Evaluating the potential of language-family-specific generative models for low resource data augmentation: a Faroese case study

Barbara Scalvini, Iben Nyholm Debess
The University of the Faroe Islands
J. C. Svabosgøta 14, 100 Tórshavn
barbaras@setur.fo, ibennd@setur.fo

Abstract

We investigate GPT-SW3, a generative language model for the Nordic languages, to assess its understanding of the low-resourced Faroese language. Our aim is to demonstrate the advantages of using language-family-specific generative models to augment data for related languages with fewer resources. We evaluate GPT-SW3 by prompting it for Faroese to English translation in a zero, one, and few-shot setting. We assess such translations with an ensemble score consisting of an arithmetic average between the BLEU and a semantic similarity score (SBERT). Moreover, we challenge the model's Faroese language understanding capabilities on a small dataset of curated Faroese trick sentences. There, we make a qualitative comparison of the model's performance with respect to Open AI's GPT-3.5 and GPT-4, demonstrating the advantages of using a language-family-specific generative model for navigating non-trivial scenarios. We evaluate the pipeline thus created and use it, as a proof of concept, to create an automatically annotated Faroese semantic textual similarity (STS) dataset.

Keywords: Semantic Textual Similarity, Low-resource, Machine Translation, Data Augmentation

1. Introduction

In recent times, the popularity and performance of GPT-like language models has seen a dramatic increase. However, we are only beginning to explore the potential of such models for low-resource languages. Generative language models have shown promising results for translation in a zero or few-shot learning settings, among other types of tasks (Brown et al., 2020). Here, we investigate the language understanding capabilities of GPT-SW3 (Ekgren et al., 2022, 2023), a large-scale generative language model for the Nordic languages, in the context of Faroese – a low-resource Scandinavian language. We do so by 1) prompting GPT-SW3 for translating the Faroese split of the FLORES 200 test set (Team et al., 2022) to English and 2) testing its Faroese understanding over a new dataset of 50 carefully crafted Faroese trick sentences. Here, we compare here GPT-SW3's performance with OpenAI's GPT-3.5 (OpenAI (ChatGPT), 2021) and GPT-4 (OpenAI, 2023).

GPT-SW3 was not originally trained on Faroese language data; however, the inherent linguistic similarities between Faroese and its close Nordic relatives can ensure efficient transfer learning. We perform empirical evaluation of the translations by comparing the translated English sentences with the original English FLORES 200 dataset via an ensemble score considering translation quality and semantic similarity.

Current trends in Natural Language Processing have witnessed a shift from using task specific models to employing general purpose Large Language Models (LLMs) in a few-shot setting. However, such models are generally highly expensive in terms of hardware requirements and computational resources, which might also be lacking in a low-resource setting. Therefore, we believe smaller, lighter ad-hoc models still have a role in language technology. General purpose LLMs can then be exploited for knowledge distillation and data augmentation, for the development and training of such task-specific models. As a proof of concept of the potential of the method to annotate and augment data for low-resource languages, we exploit our pipeline to create a new Faroese semantic textual similarity (STS) dataset, by translating and automatically annotating the Faroese BLARK corpus (Debess et al., 2022). The dataset thus created represents, to our knowledge, the first native Faroese STS dataset.

2. Background and related work

Faroese has for the most part been overlooked by advancements in NLP, because of its relatively small native speaker population (55,000[1]). A notable effort to enhance progress in this direction was the compilation of a Basic Language Resource Kit for Faroese (Simonsen et al., 2022). Despite the low resource availability, the close relation between Faroese and its Scandinavian relatives makes it a prime candidate for transfer learning (Mena et al., 2023; Snæbjarnarson et al., 2023). Therefore, we want to demonstrate the potential of a Scandinavian

[1] https://hagstova.fo/fo/folk/folkatal/folkatal

LREC-COLING 2024, pages 6496–6503
20-25 May, 2024.

large scale language model, GPT-SW3 (Ekgren et al., 2022, 2023), for the Faroese language. Our approach stems from recent methodologies that employ LLMs instead of custom-designed models for, e.g., machine translation (Brown et al., 2020). The ability of GPT models for translation is currently being evaluated (Hendy et al., 2023; Peng et al., 2023) and sometimes expanded with the use of human-readable dictionaries in a low-resource setting (Elsner and Needle, 2023). Moreover, GPT models can be used contingently with machine translation models to improve performance, e.g., for sentence augmentation (Sawai et al., 2021) or post-editing (Raunak et al., 2023). Synthetic datasets created via machine translation have been proven to be highly valuable for low resource data augmentation (Tars et al., 2021).

2.1. Typological background

Faroese is an Insular Scandinavian language, closely related to the other Nordic languages. Historically developed from Old Norse, parallel to Icelandic and Norwegian, modern Faroese shares many linguistic similarities to these languages regarding morphology (e.g. similar inflectional paradigms), phonology (e.g. pre-aspiration of stops) and syntax (e.g. adverbial placement). Being under Danish rule since 1814 has resulted in language contact and similarities between Faroese and Danish (and Mainland Scandinavian), especially on a lexical and syntactic level (see more in Thráinsson et al. (2012)). These linguistic features make Faroese an interesting case for investigating transfer learning from related languages, as Faroese is typologically situated at the centre of all the Nordic languages.

3. Method

3.1. Prompting GPT-SW3 for translation

We carried out Faroese to English translations on the test split of the Faroese FLORES 200 dataset using zero, one, and few-shot approaches. The examples used originated from the Sprotin sentences dataset, an English-Faroese parallel corpus (Mikkelsen, 2021). Detailed information on all prompts can be found in Supplementary Materials A. For the one-shot prompts, we introduced two variants: 1) random selection of examples from the first 100 entries of the Sprotin dataset for every translation query and 2) a selection from a pool of 5 hand-picked, high quality examples for every translation query. High quality is defined in the following terms: 1) a similarity score of 5 (see scale in 3.2), as assigned by a linguist, 2) unambiguous meaning of all words, 3) simple syntax (declarative sentences or interrogative sentences, excluding subordinate clauses or sentences), 4) lack of typographical and inflectional errors. For the few-shot prompt, the five hand-picked examples were combined together.
We employed AI Sweden's GPT-SW3 models for all translations, experimenting with models of various sizes (6.7B, 20B, and 40B parameters), and different temperature parameters. All results presented in the paper were calculated with the 40B parameter model and the best performing temperature value: 0.1 (Supplementary Material B). We found that increased temperature often introduces additional details with respect to the source into the translations, compromising their quality.
The English translation was then compared sentence by sentence with the original English FLORES 200 dataset, by automatic score.

3.2. Evaluating translation quality

In order to quantify translation quality automatically, three different scores were tested and benchmarked against human evaluation: a semantic similarity score calculated via Sentence-BERT (Reimers and Gurevych, 2019), the BLEU (Papineni et al., 2002) and the chrF score (Popović, 2015). Two annotators evaluated manually by assigning scores using the scale detailed in Cer et al. (2017). This scale spans from 0, entirely dissimilar sentences, to 5, complete equivalent sentences. The final human score was determined by averaging the values given by both annotators. To validate the quality assessment method used throughout the paper, we proceeded as follows:

- We performed zero shot translation over the dataset as described in section 3.1, and assigned a preliminary SBERT sentence similarity score by comparing the English translation with the original FLORES 200 English split.
- This preliminary score allowed us to select a subset of 50 translated sentences to be manually evaluated. To ensure a balanced representation of both poor and good translations, we divided the dataset into three bins based on semantic similarity scores: 0 - 0.33, 0.33 - 0.66, and 0.66 - 1.0. Sentences were then randomly sampled from these bins in equal proportions.
- The subset of selected translations was evaluated manually by the annotators, and the manual scores were correlated with the SBERT, BLEU and chrF scores.

The SBERT semantic similarity score seems to overestimate poor translations (Figure 1, central panel), while the BLEU (Figure 1, left panel) and the chrF (Supplementary Materials C) score underestimate good translations - possibly because they both rely on string matching and do not take into account synonyms. Since BLEU and chrF present

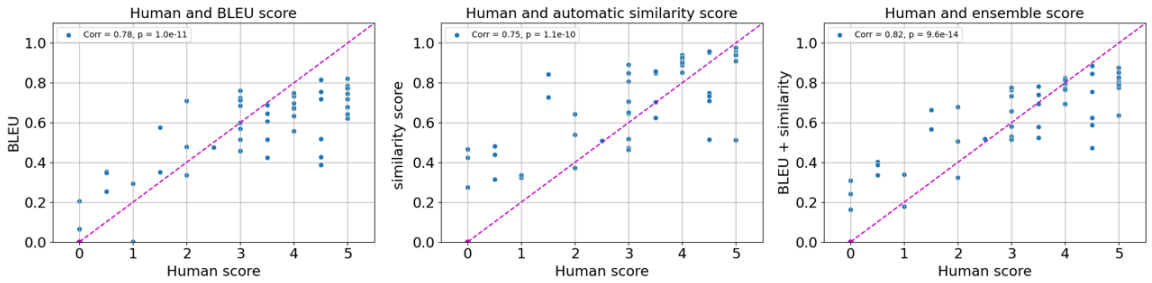

Figure 1: Correlations between semantic similarity, BLEU and ensemble score with human evaluation. The legend displays the Pearson correlation coefficient and its p value. The reference dotted line displays a 1:1 correspondence between automatic and human score.

a similar pattern, and the former shows slightly better correlation with the human score (0.78 versus 0.71), we decided to only consider the BLEU score for the analysis. We then considered combining the BLEU and semantic similarity score into an *ensemble* score, as a way to mitigate the respective pitfalls. By averaging the two scores, we obtain a better correlation between the automatic and human score than by using the two scores separately (Figure 1). We express the goodness of the translation with the following parameters:

- The percentage of sentences displaying ensemble score higher than 0.7 (successful).
- The percentage of sentences displaying ensemble score between 0.5 and 0.7 (inconclusive).
- The percentage of sentences displaying ensemble score lower than 0.5 (unsuccessful).
- The percentage of entries which failed to yield any English translation (failed).

We observed how all translations yielding an ensemble score higher than 0.7 fell in the range of human score between 3 and 5 (Figure 1). This range indicates translations which span from roughly to completely equivalent to the reference target. On the other hand, all translations yielding ensemble scores below 0.5 were labelled with human scores ranging from 0 to 2, indicating that the two sentences are not equivalent. After validating these metrics on the subset of 50 sentences, these were then used for assessing the full Faroese FLORES 200 dataset (1012 sentences).

3.3. Analysing linguistic nuances in translation quality

When assessing the back-translations of the sentences from FLORES 200, we found that uncommon Faroese words negatively influence the translation quality. According to our observations, the models handle rare words in mainly two ways, due to statistical bias stemming from training data and model size: a) context-based generalisation **(type 1)**, or b) linguistically-informed generalisation, based on token commonalities with the other Nordic languages **(type 2)**. To investigate these observations in detail, we hand-crafted a small **trick sentence dataset** (50 sentences). The sentences were categorised as follows:

- **Baseline sentences**: common language and content. Example: *'A dog usually has four legs, two ears and a tail.'*
- **Trick sentences**: same as baseline, with one word changed to be out of context, logically unlikely, or surprising. Example: *'A dog usually has **beautiful (føgur)** legs, two ears and a tail.'*
- **Nonsensical sentences**: same as baseline, with one word changed to be nonsensical, though linguistically well structured. Example: *'A dog usually has **[nonsense] (frúnk)** legs, two ears and a tail.'*

In the trick sentence, a **type 1** translation of *føgur* would be 'four'. An example of **type 2** translation would be 'beautiful', as the word 'føgur' has cognates with the same meaning in the forms of 'fagur' in Icelandic and 'fager' in the other Scandinavian languages. The translations of baseline and trick sentences were manually labelled as successful or unsuccessful. Translations of trick and nonsense sentences were also marked for translation bias: type 1 or type 2. We translated all sentences through GPT-3.5, GPT-4[2] and GPT-SW3 using the same prompt configuration.

4. Results and Discussion

4.1. Translating Faroese to English with GPT-SW3

Table 1 presents the outcomes—successful, unsuccessful, and failed translations—across five runs for

[2]GPT-3.5 and GPT-4 were accessed in September 2023 via ChatGPT.

	Zero-shot	One-shot_1	One-shot_2	Few-shot	Few-shot Sw
S (%)	79.31 ± 0.78	81.71 ± 0.51	82.65 ± 0.74	84.82 ± 0.80	97.17 ± 0.30
I (%)	18.45 ± 0.80	16.20 ± 0.52	16.54 ± 0.80	14.74 ± 0.83	2.65 ± 0.25
U (%)	1.91 ± 0.23	0.84 ± 0.24	0.69 ± 0.22	0.43 ± 0.05	0.12 ± 0.11
F (%)	0.31 ± 0.08	1.22 ± 0.34	0.12 ± 0.11	0 ± 0	0.06 ± 0.05

Table 1: Percentages of successful (S, ensemble score > 0.7), inconclusive (I, 0.5 < ensemble score < 0.7), unsuccessful (U, ensemble score < 0.5) and failed (F) translation for zero, one, and few-shot settings.

each translation setting. Additionally, we've added a Swedish baseline in the final column of Table 1, as the model was primarily optimised for Swedish. Here, translations were performed from Swedish to English, by using the same optimal settings identified for Faroese (few-shot).

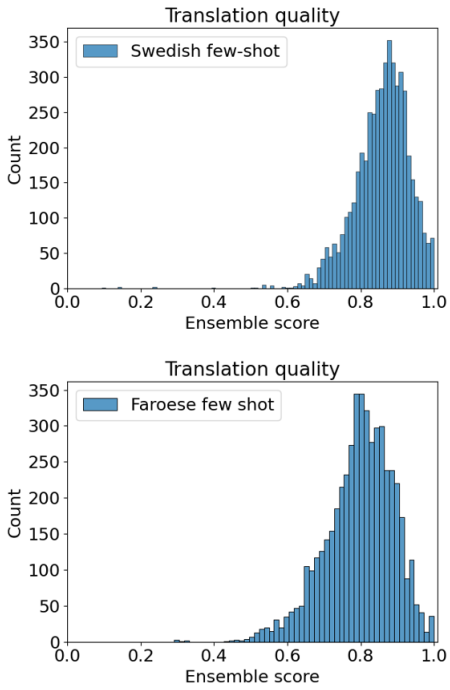

Figure 2: Distribution of the ensemble score calculated by translating the FLORES 200 dataset with GPT-SW3 in a few-shot setting, for Swedish (top panel) and Faroese (bottom panel).

Many sentences get a lower ensemble score because the model generates extra pieces of sentence on top of the English translation. This aspect can be mitigated by feeding examples of the translation task, as demonstrated by the higher performance of the few-shot setting (Table 1). The full distribution of the few-shot ensemble scores over the dataset is presented in Figure 2, together with its Swedish counterpart. The overall performance of GPT-SW3 clearly indicates that the model does understand Faroese. These results might be due to enhanced linguistic transfer between Faroese and its Scandinavian relatives. Another possible reason could be the presence of spurious Faroese in the training set, possibly mislabelled as Icelandic. We did find evidence of some Faroese words in the Icelandic Gigaword Corpus (Steingrímsson et al., 2018), which is part of GPT-SW3's training data. However, it is difficult to estimate the amount of Faroese present and whether it is enough to explain our results.

We might want to investigate which factors, in the source sentences, play a role in yielding good translation quality. A key consideration is the tokenisation of Faroese words versus Scandinavian counterparts. Analysis of FLORES 200 shows that successful translations often have a higher token overlap between Faroese and other Scandinavian languages (p value = 4×10^{-4}, as calculated by Mann-Whitney U test) . Moreover, we find that longer sentences, statistically, yield better translation quality (p value =2×10^{-7}, as calculated by T-test). This could be attributed to the enhanced contextual insight that models derive from extended sentences. FLORES 200, initially composed in English, might have cultural nuances which are not typical to Faroese. Many of its Faroese terms, especially technical ones, aren't frequently used by natives and may be limited in scope. Such discrepancies, part of a phenomenon called linguistic formality gap (Jacobsen, 2021), affect Faroese users and potentially also translation quality.

4.2. Qualitative assessment of translation biases

We assessed translation quality and biases on the trick sentence dataset. All three models were successful over the 23 baseline sentences with 21 (GPT-3.5), 21 (GPT-4) and 23 (GPT-SW3) good translations. GPT-SW3 outperformed GPT-3.5 and GPT-4 over the 19 trick sentences, yielding 13 successful translations, against the 5 and 6 produced respectively by the other two models. For both trick and nonsense sentences, GPT-SW3 showed a preference for type 2 generalisation (9 versus 21 translations), while GPT-3.5 and GPT-4 showed a preference for type 1 (24 versus 6 and 22 versus 8 translations respectively).

For example, "Tey fýra elementini eru vatn, **hiti**, jørð og luft." - *The four elements are water, **heat**, earth and air.*, translated by the models as follows: **GPT-3.5**: "The four elements are water, **fire**, earth,

and air.", **GPT-4**: "The four elements are water, **fire**, earth, and air., and **GPT-SW3**: "The four elements are water, **heat**, earth, and air.
The example shows how GPT-SW3 recognises 'hiti' correctly as 'heat', a word having similar translations in other Nordic languages. The model tokenises the words identically in Faroese and Icelandic.
It is worth noting that GPT-4 provided additional comments to some of the translations, letting us know that it was aware of the translations being linguistically inaccurate and explaining the reason for translating otherwise. This preliminary qualitative analysis suggests that GPT-SW3, GPT-3.5 and GPT-4 have different biases in the translation task from Faroese to English. Moreover, our results indicate that cultural bias is a relevant factor: the model based on Nordic language data, GPT-SW3, demonstrated a higher ability to consider linguistic nuances in translation. The size of this trick dataset limits any statistical conclusions. However, the specifics of the sentences allow for a valuable qualitative analysis and results, which should be further investigated in future research. Awareness of the inference mechanisms and generalisation performed by the model is crucial, especially in a low-resource setting. Linguistic transfer can be leveraged as a whole, but it should be used with caution when handling culturally specific elements.

4.3. Creation of the STS dataset

As a proof of concept of the potential of the methods here described for data augmentation, we created an automatically annotated semantic textual similarity (STS) dataset[3]. In order to do so, we filtered sentences from a native Faroese corpus, the BLARK. The first selection was based on sentence size: we kept sentences from 29 to 59 token long (as tokenized by GPT-SW3's tokenizer), as this range matches the distribution of sentence size present in the FLORES 200 dataset (average = 44.3, standard deviation = 15.4 tokens), for which translation evaluation was performed. This procedure yielded about 100,371 sentences, that were subsequently translated with GPT-SW3 (few shot) to English. The translations were then filtered for linguistic correctness using heuristics: presence of a root in the dependency parse tree, verbs, capitalisation and punctuation respectively at the beginning and end of a sentence. Annotation of the remaining 68,260 translated sentences was then performed as follows:

- 5000 unique sentences were randomly extracted from the translated dataset.
- each unique sentence was then compared for semantic similarity (SBERT, multi-qa-mpnet-base-dot-v1 model) with 7000 unique sentences extracted randomly.
- This procedure yielded 5000 x 7000 unique pairings with correspondent similarity score.
- The scores were then mapped to discreet labels (0 to 5). The annotations thus created were projected back to the original Faroese sentences.

In order to create a balanced dataset, we enriched the sentence pairs thus produced with 'equivalent' sentences (label 5). We did so by randomly selecting 400 sentences from the English translations and and back-translating them to Faroese with GPT-SW3. These sentences were checked and manually corrected by a human expert, as proficiency of GPT-SW3 for English to Faroese translation is still to be assessed. Finally, we randomly sampled 200 sentence pairs for each class, to have equal representation of all classes.

5. Conclusion

For data augmentation in low-resource languages, our Faroese study suggests the benefit of using smaller, family-specific generative models over vast multilingual ones. This method may limit broad-task reasoning but boosts culturally relevant knowledge transfer. In a "bigger is better" age, these insights are vital for equitable NLP resource distribution and decentralisation. When comparing GPT-SW3 with OpenAI's models, it is important to note the difference in dataset transparency. While the dataset used to train GPT-SW3, the Nordic Pile (Öhman et al., 2023), is openly available, Open AI does not disclose the specifics of its datasets. Working with GPT-3.5 and 4 raises transparency issues, so full quantitative comparison of the two models cannot be made - as we cannot exclude that these models already had access to the FLORES 200 dataset in their training phase. We can, however, exclude that these models accessed our trick sentences dataset, as it was crafted specifically for this study. Other possible experimental settings could be explored in the future to exploit LLMs for data augmentation, such as using the instruction-tuned version of the model to directly assign labels for language understanding tasks, such as, for example, POS, NER and STS.

6. Supplementary Materials

A. Prompt engineering

The following prompts were found, via iteration and refining to be the most effective for Faroese to English translation:

- zero shot prompt: *Translate the following sentence to English:* ‹*sentence*›

[3] https://huggingface.co/datasets/barbaroo/STS

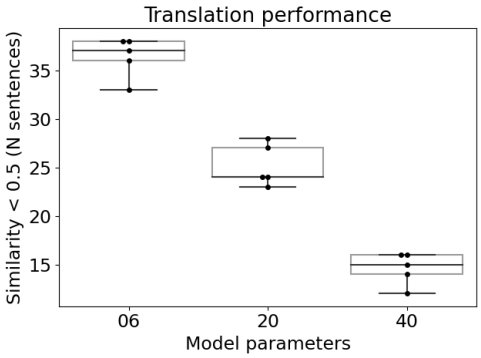

Figure 3: Number of unsuccessful translations (ensemble score < 0.5) obtained with zero-shot translation performed with three different sizes of the GPT-SW3 model: 6.7, 20 and 40 billion parameters.

- one/ few shot prompt: *I need you to translate sentences from Faroese to English.*
  1. *The Faroese sentence <Faroese example> is translated to English as <English example>.*
  2. *The Faroese sentence <Faroese example> is translated to English as <English example>.*
  3. *The Faroese sentence <Faroese example> is translated to English as <English example>.*
  4. *The Faroese sentence <Faroese example> is translated to English as <English example>.*
  5. *The Faroese sentence <Faroese example> is translated to English as <English example>.*
  6. *The Faroese sentence <sentence> is translated to English as*

	T = 0.1	T = 0.3	T = 0.6	T = 0.9
S (%)	79.54	78.16	68.37	49.60
I (%)	18.18	19.17	26.58	39.03
U (%)	1.87	2.27	4.25	8.40
F (%)	0.39	0.39	0.79	2.96

Table 2: Percentages of successful (S, ensemble score > 0.7), inconclusive (I, 0.5 < ensemble score < 0.7), unsuccessful (U, ensemble score < 0.5) and failed (F) for zero shot translation and four different settings of temperature: T = 0.1, 0.3, 0.6, 0.9.

B. Model settings

We tested GPT-SW3 models of sizes 6.7, 20 and 40 billion parameters (Figure 3). Unsurprisingly, the biggest model was found to be the best performing one. Tested temperature values were 0.9, 0.6, 0.3 and 0.1. The best results were achieved for temperature = 0.1 (Table 2), indicating that a more deterministic output is better for the translation task. All experiments in this study were performed with nucleus sampling $top_p = 1$ and maximum new tokens produced by the model equal to 120.

C. Validation: chrF score

The CHaRacter-level F-score (chrF) was tested for translation evaluation and benchmarked against human evaluation (Figure 4) by Pearson correlation coefficient. When benchmarked against human evaluation, the chrF score presents a trend similar to the BLEU score: it tends to overestimate poor translations and underestimate good translations.

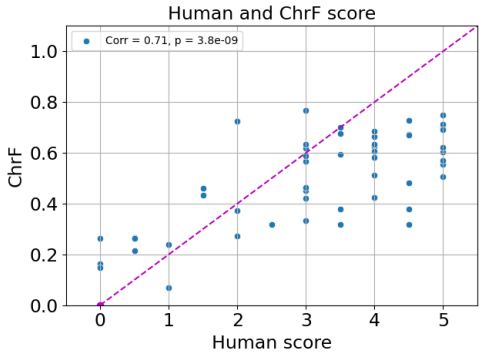

Figure 4: Correlation between chrF score and human evaluation. The legend displays the Pearson correlation coefficient and its p value. The reference dotted line displays a 1:1 correspondence between automatic and human score.

7. Bibliographical References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual

and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ariel Ekgren, Amaru Cuba Gyllensten, Evangelia Gogoulou, Alice Heiman, Severine Verlinden, Joey Öhman, Fredrik Carlsson, and Magnus Sahlgren. 2022. Lessons learned from GPT-SW3: Building the first large-scale generative language model for Swedish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3509–3518, Marseille, France. European Language Resources Association.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Alice Heiman, Judit Casademont, and Magnus Sahlgren. 2023. Gpt-sw3: An autoregressive language model for the nordic languages. Https://arxiv.org/abs/2305.12987.

Micha Elsner and Jordan Needle. 2023. Translating a low-resource language using GPT-3 and a human-readable dictionary. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–13, Toronto, Canada. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Jógvan í Lon Jacobsen. 2021. *Føroysk purisma: Føroysk orð ella orð í føroyskum*. Fróðskapur, Faroe Islands.

Carlos Hernández Mena, Annika Simonsen, and Jon Gudnason. 2023. Asr language resources for faroese. pages 32–41. University of Tartu Library.

OpenAI. 2023. Gpt-4 technical report.

OpenAI (ChatGPT). 2021. Gpt-3.5: A large-scale language model. Accessed on 2023-10-08.

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.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *ArXiv*, abs/2303.13780.

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.

Vikas Raunak, Amr Sharaf, Hany Hassan Awadallah, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ranto Sawai, Incheon Paik, and Ayato Kuwana. 2021. Sentence augmentation for language translation using gpt-2. *Electronics*, 10(24).

Annika Simonsen, Sandra Saxov Lamhauge, Iben Nyholm Debess, and Peter Juel Henrichsen. 2022. Creating a basic language resource kit for Faroese. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4637–4643, Marseille, France. European Language Resources Association.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on faroese. pages 728–737. University of Tartu Library.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A very large Icelandic text corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Maali Tars, Andre Tättar, and Mark Fišel. 2021. Extremely low-resource machine translation for closely related languages. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 41–52, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Höskuldur Thráinsson, Hjalmar Páll Petersen, Jógvan í Lon Jacobsen, and Zakaris Svabo Hansen. 2012. *Faroese. An overview and reference grammar*, 3 edition. Faroe University Press/Linguistic Institute of Iceland, Tórshavn/Reykjavík.

Joey Öhman, Severine Verlinden, Ariel Ekgren, Amaru Cuba Gyllensten, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, and Magnus Sahlgren. 2023. The nordic pile: A 1.2tb nordic dataset for language modeling.

8. Language Resource References

Debess, I. N. and Lamhauge, S. S. and Simonsen, A. and Henrichsen, P. J. and Hofgaard, E. and Johannesen, U. and Hammer, P. M. J. and Brimnes, G. H. and Thomsen, E. M. D. and Poulsen, B. 2022. *Basic LAnguage Resource Kit 1.0 for Faroese*. Talutøkni. OpenSLR. [link].

Jonhard Mikkelsen. 2021. Sprotin sentences. https://raw.githubusercontent.com/Sprotin/translations/main/sentences_en-fo.strict.csv. Accessed: October 13, 2023.