A Manual Evaluation Method of Neural MT for Indigenous Languages

Linda Wiechetek
UiT Norgga árktalaš universitehta
linda.wiechetek@uit.no

Flammie A. Pirinen
UiT Norgga árktalaš universitehta
flammie.pirinen@uit.no

Per E Kummervold
National Library of Norway
per.kummervold@nb.no

Abstract

Indigenous language expertise is not encoded in written text in the same way as it is for languages that have a long literal tradition. In many cases it is, on the contrary, mostly conserved orally. Therefore the evaluation of neural MT systems solely based on an algorithm learning from written texts is not adequate to measure the quality of a system that is used by the language community. If extensively using tools based on a big amount of non-native language this can even contribute to language change in a way that is not desired by the language community. It can also pollute the internet with automatically created texts that outweigh native texts. We propose a manual evaluation method focusing on flow and content separately, and additionally we use existing rule-based NLP to evaluate other factors such as spelling, grammar and grammatical richness. Our main conclusion is that language expertise of a native speaker is necessary to properly evaluate a given system. We test the method by manually evaluating two neural MT tools for an indigenous low resource language. We present an experiment on two different neural translations to and from North Sámi, an indigenous language of North Europe.

1 Introduction

Indigenous languages with few speakers are often left out in the development of high-level NLP tools that require a lot of data and have therefore not been subject to evaluation either. However, recently neural machine translation has become more effective and more available for even lesser resourced languages than before. While the technology has made the use of neural machine translators plausible, it is not clear whether the quality of the translation really is good enough for the common use cases within language communities. High-resource languages typically apply data-hungry evaluation methods. The demand for big data is known to be problematic for smaller languages. An additional factor is, that while big languages with a long literary tradition have their language expertise encoded in large amounts of written texts, typically this is not the case for indigenous languages with a much shorter literary tradition. Here language expertise is often transmitted orally and may not be reflected in written text at all, partly due to lack of literacy and tradition. It is problematic if we base our knowledge of a language on existing written text for a language community that does not have a long tradition in writing. Written texts need to be treated much more critically with regard to who wrote it (was it even a native speaker?), if it was a translation, and which genre it belongs to. Written texts can have systematic spelling and grammar errors. Their authors can be second language learners instead of language experts, or they can be synthetically created by machine translation programs. Taking into account the distribution of human resource and language expertise is an important factor in the thought process. Language communities that put a great deal of work into preserving and strengthening their language typically use a lot of resources in teaching the younger generation. That also means that expertise may be found to a great deal in oral contexts rather than being reflected in text corpora. Basing evaluation on algorithms that learn from written corpora is therefore a thinking error in these contexts.

Consequently, we find a manual evaluation of neural MT tools by language experts in this context unavoidable. By *language experts* we mean native speakers with a profound understanding of their own language, which allows them to make judgements about the grammaticality and idiomaticity of a sentence. Especially since indigenous written grammars are far from exhaustive, good language intuition is a key qualification.

In this article we suggest a grading system for a language expert evaluator that is an expert of both source and target language. The scale distinguishes between flow and content, where flow (which has a main focus on the target sentence) is evaluated before content (which again requires an analysis of the source sentence). Our main hypothesis is, we need native language/linguistic expertise to even know how good the translation is.

We do a small-scale but detailed manual evaluation of two neural MT tools for an indigenous low resource language (North Sámi). Our aim is to develop a workflow for future evaluations of similar languages and systems and those with even less resources, than the ones we work on, should they become available in the popular NMT toolkits.

2 Background

Methods of evaluating machine translation are often based on two approaches: automatic that requires high quality parallel texts and human-based, which requires a large amount of humans doing annotation or rating of large number of sentences for example. In a low-resource minority language situation, neither of these resources is easily available; there are no parallel texts and very few humans to do annotation or rating. That is to say, the amount of sentence-aligned parallel texts that is needed to automatically verify quality is larger than amount of any translated texts in the language in the foreseeable future and the amount of people required to do a meaningful comparison is well larger than available people as well, it is physically impossible to do perform such tests. The typical automatic evaluation metrics like word error rate require either post-editing or parallel corpora which typically are not available in large quantities in indigenous low-resource contexts.

Thus we will be able to identify the criteria that matter for a good translation of or into the language in question. Based on their feedback, automatic processes to perform an adequate evaluation can be developed.

Also with regard to human resources the indigenous context is a challenging one. Those that are language experts with a linguistic background and a high degree of literacy are typically recruited by schools, media, as translators or any other context where language knowledge is highly sought-after.

Generally, the machine translation use cases can be divided in two main categories: translations that can be read to understand the source texts (assimilation, gisting) and translations that can be edited for further use (dissemination). If the tools are useful as a basis for post-editing has to be decided by members of the the language communities, which is why we also think that feedback from the community is needed to evaluate the quality. Because of the systems' fluency, new machine translation tools tend to get adopted quickly by businesses (e.g. Facebook, Google reviews) and even official bodies. An early and critical evaluation by language community is therefore essential. Machine-learning MT is now almost a standard and being used in every day life without much thought. How does it look like in an extremely low resource language context? (Moorkens et al., 2018)

2.1 Languages

North Sámi is a Finno-Ugric language belonging to the Uralic language family, it is spoken in Norway, Sweden, and Finland by approximately 25,700 speakers (Eberhard et al., 2018). It is a synthetic language, where the open *parts-of-speech* (PoS) — e.g. nouns, adjectives — inflect for case, person, number, and more. The grammatical categories are expressed by a combination of suffixes and stem-internal processes affecting root vowels and consonants alike, making it perhaps the most fusional of all Uralic languages. In addition to compounding, inflection and derivation are common morphological processes in North Sámi. The Sámi languages are typically described as verb heavy languages, with at least few hundred distinct inflectional verb forms (both finite and non-finite, varies a bit based on paradigms and depending on what you include as inflectional). Sammallahti (1998) notes that in a list of the most common North Sámi words, verbs are in first place (33%), followed by 28% nouns. English and Norwegian, on the other hand, are Indo-European languages, with relatively low morphological complexity: less than 10 word-forms per word in productive inflection. The word order in English and Norwegian is stricter than in North Sámi and our hypothesis is that the distribution of parts-of-speech and derivations is different as well. We expect this to have an effect on the translated language and non-translated, as well as different profiles between machine and human translated texts.

The syntactic differences between Sámi and the two Germanic languages are notable. While the neutral word order for all of them is Subject-Verb-

Object (SVO), there are a number of mismatching features in the syntax. Unlike Norwegian and English, Sámi has pro-drop (pronoun dropping) for 1. and 2. person. Sámi uses mostly postpositions as opposed to prepositions. Other differences are adverbial positioning, word order in sub-clauses, question clauses or after adverbial extensions, etc.

2.2 Previous research

There has been a lot of research in the evaluating of machine translation. There are many ways to evaluate the machine translation quality, some are standardised like MQM (Multidimensional Quality Metrics) and others are purpose-built for one specific experiment or study. Lommel (2018) use a very fine-grained system for categorising translation errors. Popović (2018) use a less fine-grained system. OpenAI has used following criteria (Stiennon et al., 2020) for their human evaluation work of a summarisation system, we have taken some inspiration from that, for example in our 7-grade scale for judgments. The machine translation systems we evaluate are based on neural machine translation. The translation system between English and North Sámi is described in Yankovskaya et al. (2023). Mager et al. (2023) have studied machine translation in similar contexts than as we work in.

Human evaluation of machine translated texts often is based on crowd-sourced quick evaluations based on superficial reading of the sentences without context (c.f. WMT shared tasks (Weller-di Marco and Fraser, 2022), AppRaise (Federmann, 2018)). While this kind of quick eyeballing by average language users can give some impression of fluency of the translations it may be insufficient to determine if the text is translated accurately and language is truly idiomatic. A lot of evaluation approaches use scales of fluency and adequacy, in a way to measure separately the overall readability of the text from the accuracy of the translated content.

2.3 Data

The corpora available for a low resource language like North Sámi is very limited. In Table 1 we list the corpora that we have used in the experiments: the largest electronically available Sámi corpus SIKOR (2018) has been used both for training the North Sámi—Norwegian and English—North Sámi machine translation. We did not train the English—North Sámi model ourselves but used TARTUNLP that is partly trained on *SIKOR*, cf. Section 3.2.

We also use part of *SIKOR* to calculate the linguistic features of non-machine translated, open domain texts. *Alice in Wonderland*[1] (henceforth referred to as ‘Alice’; we evaluated here the first three chapters), CTV.ca news item: *What’s behind the increase in orca-human interactions, boat attacks?* (CTV), BBC.co.uk news item: *Multicancer blood test shows real promise in NHS study* (BBC) and *ILO-169 declaration of indigenous peoples’ rights*[2] (ILO-169) are texts we have manually harvested from the internet and represent different genres: fiction, news texts in two variants of English and a legal / political text respectively. These texts were used as sources for machine translation from English.

Corpus	**Size**
SIKOR	23,923,558
Alice in Wonderland	3,509
CTV	722
BBC	413
ILO-169	2,978

Table 1: Sizes of corpora in simple, space-separated tokens (`wc -w`).

The data used for training the Sámi—Norwegian machine training system is described in 3.1.

3 Methods

Despite limited amount of corpora North Sámi has in recent years gained some experimental neural machine translators. By evaluating their current state-of-the-art we present a manual evaluation method and relevant criteria. As a test case we looked at one system to and another one from North Sámi.

Previously North Sámi has been unreachable for neural approaches to language technology due to low resourcedness. The majority of resources are therefore rule-based tools. For machine translation, language pairs included other closely related Sámi languages, as well as Finnish, which is in same language family, but not closely related. There also exists translators for Norwegian, which is another majority language in North

[1] https://www.gutenberg.org/ebooks/11

[2] https://www.ilo.org/dyn/normlex/en/f?p=NORMLEXPUB:55:0::NO::P55_TYPE,P55_LANG,P55_DOCUMENT,P55_NODE:REV,en,C169,/Document

Sámi territory. Many of the existing majority-to-minority language translators are primarily developed in one direction first (Trosterud and Unhammer, 2012). The rule-based machine translators are based on other language technology resources, such as dictionaries, morphological analysers, syntactic analysers and so forth. We use these morphological analysers, as well as spell-checkers and grammar checkers as tools to find out if there are differences between the human and machine translated texts for potential spelling errors, grammatical errors as well as differences in distributions of the grammatical features. The systems for linguistic analysis and grammar and spell-checking have been acquired from the GiellaLT infrastructure[3], that contains freely available open source language technology tools for minority languages (Pirinen et al., 2023).

We used the existing neural machine translation systems as a black box, we fed in the source texts and evaluated the target translations without post-editing in between; only the cases where formatting went destructively wrong (line breaks and spaces added or disappeared in unusual places, like intra-word spaces) were corrected.

3.1 North Sámi to Norwegian NMT

In the development of the North Sámi—Norwegian machine translator, we utilized a standard sequence-to-sequence model based on mT5 (Xue et al., 2020). Our starting point was the pretrained NorthT5 checkpoint[4], a checkpoint that is additionally pretrainedfrom the mT5 checkpoint using additional Scandinavian and English data. Notably, while both these are multilingual models, North Sámi is not included in the listed training corpus.

We retrieved a set of bilingual translations from *SIKOR*. This was divided into a train and test set, and we proceeded to fine-tune a translation model on the train set with 3,800 parallel North Sámi—Norwegian sentences for 10,000 steps. After training, the model was applied to translate sentences in the test set, and a professional translator evaluated the output. As mentioned earlier, human resources are limited, which is why finding even a single adequate evaluator can be difficult.

3.2 English to North Sámi NMT

The English-North Sámi machine translation was built by university of Tartu NLP group as a part of their low resource Uralic neural machine translators[5] and it is based on North Sámi corpus SIKOR (2018) and its parallel parts have been used to train the machine translation (Yankovskaya et al., 2023). The output was analyzed by our rule-based tools. Hand-picked examples show shortcomings of the system. As we were short on human resources for this task, i.e. language experts, we were not able to apply the same method as for North Sámi to Norwegian.

4 Evaluation method

We evaluate separately for the from and to North Sámi scenarios.

4.1 North Sámi as a source language

We study the evaluation of the translations by a language expert. We want to gain an insight on how useful the translated texts are for their use cases within the speaker community: for the speakers who are proficient in the source and target languages with different levels and aims, and relevant to the user experience. We expect that the results of the neural machine translation may partially reflect the style and features of the available corpora in the language, which is not necessarily representative of the norms and standards in the same proportion as with largely resourced majority languages. We also study to what extent the translated texts look translationese versus texts written by native speakers. The commonly translated languages in a neural MT setting at the moment are Indo-European majority languages: English, Norwegian etc., that are in a whole different language family, it is possible that this reflects in the (machine) translated texts more heavily. As it is well-known that neural machine translations get more fluent-looking before they get content-accurate, we also attempt to study how expensive it is to evaluate the translations on this. A professional translator with North Sámi and Norwegian as her native languages evaluated the machine translation from North Sámi to Norwegian described in Section 3.1.

For evaluation we developed a 7-level scale for two main criteria inspired by the scale automatic summaries described in Stiennon et al. (2020, p.23)

[3] https://github.com/giellalt/

[4] https://huggingface.co/north/t5_large_NCC

[5] https://translate.ut.ee

and based on initial comments on translation quality of our professional North Sámi translator. In developing categories for MT evaluation and looking at actual translations we found to main categories: flow and content. First reactions to the quality of a translation typically focus on the output and if there is a good flow in the target language, rather than meticulously comparing the input to the output. However, when knowing the source language in addition to the target language, one will have a second look at the source sentence, and be more critical to the well-sounding translation when parts of the source sentence are missing or incorrectly translated.

A professional translator who is trained in exactness, idiomaticity, and polysemy will quickly be able to identify not only critical errors that change the whole meaning of the sentence, but also other errors that reduce the quality of the translation.

We will therefore distinguish between the first impression of the output with regard to idiomaticity, grammatical and semantic coherence of the text on the one hand, and the exactness of which grammatical structures and content are transferred from the source language into the target language on the other hand. In order to get an unbiased result, the method is the following:

1. read the target translation and evaluate the flow
2. read the Sámi translation and decide on the quality of the translation of the content

The score of 1 stands for the worst possible result, while a score of 7 stands for the best possible result.

The scale for flow is the shown in Table 2. Candidates for flow errors are agreement, valency and word order errors, errors in definiteness, missing articles, morphology and spelling errors, punctuation errors, missing conjunctions and non-idiomaticity.

Grade	Description
7	Perfect flow
6	Good flow (nothing stopping it)
5	Spelling error, smaller idiomatic error
4	Grammatical error, bigger idiomatic error
3	Several grammatical/idiomatic errors
2	A lot of grammatical/idiomatic errors
1	Sentence is unintelligible, cannot be understood or unrelated to the original

Table 2: Flow grades and descriptions.

The scale for content is shown in Table 3. Error candidates are (central) verb meanings in either sub-clause or main clause, where a the meaning difference is not a slight connotation deviation as it would be with synonyms, but a bigger lexical error. Secondly participants, which change the content of a sentence. If a sentence about reindeer would suddenly refer to dogs instead, the meaning of the sentence would be critically changed. Other critical errors can involve time and place errors or errors in quantities and temporal descriptions. Lastly, relevant extra content or missing content.

Grade	Description
7	Perfect, translation contains every single detail and translates it accurately
6	Good content (good enough synonyms)
5	Smaller content errors of the type above/missing information, extra content
4	Big content error/missing information
3	Several big content errors/missing information
2	A lot of big content errors/missing information (more than 50% of the sentence)
1	Nothing is as it should be, translation is (almost) unrelated to original (more than 90% is incorrect)

Table 3: Content grades and descriptions.

The human translation of ex. (1) is exx. (2-a) and the (2-b).[6] In a blind evaluation, the evaluator gave good flow scores to both (6) and slightly better content scores to the neural translation (5) than the human translation (4). *verddevuođa sullasaš ortnegat* is translated into ‘the same system with ear clips’ which includes extra information compared to the more literal neural translation saying ‘verde-like relations’. This yields several issues:

1. If we only evaluate one sentence at a time, we may not get contextual information, where simply the distribution of content onto different sentences is different in manual translation.
2. Automatic translation evaluation based on parallel corpora will have to take into account that the output sentence may be of better quality than the target sentence.

(1) Departemeanta deattuha
department.N.SG.NOM accentuate.V.PRES.3.SG
ahte vejolašvuohta addit
that.C possibility.N.SG.NOM give.V.INF
sierralobi ii
special.dispensation.N.SG.ACC not..V.NEG.3.SG
galgga mielddisbuktit ahte
shallV.CONNEG entail.V.INF that.C

[6] Linguistic examples follow Leipzig glossing standards: https://www.eva.mpg.de/lingua/resources/glossing-rules.php

verddevuođa sullasaš
verddevuohta.N.SG.GEN like.A
ortnegat galget
arrangement.N.PL.NOM shall.V.PAST.3.PL
fas ásahuvvot.
again.ADV build.V.PASS.INF.

(2) a. The department would like to emphasise that the possibility to give special dispensations should not lead to that the same system using ear clips should be reestablished.

b. The departments accentuates that the possibility to give special dispensations should not lead to a reestablishment of *verde*-like relations.

Ex. (3) is a good example where the flow in the neural translation is good (6), and content scores low (2) in the neural translation in ex. (4-b). The reason for that is missing of substantial content, i.e. a translation of *Almmolašvuođagažaldat ja oktavuohta dábálaš láhkaprosedyraide*.

(3) Almmolašvuođagažaldat ja
publicity.question.SG.NOM and
oktavuohta dábálaš
relation.N.SG.NOM normal
láhkaprosedyraide leat
legal.procedure.PL.ILL be.V.PRES.3.PL
guovddážis dán dáfus.
central.SG.PX3SG this.SG.GEN context
'Publicity questions and relations to normal legal procedures are in the center in this context.'

(4) a. The issue of publicity and the relationship with ordinary legal procedures is central in this context.

b. This is a core point in this context.

Table 4 is based on 34 sentences and sentence fragments. It shows only slight differences between human and neural translations. It is however revealing that even human translations do not get perfect scores. This means that automatic evaluations that contrast machine vs. manual translations will not necessarily be able to make judgements about the machine translation quality, but only its similarity to the (possibly bad) human translation. One important factor that was revealed while discussing the evaluation was that in many cases sentences cannot be adequately evaluated without their context as certain terms only get their meaning from the context in which they are used. Therefore, an evaluation of out-of-context sentences' MT test sets can never be entirely satisfactory.

	Neural MT	Human
Flow	5.8	6
Content	5.5	5.6
Average	5.6	5.8

Table 4: Score for neural vs. human ML evaluation

4.2 North Sámi as a target language

For North Sámi as target language, we use the Tartu neural machine translation system for Uralic low resource languages by Yankovskaya et al. (2023). We picked samples from different genres, fiction, news, legal texts, and evaluated these both manually and with our rule-based tools. The only text in our corpora that has pre-existing translations for both North Sámi and English is the ILO declaration.

Ex. (5) from a news text is translated into ex. (6). Flow scores 3 and content scores 4. Flow is negatively affected by both, a case error and word order error *boazodolliid (Gen Pl) orohagaide (Ill Pl) > orohaga (Gen Sg) boazodolliid (Pl Acc)*. In addition, the output sentence contains a non-idiomatic term / lexical error *bohccofuođđut* (Nom) which should be *bieggaturbiinnaid* (Acc). It also involves a case error.

(5) "I have apologised (today) on behalf of the government to the reindeer herding districts for the fact that the permits (to build wind farms) constituted a violation of human rights," Aasland told a at a news conference.

(6) Mun lean áššáskuhttán (dál)
I.NOM have.PRES.SG1 accuse.PTCP (now)
ráđđehusa bealis
government.SG.GEN side.SG.LOC
boazodolliid orohagaide
reindeer.herder.PL.ACC dwelling.PL.ILL
dan go lobit
it.SG.GEN QST permit.PL.NOM
(bohccefuođđuid hukset)
(wild.reindeer.PL.GEN build.INF)
ledje olmmošvuoigatvuođaid
have.PAST.3.PL human.right.PL.ACC
rihkkun," Aasland
violation.SG.GEN," Aasland.SG.NOM,
muitalii ođaskonferánssas.
tell.PAST.3.SG news.conference.SG.LOC.

‘I have accused (now) on the side of the government the reindeer herders dwellings as the permits (to build wild reindeer) were a violation of the human rights,” Aasland told on the news conference.’

We evaluate the translations on linguistic level using several approaches. We use spelling checking and correction to find out where machine translation has created non-words and whether those are near to right words by automatic spelling corrections, we also use grammatical error correction to find out some of the grammatical errors and suspicious constructions the MT system has constructed, we evaluate the errors found this way using linguistic and language understanding. We also calculate some linguistic metrics such as morpho-syntactic form distributions from the translated texts and compare those to texts that are not machine translated; to see if machine translation uses same kind of word-forms and grammatical structures as non-translated or professionally translated texts.

As is expected, the output text of *Alice* involves a number of non-word and probably also real word spelling errors, the latter of which are not handled entirely by the grammar checker yet. There are several spelling errors such as **teleskopa* for *teleskohpa* and **beallahemiin* for *bealjahemiin*.

Grammatical errors include incorrect attributive forms such as **golmmageardánis* for *golmmageardán* in ex. (7), although here the main error is a lexical error. Three-legged in the original sentence ex. (8) is translated with *golmmageardánis* ‘three-times’.

(7) Fáhkka son bođii unna
suddenly s/he.NOM come.PAST.3.SG small
golmmageardánis beavdái, buot
three-times.SG.LOC table.SG.ILL, all
duddjojuvvon čavga *glássas
craft.PASS.PTCP tight glass.
‘Suddenly she came to a three-time table, all crafted in tight glass.’

(8) ‘Suddenly she came upon a little three-legged table, all made of solid glass’

In ex. (9), both flow and content are affected. The sentence sounds weird as such even from a logical point of view as to using future tense and the adverb *ikte* in the same sentence. The comparison with the source sentence (10) shows that the adverb is a wrong translation of *never* and *fall* is wrongly translated as *čakča* ‘autumn’ instead of a form of *gáhččat* ‘to fall’. I.e. when translating a word with polysemy to a target language without the same polysemy, the MT system fails. The verb *loahpahuvvat* has a spelling error, it should be loahpahuvvot and is therefore erroneously analyzed as a compound noun with possessive suffix ending instead of as a passive verb.

(9) Boahtá go čakča
come.PRES.3.SG QST autumn.SG.NOM
ikte loahpahuvvat?
yesterday be.finished.SG.NOM.PX2SG?
‘Will autumn be finished yesterday?’

(10) Would the fall never come to an end?

Table 5 shows translation errors by type.

4.3 Some automatic measures

The emphasis in our study is in the linguistic evaluation of the translations, but we were also interested if we can quantify if the translations are similar to texts written by native speakers in terms of grammatical features, and also how many errors there are.

Table 6 shows how many spelling and grammar errors are detected in the target text. Grammatical errors include subject-verb agreement errors, compound errors.

The amount of non-words that the system has generated is quite notable, although several of these are reflected in non-translated corpus as well, for example confusion between á and a. It is more surprising that the neural MT has not generated many grammatical errors, at least ones that can be automatically detected.

Table 7 contains distributions of grammatical features in machine translated texts and large corpus.

There does not appear to be large difference between the machine translated and reference corpus, with the exception of lack of dual forms. This is not totally unsurprising, the forms are rare in use in general and do not have any comparable equivalent in source language: virtually all word-forms that concern two individuals fall under generic plurals in English, very few lexical selections can be used to refer two people specifically.

Type	error	correct
Nonsense words based on orthographic similarity	*Rabihtta-Hole*	*njoammilbiedju* ‘rabbit hole’
	”Vel!” for ”Well!”	*de*
Postpostition vs. preposition	*haga govaid*	*govaid haga* ‘without pictures’
Wrong PoS	*hui oađđin* ‘very sleep’ (noun)	*hui váiban* ‘very tired’ (adjective)
Lexical error	*álggii čuožžut su bálgáide* ‘started to stand his paths’	*álggii čuovvut su bálgáide* ‘started to follow his paths’
	su ***čivga*** *lei lohkame* ‘baby animal’	*su* ***oabbá*** *lei lohkame* ‘sister’
Literal/Non-idiomatic	*Aliceas ii lean boddu smiehttat* ‘Alice did not have a break to think’	*Alice ii ribahan smiehttat*
Polysemy error	*girjái* ***ahte*** *(subjunction ‘that’) su čivga lei lohkame*	*girjái maid (relative pronoun ‘that’) su čivga lei lohkame*
	mii lea girjji geavaheapmi ‘how can the book be used’	*mii lea girjji* ***ávki*** ‘what is the use of the book’
Periphrastic > synthetic construction	*ALICE lei šaddagoahtán váiban čohkkedit*	*ALICE lei váibagoahtán čohkkedeamis* ‘Alice started to be tired of sitting’
Valency error	*váiban čohkkedit* (infinitive)	*váiban čohkkedeamis* (locative) ‘tired of sitting’
Agreement error	*das eai lean govat* ***iige*** *ságastallamat* ‘there weren’t pictures and there wasn’t conversations either’	*das eai lean govat* ***eaige*** *ságastallamat* ‘there were neither pictures and there weren’t conversations either’

Table 5: Error types found in English-North Sámi neural MT

Text	Spelling (%)	Grammar (%)
Alice	232 (5%)	9 (0.1%)
BBC	23 (5%)	0
CTV	33 (4%)	1 (0.1%)
ILO-169	0	3 (0.1%)
SIKOR	399,282 (1.8%)	59,611 (0.3%)

Table 6: Automatically detected spelling (non-word) and grammar errors (real-word) in machine translated texts

5 Conclusion

We manually evaluated two neural machine translation systems in an indigenous low-resource context, one of which has North Sámi as a source language and the other of which has North Sámi as a target language. Translation is done either into or from a higher resource language, i.e. Norwegian and English, which are both morphologically simple compared to North Sámi. The Sámi to Norwegian evaluation is done by a native North Sámi speaker who has worked as a professional translator. We developed a scale according to which first the flow of the target language is evaluated and then the representation and exactness of the source language content in the target language. Both scales have 7 grades. Flow and content evaluation can differ very much from each other as flow mostly focuses on the target sentence, while content takes into account the source sentence to a much higher degree. The evaluation shows that flow typically scores higher than content, which means that a clear understanding of both source and target sentence is necessary to evaluate how well the matching is done. This supports our hypothesis that high-level language expertise is necessary to evaluate the quality of a translation.

For the English to Sámi evaluation we applied a different evaluation method. We applied high-quality rule-based proofing tools for Sámi for spellchecking and basic grammar checking of the target text. As human resources for indigenous languages are typically low, we find that this method - while it cannot replace human evaluation - can be revealing as regards certain shortcomings of the MT system, which affect its quality. We discovered that spelling errors in the neural translation are more than twice as much as in the Sámi text collection SIKOR. Additionally, a low-scale manual evaluation of the fictional text *Alice*, showed that shortcomings of the system included a variety of different morpho-syntactic errors as well of non-idiomatic constructions and nonsense translations.

The second system evaluation regards the newly released multi-lingual neural MT tool by Tartu university, where we had a look at English-North Sámi machine translation. None of the developers has knowledge of North Sámi and is therefore not able to properly evaluate the results in all its relevant details. We regard it as important that these systems are evaluated by those that have knowledge of the language, and give a reliable picture of what can and what cannot be expected of such a system. As a user can have varying knowledge themselves about either source or target language,

Text	**Poss**		**Dual**		**Actio**	
	n	%	n	%	n	%
Alice	34	0.8%	0	0	26	0.6%
BBC	1	0.2%	0	0	1	0.2%
CTV	4	0.5%	0	0	2	0.2%
ILO-169	23	0.7%	1	0.0%	3	0.1%
SIKOR	130,257	0.5%	59,623	0.2%	58,850	0.2%

Table 7: Distribution of grammatical features in machine translated documents (first four) and the large corpus (SIKOR).

expectations to the system can be different. We apply our rule-based proofing tools to test both spelling and grammar, provide an overview of prevailing error types of the MT tool, and show if the outcome reflects the morpho-syntactic reality of the monolingual Sámi corpus SIKOR written by native language users.

In the future we would like to manually evaluate neural MT both from and to an indigenous language (starting with North Sámi) on a larger scale in order to get more insights in refining the criteria of our evaluation method to come to adequate conclusions of the systems' quality. As this highly depends on human resources and language expertise, we also plan to focus on recruitment of language experts.

Acknowledgments

Máret Láila Anti contributed with her language expertise in North Sámi and Norwegian, as well as her professional knowledge regarding translation practices and Sámi linguistics.

This research was supported by Cloud TPUs from Google's TPU Research Cloud (TRC).

References

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2018. *Ethnologue: Languages of the World*, twenty-fifth edition. SIL International, Dallas, Texas.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Arle Lommel. 2018. *Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies*, pages 109–127. Springer International Publishing, Cham.

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers.

Marion Weller-di Marco and Alexander Fraser. 2022. Findings of the WMT 2022 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 801–805, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors. 2018. *Translation Quality Assessment*, volume 1 of *Machine Translation: Technologies and Applications*. Springer.

Flammie Pirinen, Sjur Moshagen, and Katri Hiovain-Asikainen. 2023. GiellaLT — a stable infrastructure for Nordic minority languages and beyond. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 643–649, Tórshavn, Faroe Islands. University of Tartu Library.

Maja Popović. 2018. *Error Classification and Analysis for Machine Translation Quality Assessment*, pages 129–158. Springer International Publishing, Cham.

Pekka Sammallahti. 1998. *The Saami languages – An Introduction*. Davvi Girji, Kárášjohka.

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.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Trond Trosterud and Kevin Brubeck Unhammer. 2012. Evaluating North Sámi to Norwegian assimilation RBMT. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 13–26, Gothenburg, Sweden.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

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.