Tokenization on Trial: The Case of Kalaallisut–Danish Legal Machine Translation

Esther Ploeger[1] **Paola Saucedo**[1] **Johannes Bjerva**[1]
Ross Deans Kristensen-McLachlan[2] **Heather Lent**[1]
[1]Department of Computer Science, Aalborg University
[2]Department for Linguistics, Cognitive Science, and Semiotics, Aarhus University
{espl,hcle}@cs.aau.dk

Abstract

The strengths of subword tokenization have been widely demonstrated when applied to higher-resourced, morphologically simple languages. However, it is not self-evident that these results transfer to lower-resourced, morphologically complex languages. In this work, we investigate the influence of different subword segmentation techniques on machine translation between Danish and Kalaallisut, the official language of Greenland. We present the first semi-manually aligned parallel corpus for this language pair[1], and use it to compare subwords from unsupervised tokenizers and morphological segmenters. We find that Unigram-based segmentation both preserves morphological boundaries and handles out-of-vocabulary words adequately, but that this does not directly correspond to superior translation quality. We hope that our findings lay further groundwork for future efforts in neural machine translation for Kalaallisut.

1 Introduction

In contrast to many of the world's indigenous languages facing challenges in revitalization as a result of colonialism (Meakins and O'Shannessy, 2016), Kalaallisut (West Greenlandic) has a vibrant linguistic ecosystem. Spoken as a first language by people of all ages (Grenoble and Whaley, 2021), Kalaallisut is used in all aspects of daily life by most of the population (Nielsen, 2021), from teenagers texting (Grenoble, 2011) to everyday communication (Ravn-Højgaard et al., 2018). It is also supported by language policies that prioritize its use in education and administration (Møller, 1988; Valijärvi and Kahn, 2020), and boasts a wide range of linguistic resources (*e.g.*, word lists and dictionaries) and existing language technologies (*e.g.*, a spell-checker and a grapheme-to-phoneme converter) from Oqaasileriffik, the Language Secretariat of Greenland.

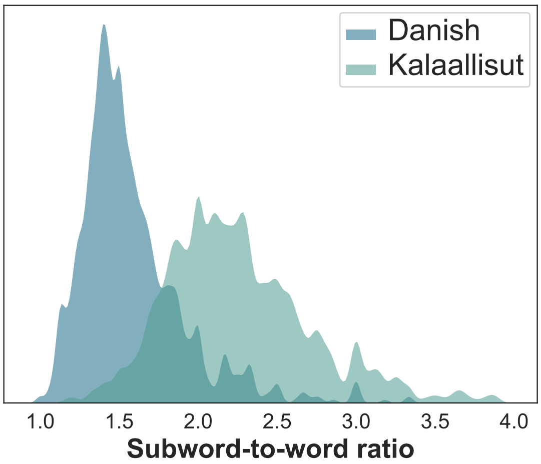

Figure 1: When applying BPE to our test dataset, Kalaallisut generally has higher subword-to-word ratios than Danish; *KDE plot, capped at 4.*

Despite the vitality of the language, however, Kalaallisut – like most of the world's languages – does not have sufficient resources for the data-intensive methods of contemporary NLP (Joshi et al., 2020). Specifically in the context of neural machine translation (NMT), Kalaallisut lacks the large-scale aligned parallel corpora required for contemporary machine learning methodologies, and is thus considered a low-resource language. Consequently, Kalaallisut trails behind higher-resourced languages in terms of NMT.

Beyond the limited availability of high-quality parallel corpora, Kalaallisut's high degree of morphological inflection poses additional challenges for NMT. Commonplace tokenization methods often lead to large, sparse vocabularies for morpho-

[1] https://github.com/esther2000/tokenization-on-trial

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 480–491
March 3-4, 2025 ©2025 University of Tartu Library

logically rich languages (Vylomova et al., 2017; Gerz et al., 2018; Akın Özçift and Söylemez, 2021). To illustrate the difference with a morphologically simple language: Figure 1 shows that BPE tokenization yields many more subwords per word for Kalaallisut than for Danish. It is likely more difficult for models to learn systematic patterns in structure between source and target languages for morphologically rich languages than for morphologically-poor ones (Gutierrez-Vasques et al., 2023). While previous work has compared subword tokenization and segmentation strategies for polysynthetic languages of the Americas (Mager et al., 2022), this work found no single best solution across languages. Thus, NMT for Kalaallisut stands to benefit from a dedicated investigation.

Despite the desire for machine translation by speakers of Kalaallisut (Oqaasileriffik, 2023), however, there is a marked scarcity in research attention (Kristensen-Mclachlan and Nedergård, 2024). As a result, adequate benchmarking datasets for Kalaallisut NMT are extremely scarce.[2] This work aims to provide practical insights for improved NMT for Kalaallisut, by comparing the efficacy of different segmentation strategies. To this end, we provide the following contributions:

- We present the first semi-automatically aligned Danish–Kalaallisut parallel dataset, in the legal domain;
- We present the first open-science initiative to benchmark NMT from Danish *into* Kalaallisut;
- We compare subwords from four segmentation models and relate the insights to downstream NMT performance;
- We provide discussion and recommendations for future research on Kalaallisut NMT.

Ultimately, we hope that this work can be helpful for the development of open-science NMT systems for Kalaallisut going forward.

2 Background

Greenlandic Language Kalaallisut is the largest member of the Inuit-Yupik-Unangan family. Among the world's languages, it is one of the more morphologically rich, described as *"typologically extreme"* in the number and variety of suffixing morphemes available for marking nominal and verbal stems (Fortescue and Olsen, 2022). While effectively segmenting languages with greater morphological complexity is notoriously difficult in NLP (Klavans, 2018b), additional linguistic characteristics of Greenlandic may further complicate subword tokenization. Specifically, De Mol et al. (2020) point to three salient features of Greenlandic morphology: 1) some morphemes are polysemous (*e.g.*, no distinction between present and past tense); 2) unbound morphemes can sometimes be incorporated, resulting in ambiguous morpheme boundaries; and 3) some morphemes undergo phonological changes depending on the subsequent context, in order to avoid illegal morphophonemic sequences. The combination of these factors underscore the utility of a targeted investigation into optimal tokenization strategies for Kalaallisut NMT.

Polysynthesis in NLP In linguistic typology, polysynthesis is a high-level categorization for languages relying heavily on morphological inflection to convey meaning.[3] As a result, individual utterances in polysynthetic languages tend to be relatively longer than their non-polysynthetic counterparts. In other words, where non-polysynthetic languages might add a pronoun or preposition, polysynthetic languages incorporate additional morphemes. This results in a kind of *holophrasis*, with a single word encoding both predicate and arguments of a clause within the verb itself (Mithun, 2017).

While polysynthetic languages can be found across the globe (*e.g.*, Quechua in South America and Ainu in Asia), many of them are endangered (Klavans, 2018a), and thus lack representation in NLP (Joshi et al., 2020). Indeed, the fact that most polysynthetic languages are low-resource has meant that the development of language technology for these languages continues to lag behind (Klavans, 2018b). At the same time, polysynthesis brings with it unique challenges for NLP (Eskander et al., 2019). For example, in the context of NMT, Mager et al. (2018) observe marked information loss between polysynthetic and fusional languages, as a consequence of alignment. Specifically, the NMT systems omit the parts of

[2] The OPUS collection (Tiedemann, 2009) contains 291 parallel Danish ↔ Kalaallisut samples, most of which consist of a single word.

[3] It should be noted that, although widely-used across typology, polysynthesis as a proper typological categorization is contested by some linguists (Zúñiga, 2019).

the polysynthetic languages, where morpheme-to-morpheme alignment yielded no equivalent counterpart in the fusional language.

As MT has moved towards neural approaches, subword segmentation has become standard for leveraging large datasets. However, subword tokenizers like BPE have been met with skepticism for polysynthetic languages, as they do not accurately capture morpheme boundaries (Vylomova et al., 2017; Gerz et al., 2018; Kann et al., 2018; Akın Özçift and Söylemez, 2021; Saleva and Lignos, 2021). In their study on tokenization for polysynthetic languages, Mager et al. (2022) compare BPE versus morphological segmentation across four polysynthetic languages (*i.e.*, Nahuatl, Raramuri, Shipibo-Konibo, and Wixarika). For three languages, morphological segmenters outperform BPE, except for Nahuatl, where BPE yields better results.

MT for Kalaallisut The rich media ecosystem surrounding Kalaallisut means that there exists a reasonable volume of data for the language, compared to many other indigenous languages. Nevertheless, much of this data is not immediately suitable for tasks such as training NMT systems. Taken along with the perceived challenges of working with the language outlined above, this means that MT systems for Kalaallisut have historically relied on rule-based (Oqaasileriffik, 2017) and hybrid approaches (Oqaasileriffik, 2023). Early works for NMT of Kalaallisut were developed in relation to Inuktitut, in attempts to benefit from cross-lingual transfer. Le and Sadat (2020) demonstrate that the use of (bi)character-based and word-based pre-trained embeddings can improve NMT performance for Inuktitut (an indigenous language of eastern Canada), suggesting similar possibilities for other Inuit languages. Nonetheless, the addition of Kalaallisut shows limited usefulness in transfer learning for Inuktitut-English MT thus far (Roest et al., 2020).

More recently, Kristensen-Mclachlan and Nedergård (2024) introduced the first benchmark for Kalaallisut-Danish NMT, containing over 1.2 million words of Kalaallisut and 2.1 million words of parallel Danish translations. However, the authors note limitations related to "crude" sentence level alignment, noting that future data collection efforts are still necessary. In experiments, they use a BiLSTM encoder-decoder architecture with BPE tokenization (5k, 10k, 30k, and 50k vocabulary size), finding best results with 5k BPE. While the authors discuss potential concerns about subword tokenization for the morphologically rich Kalaallisut, their results demonstrate that BPE is reasonably amenable to the language. Still, they do not experiment with other tokenization strategies.

3 A Reliably Parallel Dataset

Aligning parallel datasets is non-trivial in the case of highly inflectional, low-resource languages. Popular alignment methods require pre-trained language embeddings (Thompson and Koehn, 2019), pre-suppose tokenized text (Varga et al., 2007), or assume that sequence lengths correspond directly across languages (Gale and Church, 1993). Kelly (2020) conducted extensive experiments on alignment of polysynthetic languages, but found that their result for Danish-Kalaallisut was too noisy and thus not useful downstream. Their data was sourced from magazines, however. We hypothesize that choosing a more structured domain (*e.g.*, legal) may make alignment more feasible.

Data Collection Oqaasileriffik referred us to the collection of parallel legal texts, hosted by the Greenlandic Government.[4] Although Kalaallisut is the most widely spoken language (United Nations, 2023), Greenland's legal system is bilingual, and laws and legal documents are often drafted in Danish. In this work, we use the *Law Collections*, which is an archive of the legislation of Greenland's Self-Government, Danish legislation applicable to Greenland, and international regulations that are relevant to Greenland.[5] In total, it consists of 2,545 publicly available documents in HTML format, originally written between 1908 and 2024, many of which are manually translated.

Alignment and Filtering Through scraping, we retrieve parallel documents, filtering out any non-translated documents. However, to obtain parallel *sentences*, we need to align the text. As mentioned, this assumes data or experimental consensus which is not available for Kalaallisut (*i.e.*, language embeddings and tokenized text). Fortunately, legal text is highly structured: our scraped data contains strict paragraph markers (*e.g.*, *§ 2*)

[4] Available at https://nalunaarutit.gl

[5] https://nalunaarutit.gl/om-nalunaarutit

and clause enumerations (*e.g.*, *a)*), equally across source and target. We leverage this structure by aligning through enumeration: for each document, we retrieve all enumerated text segments and align accordingly in case of 1:1 correspondence with enumeration markers. As an additional advantage, alignment on enumerated clauses furthermore serves as a filtering step. For example, less-structured introductory texts and sensitive information such as email addresses and full names are automatically filtered out. We strip the enumeration token from each line, apply deduplication and subsequently extract 1,000 lines for the validation set, and 1,000 other lines for the test set. We use the remaining lines as the training set. Importantly, we make the design choice to not remove near-duplicates. Legal texts can be highly formulaic, and since we perform an in-domain evaluation which cannot be expected to be widely generalizable regardless (see: Limitations), we decide to leave them in.

Dataset Size In Table 1, we show the size of the resulting corpus. The dataset consists of more than 40,000 parallel phrases. Unsurprisingly, due to Kalaallisut's inflections, the number of separate words (whitespace delimited strings of characters, obtained with the `wc -l` command) is much higher for Danish than for Kalaallisut.

Split	**# Lines** *GL*	*DA*	**# Words** *GL*	*DA*
Training	39,936	39,936	663,734	929,904
Validation	1,000	1,000	16,594	23,021
Testing	1,000	1,000	16,665	23,846
Total	**41,936**	**41,936**	**696,993**	**976,771**

Table 1: Size of parallel legal text dataset.

While small compared to what is available for high-resource languages, the size of the dataset is larger than that used in a comparable low-resource neural MT study (Mager et al., 2022). It is smaller than the other open, parallel Danish-Kalaallisut dataset (Kristensen-Mclachlan and Nedergård, 2024), but as ours is aligned based on human alignments, we expect that ours includes considerably less noise. This leaves us with a small, but high-quality in-domain dataset for legal translation.

4 Experiments

Since we are interested in isolating the effects of subword segmentation on NMT performance, we train dedicated bilingual MT models from scratch. Our experimental set-up consists of three steps: subword segmentation, machine translation, and evaluation, each described in more detail below.

4.1 Subword Segmentation

We experiment with two types of unsupervised segmentation for Kalaallisut: traditional MT subword tokenizers, and morphological segmenters. Following Mager et al. (2022), we keep the Danish side of the parallel corpus consistent across experiments, as this allows us to isolate the effects of Kalaallisut segmentation. We apply BPE to the Danish text, trained on the Danish training set of our corpus. We use a vocabulary size of 5k, as this was found to be optimal in the three most similar research initiatives (Saleva and Lignos, 2021; Mager et al., 2022; Kristensen-Mclachlan and Nedergård, 2024).[6]

Traditional MT Tokenization Following Mager et al. (2022), we train and apply Byte-Pair Encoding (BPE; Sennrich et al., 2016). Originally introduced as a data compression algorithm (Gage, 1994), the segmenter is trained bottom-up by merging frequently co-occurring vocabulary items. In addition, we experiment with Unigram language modeling (Kudo, 2018). Rather than constructing the vocabulary bottom-up, it starts from the largest vocabulary, which is subsequently pruned. This method has been shown to preserve morphological segmentation better than BPE (Bostrom and Durrett, 2020), making it especially relevant for our study. We use both algorithms as implemented in SentencePiece (Kudo and Richardson, 2018).

Morphological Segmentation Segmenting text according to (predicted) morpheme boundaries may be particularly beneficial for low-resource MT, as a means to counter the data scarcity of co-occurring characters that inflections may introduce. As we do not have a large-scale in-domain annotated dataset of morphological segmentations for Kalaallisut, we are constrained to unsupervised segmenters. Specifically, we follow Saleva and Lignos (2021) in using Morfessor 2.0 (Smit et al.,

[6] For Kalaallisut, we also experimented with vocabulary sizes 1k, 3k, 7k, 9k and 11k, but found no improvement.

	Kalaallisut Segmentation		***Machine Translation***			
			Danish→Kalaallisut		Kalaallisut→Danish	
Method	**Fertility**	**% Cont.**	**chrF2**	**BLEU**	**chrF2**	**BLEU**
None	*1.000*	*0.00*	*44.6*	*3.4*	*61.5*	*15.4*
BPE	2.294	**66.10**	56.4	8.7	**64.2**	**21.5**
Unigram	2.290	66.27	61.4	**10.1**	58.9	17.1
Morfessor	1.925	70.72	56.7	7.9	58.3	17.2
FlatCat	**1.870**	69.30	**63.2**	9.6	57.0	15.1

Table 2: Comparison of segmentation and translation quality metrics on the Kalaallisut test set.

2014), henceforth simply *Morfessor*. In addition, we use FlatCat (Grönroos et al., 2014), which is an extension over Morfessor that uses a Hidden Markov model. After applying morphological segmentation, we post-process the data such that the SentencePiece output format is replicated (words separated by the underscore symbol, and subwords separated by spaces).

Each of the segmentation methods is trained on the training set and applied to all sets (training, validation, and test set) of the Kalaallisut part of the parallel corpus data only. As a baseline, we add the case of applying no segmentation whatsoever to the Kalaallisut side.

4.2 Machine Translation

We train bilingual NMT models for both translation directions separately, with the Transformer architecture (Vaswani et al., 2017). We use the Fairseq toolkit (Ott et al., 2019). Because of the limited data availability in our scenario, we tailor the hyperparameters to those typically found to be effective in low-resource translation, such as using a higher dropout rate (Sennrich and Zhang, 2019; Araabi et al., 2022). We use a learning rate of 0.0001, cross entropy as a criterion with label smoothing (0.2), and apply a dropout rate of 0.3. Each model is trained for a maximum of 100 epochs, with a patience setting of 5 epochs monitoring the validation loss. For generation we use the best checkpoint.

4.3 Evaluation

Subword Metrics To compare segmentation methods, we use two metrics proposed by Rust et al. (2021): *subword fertility* and *continued word proportion*. Subword fertility is the average number of subwords per word. This metric provides insight into "*how aggressively a tokenizer splits*". The proportion of continued words measures the percentage of words that are divided into more than one subword, indicating how *often* words are split. For "words", we use the whitespace delimited character strings. Intuitively, lower scores are preferred, as high values signal weak compression efficacy, which could lead to oversegmentation.

Translation Quality For assessing the quality of the output translations, we report the chrF2[7] (Popović, 2015) and BLEU[8] (Papineni et al., 2002) scores, as implemented in SacreBLEU (Post, 2018). The ChrF2 metric is especially suitable to our scenario, as it is based on character n-grams. It has been previously been used in the context of low-resource NMT on diverse languages (e.g. Tiedemann, 2020). Due to the low-resourcedness, we do not include evaluation based on language embeddings, such as COMET (Rei et al., 2020), as there are indications that they are not reliable in low-resource scenarios (Falcão et al., 2024). While human evaluation would likely provide a better insight into the usefulness for speakers, the absolute number of native translation professionals is much lower than for many, higher-resourced, language pairs. At the same time, this highlights the need for research into reliable MT systems for Kalaallisut.

4.4 Main Results

Table 2 lists the subwords metrics and downstream MT performance for each of the segmentation methods. It should be noted that the results for Danish→Kalaallisut and Kalaallisut→Danish cannot be compared directly, because of the un-

[7] Signature: `nrefs:1|case:mixed|eff:yes|nc:2|nw:0|space:no|version:2.4.3`

[8] Signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.3`

even number of character and word n-grams. Instead, systems should be compared column-wise.

First, we observe that not using any segmentation method leads to suboptimal downstream MT results. Especially in the case of Danish→Kalaallisut, performance trails considerably behind that obtained with segmenters. BPE obtains the highest scores for translation into Danish, but this is not the case for translation into Kalaallisut, where both the Unigram and FlatCat approaches obtain higher chrF2 scores.

We do not observe clear patterns as to the subword metrics and MT performance. While the morphological segmenters, Morfessor and FlatCat, obtain the lowest fertility scores, this does seem to correspond directly to higher MT quality. While this corroborates earlier findings (Saleva and Lignos, 2021; Mager et al., 2018), more data points are needed to draw robust conclusions.

5 Analysis

To add more context to our findings, we perform additional analyses.

5.1 Subwords vs. Morphological Boundaries

To what extent do the subword segmenters preserve morphological boundaries? To analyze this, we apply each segmenter to a list of words, for which we have gold-standard annotations. We use the data from De Mol et al. (2020), who compiled a set of Kalaallisut words and phrases, and their morphological segmentations. These annotations originate from courses on Kalaallisut, and were corrected by a native speaker. Their data contains both short (e.g. *"he drinks"*) and long (e.g. *"it can be expected to have been eating jellyfish"*) general-domain examples. Since this is out-of-domain for the trained segmenters, it requires a degree of generalization. In total, we use 499 of these examples for our evaluation. We apply the segmenters to each of these examples, and evaluate the resulting subwords using precision (Eq. 1) and recall Eq. 2).[9] The F1-score is then calculated as the harmonic mean between the average precision and recall.

$$P = \frac{|\{gold\ morphemes\} \cap \{subwords\}|}{|\{subwords\}|} \quad (1)$$

$$R = \frac{|\{gold\ morphemes\} \cap \{subwords\}|}{|\{gold\ morphemes\}|} \quad (2)$$

Table 3 contains our results. For all segmenters, we find that morphological boundaries are only preserved modestly, with F1 scores all under 35 percent. The lowest score is found with BPE, with precision, recall and F1 only slightly above 10%. Relating this to the downstream results in Table 2, where best results for translation to Danish were obtained with BPE, it seems that preserving morphemes does not directly lead to optimal downstream NMT performance. This is in line with previous findings (Saleva and Lignos, 2021).

A second observation is that Unigram is (at least) on par with FlatCat and Morfessor when it comes to preserving morphological boundaries. This may be somewhat surprising, as Unigram is not a dedicated morphological segmenter. Yet, given its top-down pruning approach, morphemes are better preserved than with BPE's bottom-up approach. This is in line with findings from Bostrom and Durrett (2020).

Method	**Prec. (%)**	**Rec. (%)**	**F1 (%)**
BPE	10.81	12.42	11.56
Unigram	30.88	**37.68**	**33.94**
Morfessor	**31.08**	31.61	31.34
FlatCat	29.58	29.40	29.49

Table 3: Comparison of morphological boundaries and subword segmentation.

5.2 Out-of-Vocabulary Words

One of the core motivations for subword segmentation, is that it enables better representations of out-of-vocabulary (OOV) words. This has been argued to improve downstream performance, for instance in the case of MT (Sennrich et al., 2016). We explore how prominent OOV words are, when processed with varying segmentation techniques. We report the percentage of unknown items (UNKs) in the test portion of our parallel corpus, as shown in the logs of `fairseq-preprocess`. The results are listed in Table 4.

First, we observe that applying subword segmentation drastically reduces the number of UNKs. When not applying any segmentation, more than 14% of the words are OOV. This high number reflects Kalaallisut's highly inflectional characteristics. Moreover, this observation may provide an explanation for why down-

[9] Equations adapted from Nouri and Yangarber (2016).

Method	% UNK
None	14.30000
BPE	0.005080
Unigram	0.005050
Morfessor	0.196000
FlatCat	0.295000

Table 4: Proportion of OOV words in the Kalaallisut test set.

stream MT performance, specifically *into* Kalaallisut, lags when not applying any segmentation (Table 2). Secondly, we observe a difference between the morphological segmenters (Morfessor, FlatCat) and the traditional MT tokenizers (BPE, Unigram): using the latter results in fewer UNKs than the former. Notably, it is interesting that Unigram segmentation somewhat preserves morpheme boundaries (Table 3), while also resulting in relatively few UNKs.

6 Discussion and Recommendations

Given the lack of research for Kalaallisut NMT, we posit that collaboration between NLP researchers, Greenlandic language experts, and non-specialist native speakers of the language is crucial. In addition to a parallel dataset and experimental documentation, we aim to contribute to NMT for Kalaallisut by providing some high-level recommendations below.

Explore Additional Resources Beyond the legal domain, Kalaallisut boasts a wealth of traditional linguistic resources, like dictionaries (Berthelsen, 1997) and formal grammars (Fortescue, 1984; Sadock, 2003; Berge, 2011; Kahn and Valijärvi, 2021; Nielsen, 2022). Due to Greenland's relationship with Denmark, national newspapers and official government resources are often available in both Kalaallisut and Danish, which allows "pseudoparallel" corpora to be compiled through webcrawling (Jones, 2022). Similar efforts could be applied to other domains. Additional digital resources for Kalaallisut include a spell-checker, text-to-speech system, and grapheme-to-phoneme converter (Oqaasileriffik), a hand tagged corpus (Per Langgård and VISL Team), and recent NMT benchmark dataset (Kristensen-Mclachlan and Nedergård, 2024). With the exception of the latter, no previous works make use of this wealth of resources, and thus practitioners may benefit from their inclusion going forward.

Consider Other Dialects Even among low-resource languages, the majority of research attention is paid to standard language varieties, with the risk that non-standard dialects are left behind (Faisal et al., 2024). This holds true for Greenlandic, where works on non-standard dialects are far outnumbered by those for Kalaallisut. The Greenlandic language contains three main dialects: Kalaallisut (the western dialect, and the standard form), Tunumiisut (spoken in eastern Greenland), and Inuktun (used in the northern region).[10] While Kalaallisut is predominant, all dialects are vital to understanding Greenland's linguistic diversity. Only a few grammar books are available for Tunumiisut (Robbe and Dorais, 1986; Mennecier, 1995; Tersis, 2008) and Inuktun (Fortescue, 1986), however. No NLP datasets have as yet been published, despite their apparent presence on social media. This suggests that the language's integration into advanced language technologies is still limited (Siminyu et al., 2020), and future works for Greenlandic NLP could thus benefit from curation of resources and experimentation across dialects.

Mind the Historical Context The colonization of Greenland involved Denmark's efforts to "civilize" the Inuit population, primarily through educational programs aimed at reshaping their culture (Rud, 2009) and "modernization" efforts in the 1950s also prioritized the Danish language (Gad, 2017). These initiatives reflected broader colonial views that objectified Greenlanders based on race, gender, and class (Thisted, 2021). Even after World War II when decolonization began, they were often framed within medical and social research as "controllable subjects" (Rud, 2021). In spite of these pressures, the Greenlandic language remains widely spoken and serves as a symbol of national identity. In 2009, Greenlandic was declared the sole official language, but Danish remains prevalent in the public administration and essential for higher education (Faingold, 2023), and language policy is a recurring debate in Greenlandic politics (Gad, 2017). Despite this progress for the Greenlandic language, the legacy of colonialism still has consequences for indigenous lan-

[10] https://en.wikipedia.org/wiki/Greenlandic_language.

guages in NLP, which researchers must face. For in-depth conversations on this topic, we refer readers to Bird (2020) and Mager et al. (2023).

Avoid Extractivism Indigenous people have often been treated as research subjects rather than active participants in decision-making processes, particularly under colonial rule (Guillemin et al., 2016). While Greenland has made strides towards self-government (Kuokkanen, 2017), colonial legacies persist in imaginaries[11] of an "empty" Arctic whose resources can be readily exploited and its people trivialized (Hanrahan, 2017). In terms of research, this dynamic is enacted through the extraction of knowledge from marginalized communities for academic or bureaucratic consumption (Gaudry, 2011).

This historical context of exploitation raises ethical concerns about modern data collection practices in NLP. As in previous extractive practices, the potential misuse of data, along with issues surrounding privacy, consent, and bias, mirrors ongoing debates in the field of NLP regarding the ethical implications of data mining (Žliobaitė, 2017; Hassani et al., 2020; Watson and Payne, 2020; Singh, 2020; Rogers et al., 2021; Liu et al., 2023). For an in-depth conversation on extractivism in NLP, we refer readers to Bird (2024).

7 Conclusion

In this paper, we build upon the current state of Danish↔Kalaallisut NMT research, noting a sparsity of benchmarks and open-science experimental groundwork. We then introduce a new semi-manually aligned corpus of parallel legal texts for this language pair. Leveraging this, we conduct systematic experiments on subword segmentation, analyzing the impact of both traditional subword tokenizers (BPE, Unigram), and morphological segmentation (Morfessor 2.0, FlatCat) on downstream NMT performance. While segmentation techniques generally improve translation, we do not find one segmenter that beats the others in all aspects. Ideally, more data and evidence are needed to draw more robust conclusions.

Limitations

In this study, we do not examine any (massively) multilingual MT models. As a result, it is possible our work misses out on some of the benefits of transfer learning. However, the goal of this work not to create a new state-of-the-art, but rather investigate the isolated effects of subword solutions, relating to Kalaallisut. Accordingly, the findings in this paper can still serve as a starting point for those who continue this work in the future. Moreover, our work investigates isolated subword segmentation techniques, while segmentation methods are not necessarily mutually exclusive. For example, future work could look into applying BPE after morphological segmentation.

Another limitation of this work is its highly-specific legal domain. On the one hand, leveraging legal texts allows us to avoid extractivism, as these data are not taken from Greenlandic writings with deep cultural significance. On the other hand, the use of legal data can also be criticized as reinforcing colonial systems of authority. To avoid the latter, our work is exploratory in nature, and does not seek to create deployable, culturally-appropriate NMT systems for Greenlandic speakers. Instead, we aim to provide a methodology and results pertaining to segmentation, which can still transferable to works in NMT for Greenlandic, outside of the legal domain.

This work focuses solely on the dominant Kalaallisut dialect of Greenlandic. While the inclusion of more dialects is the subject of increasing awareness in NLP, text for other Greenlandic dialects is not supported by the platform through which we sourced our methodology.

Finally, future work in Kalaallisut MT would hugely benefit from human quality assessment. While we assume that automatic, reference-based metrics can give a decent basic estimate of translation quality, human annotations of translation errors would for example enable more fine-grained analysis.

Acknowledgements

We are thankful to Oqaasileriffik for answering our initial questions, and to the anonymous reviewers for their helpful feedback. EP, JB and HL are funded by the Carlsberg Foundation, under the Semper Ardens: Accelerate programme (project nr. CF21-0454).

References

Fatma Yumuk Akın Özçift, Kamil Akarsu and Cevhernur Söylemez. 2021. Advancing natural lan-

[11] The concept of "imaginaries" refers to the collective symbols, ideas, and images that shape a society's understanding (Taylor, 2004).

guage processing (nlp) applications of morphologically rich languages with bidirectional encoder representations from transformers (bert): an empirical case study for turkish. *Automatika*, 62(2):226–238.

Ali Araabi, Christof Monz, and Vlad Niculae. 2022. How effective is byte pair encoding for out-of-vocabulary words in neural machine translation? In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 117–130, Orlando, USA. Association for Machine Translation in the Americas.

Anna Berge. 2011. *Topic and discourse structure in West Greenlandic agreement constructions*. U of Nebraska Press.

Christian Berthelsen. 1997. *Grønlandsk dansk ordbog*. Atuakkiorfik, Uddannelsesforl.

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

Steven Bird. 2024. Must NLP be extractive? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Bangkok, Thailand. Association for Computational Linguistics.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Barbera De Mol et al. 2020. A comparison of data-driven morphological segmenters for low-resource polysynthetic languages: A case study of greenlandic.

Ramy Eskander, Judith Klavans, and Smaranda Muresan. 2019. Unsupervised morphological segmentation for low-resource polysynthetic languages. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–195, Florence, Italy. Association for Computational Linguistics.

Eduardo D Faingold. 2023. Language rights and the law in greenland. In *Language Rights and the Law in Scandinavia: Sweden, Denmark, Norway, Iceland, the Faroe Islands, and Greenland*, pages 241–264. Springer.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An NLP benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3553–3565, Torino, Italia. ELRA and ICCL.

Michael Fortescue. 1986. *Inuktun – An Introduction to the Language of Qaanaaq, Thule*. Institut for Eskimologi, University of Copenhagen, Copenhagen.

Michael Fortescue and Lise Lennert Olsen. 2022. The acquisition of west greenlandic. In *The crosslinguistic study of language acquisition*, pages 111–219. Psychology Press.

Michael D Fortescue. 1984. *West greenlandic*. Croom Helm London.

Ulrik Pram Gad. 2017. What kind of nation state will greenland be? securitization theory as a strategy for analyzing identity politics. *Politik*, 20(3).

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Adam James Patrick Gaudry. 2011. Insurgent research. *Wicazo Sa Review*, 26:113 – 136.

Daniela Gerz, Ivan Vulić, Edoardo Ponti, Jason Naradowsky, Roi Reichart, and Anna Korhonen. 2018. Language modeling for morphologically rich languages: Character-aware modeling for word-level prediction. *Transactions of the Association for Computational Linguistics*, 6:451–465.

Lenore A Grenoble. 2011. On thin ice: language, culture and environment in the arctic. *Language Documentation and Description*, 9.

Lenore A Grenoble and Lindsay J Whaley. 2021. Toward a new conceptualisation of language revitalisation. *Journal of Multilingual and Multicultural Development*, 42(10):911–926.

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2014. Morfessor flatcat: An hmm-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185.

M. Guillemin, L. Gillam, E. Barnard, P. Stewart, H. Walker, and D. Rosenthal. 2016. "we're checking them out": indigenous and non-indigenous research participants' accounts of deciding to be involved in research. *International Journal for Equity in Health*, 15.

Ximena Gutierrez-Vasques, Christian Bentz, and Tanja Samardžić. 2023. Languages Through the Looking Glass of BPE Compression. *Computational Linguistics*, 49(4):943–1001.

Maura Hanrahan. 2017. Enduring polar explorers' arctic imaginaries and the promotion of neoliberalism and colonialism in modern greenland. *Polar Geography*, 40(2):102–120.

H. Hassani, C. Beneki, S. Unger, M. Mazinani, and M. Yeganegi. 2020. Text mining in big data analytics. *Big Data and Cognitive Computing*, 4:1.

Alex Jones. 2022. Finetuning a kalaallisut-english machine translation system using web-crawled data. *arXiv preprint arXiv:2206.02230*.

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.

Lily Kahn and Riitta-Liisa Valijärvi. 2021. *West Greenlandic: an essential grammar*. Routledge.

Katharina Kann, Jesus Manuel Mager Hois, Ivan Vladimir Meza-Ruiz, and Hinrich Schütze. 2018. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 47–57, New Orleans, Louisiana. Association for Computational Linguistics.

Kevin Kelly. 2020. An evaluation of parallel text extraction and sentence alignment for low-resource polysynthetic languages.

Judith L. Klavans. 2018a. Computational challenges for polysynthetic languages. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Judith L. Klavans, editor. 2018b. *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*. Association for Computational Linguistics, Santa Fe, New Mexico, USA.

Ross Kristensen-Mclachlan and Johanne Nedergård. 2024. A new benchmark for Kalaallisut-Danish neural machine translation. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 50–55, Mexico City, Mexico. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

R. Kuokkanen. 2017. 'to see what state we are in': First years of the greenland self-government act and the pursuit of inuit sovereignty. *Ethnopolitics*, 16:179 – 195.

Tan Ngoc Le and Fatiha Sadat. 2020. Revitalization of indigenous languages through pre-processing and neural machine translation: The case of inuktitut. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4661–4666.

Yu Lu Liu, Meng Cao, Su Lin Blodgett, Jackie CK Cheung, Alexandra Olteanu, and Adam Trischler. 2023. Responsible AI considerations in text summarization research: A review of current practices. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

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. *arXiv preprint arXiv:2305.19474*.

Manuel Mager, Elisabeth Mager, Alfonso Medina-Urrea, Ivan Vladimir Meza Ruiz, and Katharina Kann. 2018. Lost in translation: Analysis of information loss during machine translation between polysynthetic and fusional languages. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 73–83, Santa Fe, New Mexico, USA. 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.

Felicity Meakins and Carmel O'Shannessy. 2016. *Loss and renewal: Australian languages since colonisation*, volume 13. Walter de Gruyter GmbH & Co KG.

Philippe Mennecier. 1995. *Le tunumiisut, dialecte inuit du Groenland oriental: Description et analyse*, volume 78. Peeters Publishers.

Marianne Mithun. 2017. Argument marking in the polysynthetic verb. In *The Oxford Handbook of Polysynthesis*, pages 30–59. Oxford University Press.

Aquigssiaq Møller. 1988. Language policy and language planning after the establishment of the home rule in greenland. *Journal of Multilingual and Multicultural Development*, 9:177–179.

Flemming AJ Nielsen. 2021. Literacy and christianity in greenland. In *The Inuit world*, pages 187–206. Routledge.

Flemming AJ Nielsen. 2022. *Vestgrønlandsk grammatik*. BoD–Books on Demand.

Javad Nouri and Roman Yangarber. 2016. A novel evaluation method for morphological segmentation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3102–3109, Portorož, Slovenia. European Language Resources Association (ELRA).

Oqaasileriffik. 2017. Nutserut: The pre-2023 method. Accessed: 2024-10-14.

Oqaasileriffik. 2023. Nutserut: Hybrid artificial intelligence. Accessed: 2024-10-14.

The Language Secretariat of Greenland Oqaasileriffik. Resources - oqaasileriffik. Accessed: 2024-10-14.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Per Langgård and VISL Team. Hand-tagged closed corpus for greenlandic - panola project. Accessed: 2024-10-14.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Signe Ravn-Højgaard, Ida Willig, Mariia Simonsen, Naja Paulsen, and Naimah Hussain. 2018. *Tusagassiuutit 2018: en kortlægning af de grønlandske medier*. Ilisimatusarfik.

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.

Pierre Robbe and Louis-Jacques Dorais. 1986. *Tunumiit oraasiat = Tunumiut oqaasii = Det østgrønlandske sprog = The East Greenlandic Inuit language = La langue inuit du Groenland de l'Est*. Université Laval, Centre d'études nordiques, Québec.

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 *Fifth Conference on Machine Translation*, pages 274–281. Association for Computational Linguistics (ACL).

Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. 'just what do you think you're doing, dave?' a checklist for responsible data use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Søren Rud. 2009. A correct admixture: The ambiguous project of civilising in nineteenth-century greenland. *Itinerario*, 33:29 – 44.

Søren Rud. 2021. Governing sexual citizens: decolonization and venereal disease in greenland. *Scandinavian Journal of History*, 47:567 – 586.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Jerrold M. Sadock. 2003. *A Grammar of Kalaallisut (West Greenlandic Inuttut)*. LINCOM Europa. 21 cm.

Jonne Saleva and Constantine Lignos. 2021. The effectiveness of morphology-aware segmentation in low-resource neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 164–174, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Kathleen Siminyu, Sackey Freshia, Jade Abbott, and Vukosi Marivate. 2020. Ai4d–african language dataset challenge. *arXiv preprint arXiv:2007.11865*.

J. Singh. 2020. Natural language processing for studying consumer journey: a case study of sneaker shoppers. *International Journal of Research in Science and Technology*, 10:98–101.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24, Gothenburg, Sweden. Association for Computational Linguistics.

Charles Taylor. 2004. *Modern Social Imaginaries*. Duke University Press, Durham, NC.

Nicole Tersis. 2008. *Forme et sens des mots du tunumiisut: lexique inuit du Groenland oriental*. CNRS Editions, Paris.

Kirsten Thisted. 2021. Un-shaming the greenlandic female body. *On the Nude*.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Jörg Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

United Nations. 2023. Visit to denmark and greenland - report of the special rapporteur on the rights of indigenous peoples. United Nations General Assembly, Human Rights Council, A/HRC/54/42/Add.2.

Riitta-Liisa Valijärvi and Lily Kahn. 2020. The linguistic landscape of nuuk, greenland. *Linguistic Landscape*, 6(3):265–296.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Ekaterina Vylomova, Trevor Cohn, Xuanli He, and Gholamreza Haffari. 2017. Word representation models for morphologically rich languages in neural machine translation. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 103–108, Copenhagen, Denmark. Association for Computational Linguistics.

K. Watson and D. Payne. 2020. Ethical practice in sharing and mining medical data. *Journal of Information Communication and Ethics in Society*, 19:1–19.

Indrė Žliobaitė. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089.

Fernando Zúñiga. 2019. Polysynthesis: A review. *Language and Linguistics Compass*, 13(4):e12326. E12326 LNCO-0774.