it is still often a prohibitively large amount that is realistically required for a starting point and a large part of available corpora does not necessarily encode grammatical, idiomatic language usage; scaling this will lead to worse results, not better. From quantitative point-of-view scaling is simply impossible, because all corpora have been used and the amount of people with necessary language skills and time to write is limited, e.g. all 700 speakers would need to write non-stop for 10 years to achieve a representative corpus that is needed for baseline large language model. Qualitatively, texts on languages with small speaker communities are often of bad linguistic quality. A case is point is the set of cirucmpolar Wikipedia editions. As shown in Trosterud (2021), only 7 out of 45 circumpolar Indigenous Wikipedias had native speakers as authors. Since Wikipedia text often is the staple food of language models, this is a highly disturbing finding. From a normative point-of-view using corpus collections as a substitute for a language standard presupposes that the collection actually represents the standard. For Indigenous languages, this is not the case. For a discussion, see Trosterud (2022).

One problematic aspect that is not limited to lesser resourced languages is the reliance on text materials to encode the language use and its features; it is true perhaps for all spoken languages that the written form especially in the corpora we usually have available in free and open source format is a limited subset of the language.

Icelandic was reported to possess corpus resources of 2.7 billion words (Rögnvaldsson, 2023). Although this is only 0.5 % of what was available for ChatGPT 3.0 (Brown et al., 2020), it still seems to have resulted in a usable LLM, perhaps representing some lower boundary. We do not know of any usable LLM built on smaller corpora. Much smaller corpora are needed for uses other than generative models, the Tartu neural MT program for Uralic languages(Yankovskaya et al., 2023) is based on only a fraction of the Icelandic corpus size.

The philosophy behind the machine learning paradigm is amply described by Cuckier and Mayer-Schoenberger (2013). 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.

4. The making of indigenous language corpora

The main factor determining the amount of available text for a language is its number of speakers. Table 1 shows the size of available corpora for languages in the Nordic countries, as collected by the respective national language banks, and for the minority languages, by UiT The Arctic University of Norway. Measured by words per speaker, Inari Sámi is a league of its own. Extrapolating the same productivity level for the other languages gives the final row of the table. Note that the reason why Norwegian beats the other national languages is that the Norwegian National library has digitized every single publication published in Norway. This sets an upper bound for the possible available text in the future.

If we try to simulate for Sámi a language quality

---

[8] `https://tehiku.nz/te-hiku-tech/te-hiku-dev-korero/25141/data-sovereignty-and-the-kaitiakitanga-license` Retrieved: 2023-10-20

	Finnish	Swedish	Norw.	North Sámi	South Sámi	Kven	Inari Sámi	Lule Sámi	Lovari Romani
Corpus, million words	14,100	14,400	20,000	35	2	0,5	3.16	1.18	0.05
Words / speaker	2,350	1,440	4,000	1,750	4,000	100	10,533	905	-
If all published like the Inari	60,000	100,000	50,000	200	5	50	-	20	-

Table 1: Corpus size for different languages, in million words

situation we have for English we would have to first either remove or correct a lot of text due to the large amount of spelling errors.

A large survey of spelling errors in Icelandic, Greenlandic and three Sámi languages (Moshagen, 2014) shows that the amount of orthographic errors in South Sámi was 9.02 % (N = 181.701 and 5.2 % for Lule Sámi, as compared to 0.65 % for Icelandic (N = 163.702). With an average sentence length of 20 words this amounts to two errors per sentence for South Sámi and one per sentence in Lule Sámi, as compared to one error in every eighth sentence for Icelandic. Another study for Lule Sámi, including also grammatical errors, gave 12 % of orthographic *and* grammatical errors (Wiechetek et al., 2022). If this number is representative for South Sámi as well, the total amount of orthographic and grammatical errors would be 4 per sentence. It is clear that using corpora made up of these texts would result in unreliable normative tools.

4.1. Noise in Indigenous language corpora

In Tables 2 and 3 we describe the the corpora of Indigenous Sámi languages with special attention to the errors within; Table 2 shows that potentially 5—15% of the sentences contains a non-word type errors, we have not seen similar estimations for e.g. the English gigaword corpora, but based on quick sampling with automatic spell-checkers we can estimate the figure be closer to $< 1\%$. Table 3 shows the amount of texts that has been carefully error annotated and corrected, which already represents a large amount of expert work. The annotated grammatical errors are based on the work on annotating errors we have done earlier on and can be the more common errors in the corpus, but it should give an impression on the frequency of grammatical errors in the corpus at large.

Alternatively, they can be marked-up, corrected and linguistically analyzed. It is also a resource of the language-variation there is, a source for error-analysis.

For example, it is not feasible to read through and annotate whole corpora crawled from the internet for spelling errors, grammatical errors and so forth. However, leveraging rule-based models of the normative language can be helpful in curating proper sub-corpora.

ISO-code	Language	Sentences	Sentences with errors
sme	**North Sámi**	4,647,443	9.4%
smj	**Lule Sámi**	225,138	10.6%
sma	**South Sámi**	302,640	13.4%
smn	**Inari Sámi**	418,475	5.0%

Table 2: Corpus statistics for Sentences with ratio containing non-word type errors.

ISO-code	Language	Tokens	Marked-up errors
sme	**North Sámi**	116,623	17,221
smj	**Lule Sámi**	38,421	5,065
sma	**South Sámi**	33,222	12,681
smn	**Inari Sámi**	4,282	305

Table 3: Corpus statistics for select languages we have worked on; the number of tokens in error-annotated corpora and the portion of errors within.

4.2. Ethics of collecting Indigenous corpora

Collecting Indigenous corpora involves a number of steps that touch on ownership rights. We will illustrate this on the example of building a Sámi corpus *SIKOR*. From at least 2006[9] and on UiT and the Norwegian Sámi Parliament have been gathering corpus texts, making the corpora searchable[10] using the *Korp* interface[11]. Due to copyright issues, the corpus was divided in two parts, a public part containing texts mostly in the public domain (mainly bureaucratic texts) and a closed part containing copyrighted material. The whole process of collecting corpora required economic and human resources on the one side, and more than anything a relation of trust. Getting access to copyrighted material involved a bureaucratic process, i.e. agree-

---

[9] `https://github.com/giellalt/corpus-sme-orig/commit/901d9026a9df77c8cf1e8fec8e0c5c86656cc500`

[10] `https://giellalt.github.io/lang/common/Korp_usage.html`

[11] `https://github.com/spraakbanken/korp-frontend`

ments had to be made with publishers, newspapers, public broadcasters and individual writers. In spite of accomplishing initial agreements, both practical and communication problems slowed down the process of collecting texts. In the case of SIKOR, the developers of NLP tools are authorized by the Sámi society and their political leadership themselves, and bound by the promise to make publicly available tools for the language community. On this background, most writers were positive to sharing their work, as long as their work is only used for the development of the tools, but not made publicly available in their entirety.

The success story *per se* was an agreement with the Sámi newspaper, *Ávvir*[12], which donated its complete volumes from 1996 onwards. Texts which did not require agreements were mainly from one genre, predominantly translated, hence colored by the source language, and additionally rare, altogether not sufficient for a balanced and authentic corpus. These were mainly governmental and Sámi parliament texts. The sites of these institutions were crawled, and the language versions belonging together were connected using metadata, making it possible to create parallel corpora.

5. Ethical recipe for use of corpora

An ethical recipe for using Indigenous corpora in the making of data-driven language technology tools requires various considerations and reservations. We will in the following use South Sámi as a showcase. The previous section has illustrated the role of trust in collecting corpora, the economical and human resources required in this process, and the importance of selection and mark-up by language-experts to identify noise and general quality of texts. Previous work needs to be credited and agreements need to be kept also in secondary use of the corpora. Due to the fact that corpora are noisier than for other languages, particular attention needs to be paid to the quality of the language, meaning lexical exactness, idiomaticity (as opposed to literal equivalents that resemble the majority language and are a sign of bilingualism and interference), correctness of spelling and grammar (which can be given to a lesser extent due to the same reason). Apart from purely linguistic considerations, also cultural linguistic considerations need to be taken into account. Cultural relevance is another key factor. In a query asking for South Sámi-specific information in South Sámi language, the answer should be insider-oriented and not a mere translation of the outsider-majority perspective. Insiders are interested in other facts than outsiders.

In the following we will illustrate with two examples (one from language generation and the other one from machine translation) to illustrate which scenario we would envision and which scenario we find problematic. The language generation example is based on ChatGPT (Brown et al., 2020)[13] and the MT example based on Tartu NLP Neurotõlge (Yankovskaya et al., 2023).[14]

5.1. Language generation

Cultural relevance in language generation means that, depending on the language used for the query, the answer should be relevant for the language community in question. If ChatGPT is asked about the South Sámi area, the insider view should not resemble the English answer given by ChatGPT in Figure 1.[15] The English answer is very short and refers to the geographical area only by means of place names. While this information is relevant to an outsider that visits the area, an insider finds other information relevant, while insiders find e.g. *siida*-structures related to reindeer husbandry more relevant.

The desired insider-output would be of the following kind, illustrated in Figure 2.

As Figure 2 illustrates, insider South Sámi knowledge requires reference to Sámi family areas, *Sijte*, their family names, dialectal areas, reference to important Sámi people with language, economics, and politics with reference to their family names and area. Other interesting facts that should be covered are South Sámi centers, schools, history, political events and reindeer husbandry.

5.2. Machine Translation

In the case of corpora-based machine translation, on the other hand, linguistic form and exactness of content is the foremost objective. An article from Guardian[16] about recent South Sámi political events, i.e. protests against windfarms in a major South Sámi reindeer herding area and human rights violations by the Norwegian government, translates to South Sámi in a problematic way when translated using *TARTUNLP Neurotõlge*'s LLM,[17] cf. Figure 3.

Table 4 shows very clearly the immense amount of fails just within the two first paragraphs of the article.

The South Sámi text reveals that the content is about campaigns – Indigenous people did something in Oslo - but what did they do? Taking down

---

[12] `https://avvir.no`

[13] `https://chat.openai.com`

[14] `https://translate.ut.ee`

[15] ChatGPT 3.5 asked 16.10.2023, 11:30, `https://chat.openai.com/`

[16] `https://www.theguardian.com/world/2023/oct/11/demonstration-in-oslo-seeks-removal-of-windfarms-in-Indigenous-region` retrieved 18.10.2023

[17] `https://neurotolge.ee`

Translation fail	Example
Morpho-syntactical errors	
Case error (noun)	*aalkoealmetji* ‘Indigenous people’s’ > *aalkoealmetje* Gen. Pl.>Nom. Sg.
(pronoun)	*mejtie* ‘which’ > *mij* Acc. Pl.>Nom. Sg.
	reerenasse ‘government’>*reerenassese* Nom. Sg.>Ill. Sg.
	nænnoestæmman >*nænnoestimmeste* Ill.>Ela.
	protesten ‘protest’ >*proteste* Gen.Sg.>Nom.Sg.
Numeral treated as article	*aktem åejviehaerniem*>*åejviehaerniem* ‘one main X’>‘a main X’
	akten sjïere nænnoestæmman > *sjïere nænnoestæmman*
Active > Passive	*gohtjeme gohtjesovveme*
Wrong tempus	*feerhmedh* ‘embrace’ > *feerhmeme* Inf>Past participle
Word order	*rïjvestidh göökte bïegkefaamoeh* > *göökte bïegkefaamoeh rïjvestidh* VO>OV
Lexical issues	
Wrong word semantics	*byjresekampanjh* ‘environmental campaigns’ > *byjres* ‘environmental campaigners’
	bïegkefaamoeh ‘wind power’ > *bïegkefaamoehpark* ‘wind power park’
	bïegkefaamoevierhtieh ‘wind power ressources’ >*bïegkefaamoejårrehtsh* ‘wind power turbines’
	byjresekampanjh (should be human)
Semantic inadequacy	*gaavnehti* ‘find out’ > *nænnoesti* ‘decide, implement’
Nonsense words	*åejviehaerniem* ‘main ?’ > *åejviegæjnoem* ‘main road’
	biegkemeurhkedh > *dahpeme* ‘close’
Literal translation of idiom	‘traces its roots’ – *roehtsem geasa* ‘roots drag’ > *vualkove lea* ‘the origin is’
Non-standardized version	*Oslosne* > *Oslovisnjie*
Missing parts / Redundancy	
Missing content word	> ∅ ‘to demand’ > *krïebpesjidh*
Missing logical connector	> ∅ > *dannasinie* ‘because’
Significant meaning changes	
opposite semantics/value	*feerhmedh* ‘embrace’ (positive) > *gaertjiedamme* (negative)
	viehkine ‘with help’ (positive) > ∅

Table 4: Classification of the errors in English-South Sámi machine translation.

wind power – does that mean the abstract thing? Also, the Norwegian government has called out others for breaking human rights - but we know that in reality the Norwegian government is violating human rights.

The conclusion is that a South Sámi speaker can get a general impression of the subject area, the involved parties and the place of action, but details of these actions are unintelligible, and the worst is, the central message is not conveyed, instead the translation gives the opposite meaning. The Norwegian government stands out as a helper instead of the culprit.

When the user is proficient in both the source and target languages and is aware of the potential for such errors, their impact is unlikely to be significant. However, if the South Sámi output is used uncritically and simply published, this would have catastrophic consequences and it is unethical.

Additionally, careful reading of neural machine translation output is not an innate qualification but requires training and strong language intuitions. In a bilingual society, where typically the majority language has taken over a large amount of domains and marginalized the domains of the Indigenous languages, even manual translations can be heavily colored by the majority language. The following examples from South Sámi children’s literature shows how even highly qualified translators can choose translations that are literal translations of a Scandinavian language to the disadvantage of efficient and adequate South Sámi constructions. ‘Then Stæjna comes walking with big steps.’ is translated with an adjective and a noun phrase construction (just as in English) as in ex. (5-a) whereas South Sámi lexicalizes ’with big steps’ in a single adverb *voepsijeslaakan*, cf. ex. (5-b).

(5) a. Daelie Stæjna båata
then Stæjna come.PRES.3SG
vaedtsien **stoerre sïlligujmie**.
walking big step.PL.COM
b. Daelie Stæjna **voepsijeslaakan**
then Stæjna with.big.steps.ADV
båata vaedtsien.
come.PRES.3SG walking

Another children’s book’s translation uses the

Where is the South Sámi area?

“ChatGPT
The South Sámi area is primarily located in the northern parts of Sweden and Norway. It includes regions in both countries, and the South Sámi people are Indigenous to these areas. In Sweden, South Sámi communities are found in areas like Jämtland and Härjedalen, while in Norway, they are located in regions such as Trøndelag. These regions are part of Sápmi, the traditional Sámi territory, and are known for their unique Sámi culture, language, and traditions.”

Figure 1: ChatGPT English output to a question regarding an Indigenous area

Mij lea åarjelsaemien dajve?

Åarjelsaemien dajvesne lea åarjagommes dajve Saepmesne, gusnie leah 16 båatsoesïjhth, jïh lea juakasovveme gïelen, golme smaarehtjïerti dajvi mietie, maadtoej jïh båatsoej mietie. Åarjelsaemien dajve lea åarjagommes Saepmesne jïh lea gööktene rïjhkesne, Sveerjesne jïh Nöörjesne Skandinaavijisnie. Staati politihke leah tjuerpieslaakan gïetedalleme åarjelsaemien gïelem, kultuvrem jïh jielegem, guktie daelie garre aejhtemen nuelesne.

The South Sámi area is the southernmost area of Saepmie, which has 16 siidas and can also be divided according to three dialect areas (Northern, Mid, Southern), following families and reindeer herding traditions. The South Sámi area is in the states of Norway and Sweden in Scandinavia. The policies of these states have dealt harshly with South Sámi, South Sámi language, culture and business, so that South Sámi is today very endangered.

Figure 2: ChatGPT idealized insider-output in South Sámi and its translation to English

inadequate cognate for ‘door’ in the event of going through a door, and also translates the adpositional construction in the Scandinavian languages (and English) directly to an adpositional phrase in South Sámi. However, the meaning of this is not ‘walking’ but ‘crashing’ through the door. The South Sámi idiomatic way of saying it uses the word *okseraejkiem* ‘door frame’ in accusative case as in ex. (7) rather than *oksen tjïrrh* ‘door through’ as in ex. (6)

(6) Dïhte **oksen** **tjïrrh** båata.
PRON.3SG.NOM door.GEN through come.crashing.3SG
‘She/he/it crashes through the door’

(7) Dïhte **okseraejkiem** båata.
PRON.3SG.NOM door.frame.ACC come.3SG
‘She/he/it comes through the door.’

When efficient idiomatic South Sámi constructions are replaced with less elegant word-by-word translations from the majority language, part of the authentic South Sámi linguistic structure gets lost.

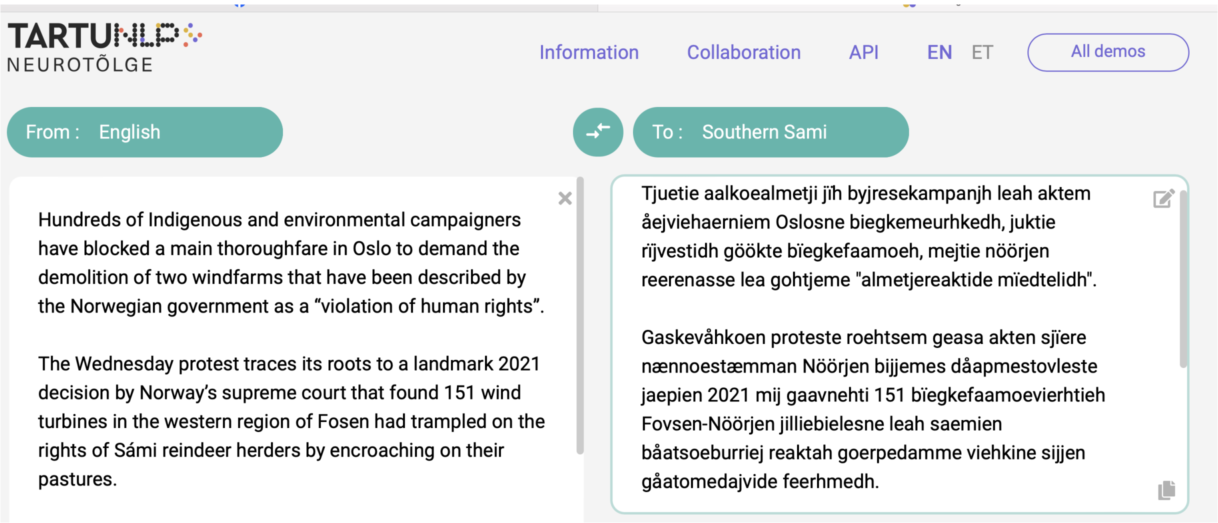
The work of a human translator is very challenging in the face of bilingualism because language happens unconsciously. Especially when there is a lack of external Indigenous language input their own language intuitions get colored by the majority language. When neural machine translation suggests syntactic and idiomatic structures that are heavily derived from majority language input this can reinforce the tendency towards language shift towards the majority language. Linguists and language experts, on the other hand, can contextualize, analyze and see constructions from an analytical standpoint and relate them to other Uralic languages, giving them a stronger starting point. Consequently, a deliberate standardization process informed by a linguistic analysis of the language’s dominant structures holds immense value.

6. Conclusion

With the emerging LLM technology a lot of recent research and development on minority and Indigenous languages is done by outsiders without knowledge of the languages in question, solely based on data harvested from the net.

In this article, we have illuminated how large language models make implicit assumptions that do not match the reality of Indigenous languages on various levels and produce results that can be problematic, both from an ethical and linguistic perspective. We have have shown this in examples from both corpus-based language generation and machine translation for the endangered South Sámi language.

As tools for language generation and machine



Hundreds of Indigenous and environmental campaigners have blocked a main thoroughfare in Oslo to demand the demolition of two windfarms that have been described by the Norwegian government as a “violation of human rights”.

The Wednesday protest traces its roots to a landmark 2021 decision by Norway’s supreme court that found 151 wind turbines in the western region of Fosen had trampled on the rights of Sámi reindeer herders by encroaching on their pastures.

Tjuetie aalkoealmetji jïh byjresekampanjh leah aktem åejviehaerniem Oslosne biegkemeurhkedh, juktie rïjvestidh göökte bïegkefaamoeh, mejtie nöörjen reerenasse lea gohtjeme "almetjereaktide mïedtelidh".

Gaskevåhkoen proteste roehtsem geasa akten sjïere nænnoestæmman Nöörjen bijjemes dåapmestovleste jaepien 2021 mij gaavnehti 151 bïegkefaamoevierhtieh Fovsen-Nöörjen jilliebielesne leah saemien båatsoeburriej reaktah goerpedamme viehkine sijjen gåatomedajvide feerhmedh.

Figure 3: *TARTUNLP Neurotõlge* English-South Sámi MT

translation are high up on the wish list of the Indigenous language communities we have worked with, we have provided an ethical recipe for developers, to ensure a digital future as it is wanted by the Indigenous language community. Harvesting corpora in itself is a time-consuming process involving a trust-based relation between developers and owners of the language data. Respecting intellectual property and crediting authorship is mandatory for both primary and secondary developers.

Language data found in the internet is of varying quality, and developers need to distinguish between original text and translated text in order to ensure that the generated language is authentic, and not just, as in our case, “Norwegian or Swedish with South Sámi words”. Considering monolingual users, it is also essential to reduce significant translation errors that could convey meanings contrary to the original or disseminate incorrect information.

In practice this means that language experts and community insiders need to be present to evaluate the output and perform quality assurance. More clearly, the Indigenous language community itself needs to be part of the process and make decisions on authentic language and desired output.

Ethical NLP for any language, and for Indigenous languages in particular, must ask for the needs of the language community and adapt the tools accordingly.

Ethics Statement

As an article about ethical ways of working with indigenous language technologies the article as a whole describes ethics such that it does not need repeating here, as a summary´:

We discuss the use of Indigenous linguistic data from within the language community. One of the authors is a member of the South Sámi community, and a mother tongue speaker.

The method of ethical data management we lay out in this article is aimed at more sustainable and just management of data. We have implemented it in our own workflow and development process.

Bibliographical References

Lene Antonsen and Trond Trosterud. 2020. Med et tastetrykk. bruk av digitale ressurser for samiske språk. volume 12 of *Samiske tall forteller. Kommentert samisk statistikk*, pages 43–68. Sámi allaskuvla, Kautokeino.

Knut Bergsland. 1982. *Sydsamisk grammatikk*. Universitetsforlaget, Oslo.

Knut Bergsland and Laila Mattsson Magga. 1993. *Åarjelsaemien-daaroen baakoegærja*. Idut.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics*, pages 3504–3519.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? opportunities and challenges for participatory ai. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO ’22, New York, NY, USA. Association for Computing Machinery.

Rogier Blokland and Cornelius Hasselblatt. 2003. *The endangered Uralic languages*. John Benjamins Publishing Company.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Kenneth Cuckier and Viktor Mayer-Schoenberger. 2013. The rise of big data. *Foreign Affairs*, pages 28–40.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Lajla Mattsson Magga, Albert Jåma, and Sigrid Stångberg. 2009. Nektelse i preteritum. Gïelenænnoestahke 2009-12-09, Saemien gïeleraerie, Kautokeino.

Sjur Moshagen. 2014. Gold standards and evaluations: an automated and extendable speller evaluation test bench.

Eiríkur Rögnvaldsson. 2023. Language report icelandic. In G. Rehm and A. Way, editors, *European Language Equality*, pages 159–162. Cognitive Technologies.

Pekka Sammallahti. 1998. *The Saami Languages: An Introduction*. Davvi Girji, Karasjohka.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. Model dementia: Generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Trond Trosterud. 2021. The circumpolar Wikipedia editions. In *Proceedings of the Arctic Knot Conference*. Wikimedia.

Trond Trosterud. 2022. Normative language work in the age of machine learning. In Jozić Željko and Sabine Kirchmeier, editors, *The Role of National Language Institutions in the Digital Age Contributions to the EFNIL Conference 2021 in Cavtat.*, pages 61–69. Nyelvtudományi Kukatóközpont, Budapest.

Linda Wiechetek, Katri Hiovain-Asikainen, Inga Lill Sigga Mikkelsen, Sjur Moshagen, Flammie Pirinen, Trond Trosterud, and Børre Gaup. 2022. Unmasking the myth of effortless big data - making an open source multi-lingual infrastructure and building language resources from scratch. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1167–1177, Marseille, France. European Language Resources Association.

Linda Wiechetek and Maja Lisa Kappfjell. 2023. A South Sámi grammar checker for stopping language change. In *Proceedings of the Constraint Grammar Workshop at 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, Tórshavn, Faroe Islands. University of Tartu Library.

Lisa Yankovskaya, Maali Tars, Andre Tättar, and Mark Fishel. 2023. Machine translation for low-resource Finno-Ugric languages. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 762–771, Tórshavn, Faroe Islands. University of Tartu Library.

Jussi Ylikoski. 2022. South Saami. In *The Oxford Guide to the Uralic Languages*, pages 113–129. Oxford University Press.

Language Resource References

SIKOR. 2018. SIKOR uit Norgga árktalaš universitehta ja Norgga sámedikki sámi teakstačoakkáldat. Veršuvdna 06.11.2018. `http://gtweb.uit.no/korp`. Accessed: 2023-06-12.