A New Benchmark for Kalaallisut-Danish Neural Machine Translation

Ross Deans Kristensen-McLachlan
Center for Humanities Computing
Department for Linguistics,
Cognitive Science, and Semiotics
Aarhus University
rdkm@cc.au.dk

Johanne Sofie Krog Nedergård
Department of Nordic Studies
and Linguistics
University of Copenhagen
jskn@hum.ku.dk

Abstract

Kalaallisut, also known as (West) Greenlandic, poses a number of unique challenges to contemporary natural language processing (NLP). In particular, the language has historically lacked benchmarking datasets and robust evaluation of specific NLP tasks, such as neural machine translation (NMT). In this paper, we present a new benchmark dataset for Greenlandic to Danish NMT comprising over 1.2m words of Greenlandic and 2.1m words of parallel Danish translations. We provide initial metrics for models trained on this dataset and conclude by suggesting how these findings can be taken forward to other NLP tasks for the Greenlandic language.

1 Introduction

Greenlandic (Kalaallisut) is an Inuit-Yupik-Unangan language spoken by around 60,000 people in Greenland and Denmark. While Greenlandic is classified as 'Vulnerable' according to the Endangered Languages Project.[1], it is nevertheless relatively healthy. Local governance, media, and schooling up to university level are conducted either purely in Greenlandic, the sole official language since 2009, or in both Greenlandic and Danish, the colonial language (Compton, 2024).

Greenlandic has a number of comparatively rare linguistic features, not least of which is the prominent use of polysynthesis (Fortescue, 2007). Morphological complexity not only manifests in word-final inflections (mood, person, number, and so on) but also the theoretically indefinite productivity of adding morphemes to stems and concatenating other morphemes. For example, *palasi* means 'priest', *palasi-nngor-poq* means 'he becomes a priest', *palasi-nngor-tip-paa* means 'he causes him to become a priest', and *palasi-nngor-tit-si-neq* means 'the act of causing someone to become a priest'.

Additionally, there is widespread morphophonological assimilation or fusion at the morpheme boundaries (Fortescue, 1980). While these changes are to a large extent predictable, they can make it difficult - both for machine learning models and humans who are not native speakers of Greenlandic - to analyze precisely which morphemes comprise any given word.

1.1 From linguistics to NLP

Greenlandic hence poses a number of specific challenges to contemporary machine learning-based approaches to NLP. For example, most contemporary NLP systems use sub-word tokenization strategies such as Byte Pair Encoding (BPE, (Zouhar et al., 2023). Given its morphological complexities, sub-word tokenization seems unsuited to working with Greenlandic data, and an informal consensus among experts in the language has been that it is hence not amenable to contemporary NLP techniques. Does this mean that the language is excluded from the fruits of recent advances in contemporary neural language technology?

We contend that this is not the case. It is true that there is still lacking scientific investigation into foundational aspects of how easily modern NLP methods can be applied to a morphologically complex, low-resource, indigenous language like Greenlandic. However, a rapidly growing body of research already exists for languages as morphologically diverse as Nahuatl, Raramuri, Shipibo-Konibo, and Wixarika (Mager et al., 2022) and indigenous languages closely related to Greenlandic (Liu et al., 2020a; Schwartz et al., 2020). Recent work has provided systematic analysis of challenges and methods involved the creation of NMT systems for these kinds of languages (Mager et al., 2023).

This paper aims to move the Greenlandic NLP more in this direction by introducing a benchmark dataset for Greenlandic to Danish and provide the

[1] https://www.endangeredlanguages.com/

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 50–55
June 21, 2024 ©2024 Association for Computational Linguistics

first set of metrics on model performance. We specifically choose to translate to Danish, since this is a meaningful task given the complex social history between these two languages and cultures (Olsen, 2011; Kleeman-Andersen, 2021). In what follows, we outline the various steps taken to construct this dataset and the results of initial simple experiments.

1.2 Current state of Greenlandic NMT

Until recently, the only available tool for machine translation for Greenlandic to Danish was Nutserut, a rule-based approach to machine translation developed and maintained by Oqaasileriffik, the Language Secretariat of Greenland.[2] Research into Greenlandic NMT is thin on the ground. Earlier work exists on Greenlandic to English NMT but this work is hampered by the synthetic nature of the training data (Jones, 2022). There exists a growing body of work on related languages such as Inuktitut which investigates whether adding Greenlandic data to training pipelines might increase performance on English–Inuktitut NMT (to mixed results) (Roest et al., 2020). Beyond this, though, a survey of the existing literature suggests that there has to date been detailed empirical studies on the prospects and limitations of Greenlandic NMT.

Since we first started work on our project, a number of interested stakeholders have moved into this space, including the largest media house in Greenland[3] and Oqaasileriffik itself.[4] This is a positive development, since greater investment and engineering is likely to lead to growth and broader adoption of contemporary machine learning for Greenlandic. However, these tools are closed-source and do not provide transparent quantitative metrics for evaluating model performance. We are currently unaware of any work which has provided quantifiable metrics for Greenlandic to Danish NMT, meaning that the results presented here are the first such results on a benchmark dataset.

1.3 The problem of data

Thanks to the work done by custodians of Greenlandic, the language punches above its weight in terms of linguistic resources. For example, Oqaasileriffik have to date developed searchable text corpora; lexical resources such as dictionaries and terminology banks; and practical tools such as spellcheckers and text-to-speech models.[5] However, the language is still greatly under-resourced relative to other languages globally. This lack is most apparent in the context of well-designed Greenlandic-Danish parallel corpora. Currently there are no gold standard corpora in this area which can be used as a reliable benchmark for NMT.

This data scarcity has meant that rule-based approaches have dominated, since these approaches resolutely *do not* require large quantities of data. Nevertheless, rule-based approaches to language are now regularly replaced by or integrated alongside machine learning developed for high performance in low-data environments (Torregrosa et al., 2019; Huang et al., 2020). To ensure that Greenlandic is not left behind, it is necessary to explore all possibilities and to make the most of the available resources, even where they might not be ideally suited for the task.

For our experiments, we collected data with permission by scraping the public facing website of *Kalaallit Nunaata Radioa* (KNR), Greenland's national public broadcasting organization.[6] As a public broadcaster, KNR's data were freely accessible, and they have an official language policy necessitating that all texts published on their websites are published in both Greenlandic *and* Danish.

2 Methods

2.1 Data preperation

In May 2023, we scraped full articles for both Greenlandic and Danish versions of all articles stretching back to the first available digital texts. This created an initial corpus of roughly 72k Greenlandic language articles and around 63k Danish language articles. This is due to the fact that, as one goes further back in KNR's archives, it appears that earlier articles were not regularly translated into Danish, with the official dual translation policy only coming into effect in 2010.[7] Different translations of the same article are linked by a unique identifier, meaning that we were able to remove Greenlandic texts for which there was no translation. Scraped HTML files were consistently structured across the translations, meaning that it was simple task to automatically extract the main body text from each document creating a raw text

[2] https://nutserut.gl/pre

[3] https://www.sermitsiaq.ag/

[4] https://nutserut.gl/

[5] https://oqaasileriffik.gl/en/

[6] https://knr.gl/kl

[7] https://knr.gl/da/sprogpolitik-i-knr

corpus of parallel documents.

To create a parallel sentence corpus, we made use of a crude and efficient alignment algorithm. Documents were first split into sentences by tokenizing on common end-of-sentence punctuation such as periods, exclamation marks, and question marks. This resulted in each document being transformed into a list of sentences. We then compared the the overall number of sentences in each article between the Greenlandic and Danish version of the document. In the case of a mismatching number of sentences per document pair, we discarded this pair of articles from the corpus. If the number of sentences in an article matched, we assumed that there was a one-to-one mapping between sentences in the different translations of the text.

The final corpus hence comprises the sentence pairs from all of those articles which have the same number of sentences per article. While this approach is of course naïve, it was necessary given the lack of available resources to otherwise create a useable parallel corpus. We expand on this problem below in Limitations.

The final result of this process is a parallel corpus of around 73k sentence pairs, comprising around 1.2m words of Greenlandic and 2.1m words of Danish. This is comparable to previous studies working in similar linguistic contexts (Schwartz et al., 2020). Of this data, a randomly drawn sample of 1k sentences were held back as test data for evaluating model performance.

2.2 Model creation

All models were trained using *OpenNMT* with a *PyTorch* backend (Klein et al., 2017)[8]. BPE tokenizers were trained using *pyonmtok,* a wrapper for OpenNMT's tokenizer.[9]

Each experimental condition used the same Bi-LSTM encoder-decoder architecture adopting the default hyperparameters outlined in OpenNMT's documentation.[10] The only exceptions are the using of the Adam for optimization and an initial learning rate of 0.001. Each model ran for 100k training steps with model checkpoints saved after every 10k steps.

Alongside the custom RNN models outlined, we also tested the performance of state-of-the-art LLMs on this task. Using the OpenAI API, we performed zero-shot testing of *GPT-3.5-turbo* and *GPT-4* with the following prompt:

> *Translate this text from Greenlandic to Danish, without any additional comments or explanations: {text}*

3 Experiment

3.1 Hardware considerations

Local models were trained on a machine running Ubuntu (18.04.6 LTS) with 40 Intel(R) Xeon(R) Silver 4210 CPU cores and four Quadro RTX 8000 GPUs.

3.2 Evaluation metrics

Evaluating machine translation quantitatively is a notoriously fraught endeavour, with a number of different metrics proposed to quantify exactly how well any given model is performing (Popović, 2015; Chatzikoumi, 2020; Rei et al., 2020). Since no one metric robustly measures translation quality in a way which is entirely in line with expectations of human readers, we evaluate model performance using of a range of standard metrics.

Surface similarity is measured via n-gram overlap using BLEU (Papineni et al., 2002) and via character overlaps with chrF (Popović, 2015). Both sets of evaluations were performed using open-source and publicly available implementations of these algorithms.[11,12]

While both the BLEU and chrF metrics evaluate slightly different aspects of the generated translations, they are both ultimately based on the amount of string overlap between machine generated text and human references. However, it is of course true that any given sentence can be translated a seemingly indefinite number of ways while retaining the same meaning. In order to capture aspects of the *semantic similarity* machine generated translations and their references, we make use of BERTScore (Zhang et al., 2020).[13] This has been applied and shown to perform well in Danish contexts, such as evaluating abstractive text summarization (Kolding et al., 2023).

3.3 Results

The results for all models are shown in Table 1 below. We see that the Bi-LSTM model with the

[8] https://github.com/OpenNMT/OpenNMT-py

[9] https://github.com/OpenNMT/Tokenizer/

[10] https://opennmt.net/OpenNMT-py/options/train.html

[11] https://github.com/mjpost/sacrebleu

[12] https://github.com/google-research/google-research/tree/master/rouge

[13] https://github.com/Tiiiger/bert_score

Model	BERTscore (F1)	BLEU	chrF2
Bi-LSTM + 5k BPE	**0.74**	**16**	32.3
Bi-LSTM + 10k BPE	0.73	13	**32.5**
Bi-LSTM + 30k BPE	0.72	12	29.9
Bi-LSTM + 50k BPE	0.71	11	27.9
GPT-3.5-turbo	0.64	3	25.4
GPT-4	0.68	4	28.3

Table 1: Results across all model types

fewest number of joins performs best on this data, with the second highest scoring model being the model with the second lowest number of joins. In general, we see that increasing the number of BPE merges descreases performance in a proportional and linear way.

Perhaps contrary to expections, the LLM solutions perform notably worse than all of our much simpler, custom RNN models. However, this is likely due to the zero-shot nature of the task; additional experimentation is necessary to test the limits of LLMs for this particular tasks.

4 Discussion

4.1 What does this show?

The most striking takeaway from these preliminary results is that models generally perform reasonably well when evaluated using BERTScore, while scores tend to be much poorer for the n-gram and character-based metrics. Put simply, it would seem as though the translations produced by these models tend to be semantically close to the human generated sentences but are otherwise lexically or stylistically divergent from the human references.

Nevertheless, the numbers presented here are not widely different from research into similar indigenous linguistic contexts, such as reported BLEU scores for Yup'ik to English (≈13, (Liu et al., 2020a)) and Inuktitut to English (≈28, (Schwartz et al., 2020); see also (Nicolai et al., 2021)). Despite the widespread perception of the linguistic uniqueness of Greenlandic, it would seem that the language is nevertheless amenable to NMT.

Crucially, though, we also demonstrate that a smaller, simpler Bi-LSTM model currently outperforms more sophisticated LLM solutions. With a few-shot prompting regime and additional finetuning this could likely be improved, but it does provide a note of caution against immediately adopting "state-of-the-art" models without detailed testing and robust scientific evaluation.

4.2 Where next?

Our initial NMT experiments with Greenlandic to Danish are limited by our use comparatively simple architectures. Immediate next steps will be to experiment with more sophisticated neural network architectures such as transformer-based models, as well as the applicability of pre-trained multilingual embeddings such as *mT5* (Xue et al., 2021) or *mBART* (Liu et al., 2020b).

This opens up a wide range of possibilities, including practical technologies such as speech-to-text models and improved research methods for linguistic analysis and language modelling. This has the potential to contribute substantially to the scientific study of Greenlandic from the perspectives of cognitive science and language psychology, such as considering the relationship between sub-word tokenization and human morphological segmentation.

Finally, we aim for this to be a stepping stone towards collaboration with researchers currently engaged with similar work on other Inuit-Yupik-Unangan languages. Given the similarities between these languages, we believe that pooling resources could lead to substantial progress in language technology for languages in this region of the world.

5 Conclusions

This paper is a preliminary step towards training neural language technology for Greenlandic and, crucially, empirically testing both the possibilities and limitations of this approach. We present a benchmark dataset for Greenlandic-Danish NMT as well as providing initial metrics from simple models trained on this dataset. These initial experiments are not intended to provide complete, industry-strength machine translation for Greenlandic to Danish. Improvements in the area of

Greenlandic NMT and NLP more generally requires greater emphasis on the curation and stewardship of high quality training data. We believe that this focus would contribute greatly to Greenland's already rich linguistic cultural heritage.

Limitations

While we view these results positively, these trained models are far from production-ready or of practical use. The process for creating the parallel corpus is crude and involves a number of pragmatic decisions by the authors, neither of whom are native Greenlandic speakers. The collection algorithm outlined in Section 2.1 was designed as a "good enough" solution for initial experiments. However, greater quality control with more human intervention is required for future work to ensure that the corpus is in fact aligned.

KNR texts have some well-known limitations (Duus, 2012a,b; Hussain, 2018; Kleeman-Andersen, 2020). Several of the texts are originally written in Danish (largely by monolingual Danish-speaking or bilingual journalists) and subsequently translated to Greenlandic. Texts hence tend to be quite "literal" or non-idiomatic translations and thus appear somewhat unnatural to a Greenlandic speaker. The quality of the Greenlandic texts at KNR generally is a point of heated public debate in Greenland with many complaining about grammatical errors, repetitive expressions, and too much influence from Danish.

References

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Richard Compton. 2024. *38 Inuit-Yupik-Unangan: An overview of the language family*, pages 843–874. De Gruyter Mouton, Berlin, Boston.

Søren Duran Duus. 2012a. Journalister: Gør noget ved sproget i vores medier. *Sermitsiaq*.

Søren Duran Duus. 2012b. Knr efter sprogkritik: Vi kan blive bedre. *Sermitsiaq*.

Michael D. Fortescue. 1980. Affix ordering in west greenlandic derivational processes. *International Journal of American Linguistics*, 46(4):259–278.

Michael D. Fortescue. 2007. The typological position and theoretical status of polysynthesis. 5:1–27.

Jin-Xia Huang, Kyung-Soon Lee, and Young-Kil Kim. 2020. Hybrid translation with classification: Revisiting rule-based and neural machine translation. *Electronics*, 9(2).

Naimah Hussain. 2018. *Journalistik i små samfund: Et studie af journalistisk praksis på grønlandske nyhedsmedier*. Ph.D. thesis, Roskilde University.

Alex Jones. 2022. Finetuning a kalaallisut-english machine translation system using web-crawled data. *CoRR*, abs/2206.02230.

Camilla Kleeman-Andersen. 2020. Plastikblomster og tungeløse grønlændere - følelser i sprogdebatten 2009-2019.

Camilla Kleeman-Andersen. 2021. Den evigt nærværende koloniale fortid. In *Sprogs status i Rigsfællesskabet 2031*, pages 38–40. Københavns Universitets Humanistiske Fakultet.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Sara Kolding, Katrine Nymann, Ida Hansen, Kenneth Enevoldsen, and Ross Kristensen-McLachlan. 2023. DanSumT5: Automatic abstractive summarization for Danish. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 248–264, Tórshavn, Faroe Islands. University of Tartu Library.

Christopher Liu, Laura Dominé, Kevin Chavez, and Richard Socher. 2020a. Central Yup'ik and machine translation of low-resource polysynthetic languages. *CoRR*, abs/2009.04087.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Manuel Mager, Rajat Bhatnagar, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2023. Neural machine translation for the indigenous languages of the Americas: An introduction. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 109–133, Toronto, Canada. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Elisabeth Mager, Katharina Kann, and Thang Vu. 2022. BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 961–971, Dublin, Ireland. Association for Computational Linguistics.

Garrett Nicolai, Edith Coates, Ming Zhang, and Miikka Silfverberg. 2021. Expanding the JHU Bible corpus for machine translation of the indigenous languages of North America. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 1–5, Online. Association for Computational Linguistics.

Carl Chr. Olsen. 2011. Sproglovgivning under grønlands selvstyre. *Sprog i Norden*, 42(1):25–30.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Christian Roest, Lukas Edman, Gosse Minnema, Kevin Kelly, Jennifer Spenader, and Antonio Toral. 2020. Machine translation for English–Inuktitut with segmentation, data acquisition and pre-training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 274–281, Online. Association for Computational Linguistics.

Lane Schwartz, Francis M. Tyers, Lori S. Levin, Christo Kirov, Patrick Littell, Chi-kiu Lo, Emily Prud'hommeaux, Hyunji Hayley Park, Kenneth Steimel, Rebecca Knowles, Jeffrey Micher, Lonny Strunk, Han Liu, Coleman Haley, Katherine J. Zhang, Robbie Jimerson, Vasilisa Andriyanets, Aldrian Obaja Muis, Naoki Otani, Jong Hyuk Park, and Zhisong Zhang. 2020. Neural polysynthetic language modelling. *CoRR*, abs/2005.05477.

Daniel Torregrosa, Nivranshu Pasricha, Maraim Masoud, Bharathi Raja Chakravarthi, Juan Alonso, Noe Casas, and Mihael Arcan. 2019. Leveraging rule-based machine translation knowledge for under-resourced neural machine translation models. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 125–133, Dublin, Ireland. European Association for Machine Translation.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Tim Vieira, Mrinmaya Sachan, and Ryan Cotterell. 2023. A formal perspective on byte-pair encoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 598–614, Toronto, Canada. Association for Computational Linguistics.