What's Wrong With This Translation? Simplifying Error Annotation For Crowd Evaluation

Iben Nyholm Debess
University of the Faroe Islands
ibennd@setur.fo

Alina Karakanta
Leiden University
karakantaa@vuw.leidenuniv.nl

Barbara Scalvini
University of the Faroe Islands
barbaras@setur.fo

Abstract

Machine translation (MT) for Faroese faces challenges due to limited expert annotators and a lack of robust evaluation metrics. This study addresses these challenges by developing an MQM-inspired expert annotation framework to identify key error types and a simplified crowd evaluation scheme to enable broader participation. Our findings based on an analysis of 200 sentences translated by three models demonstrate that simplified crowd evaluations align with expert assessments, paving the way for improved accessibility and democratization of MT evaluation.

1 Introduction

The Faroese language, with its limited resources and relatively small speaker community, currently lacks widely accepted automatic evaluation metrics akin to those available for more commonly spoken languages. At the same time, the scarcity of expert linguists and professional translators makes traditional, metric-intensive human evaluations both inviable and costly. A potential avenue for overcoming these challenges is to harness the insights and judgments of native speakers through crowdsourcing. This requires a simple and accessible framework, allowing everyday language users to effectively assess the quality of Faroese machine translation (MT) outputs.

In this study, we conducted a Multidimensional Quality Metrics (MQM)-inspired analysis to identify the most frequent error types in English-to-Faroese Machine Translation (MT) outputs from three distinct models—GPT-SW3, NLLB, and Claude 3.5 Sonnet—using a new dataset of 200 sentences. These initial explorations revealed key error patterns and categories, which guided the development of a tailored evaluation approach that accommodates Faroese linguistic nuances. Building on these insights, we designed a prototype crowd annotation framework by simplifying and adapting the error dimensions, aiming to engage a broader pool of evaluators.

These insights can inform the future development of a simplified, crowd-friendly evaluation framework. Such a framework could ultimately facilitate the collection of crowd-sourced evaluation data, fostering the creation of a Faroese MT benchmark and associated neural metrics. Over time, these resources could support the curation of open parallel data, thereby facilitating the training and enhancing the performance of upcoming Faroese MT systems.

2 Background / Related work

Faroese has been under-represented in MT research due to limited resources and scarce parallel data. Initiatives like Meta's NLLB, Google's MADLAD 400 (Kudugunta et al., 2023), and the integration of Faroese into Google Translate (Bapna et al., 2022) aim to address this. Large language models (LLMs) such as GPT-4 and Claude 3.5 Sonnet have improved Faroese translation and text generation (Debess et al., 2024; Simonsen and Einarsson, 2024; Scalvini et al., 2025b). Nordic-focused LLMs like GPT-SW3 outperform broader models (e.g. GPT-4) in culturally nuanced tasks (Scalvini and Debess, 2024), though smaller fine-tuned MT models can surpass LLMs (Scalvini et al., 2025b). The scarcity of gold-standard parallel data remains a challenge, with efforts focused on data augmentation and synthetic data creation (Scalvini and Debess, 2024; Simonsen, 2024; Scalvini et al., 2025b). Evaluating Faroese MT systems is difficult as standard automatic metrics overlook linguistic nuances (Scalvini et al., 2025a), and human evaluation is constrained by the lack of expert Faroese linguists. While benchmarks like FLORES-200 have provided some par-

Proceedings of the 1st NB-REAL Workshop 2025 - Nordic-Baltic Responsible Evaluation and Alignment of Language models, pages 42–47

Terminology	Accuracy	Linguistic conventions	Miscellaneous
Wrong term (*term-w*)	Mistranslation, major (*acc-x*)	Noun morphology (*ling-n*)	Style (*misc-s*)
Inconsistent use of term (*term-i*)	Mistranslation, minor (*acc-n*)	Adjective morphology (*ling-a*)	Localization (*misc-l*)
Foreign word/phrase	Overtranslation (*acc-v*)	Verb morphology (*ling-v*)	Named Entities (*misc-ne*)
from English (*term-fe*)	Undertranslation (*acc-u*)	Adverb morphology (*ling-d*)	Source error (*misc-c*)
from Icelandic (*term-fi*)	Addition (*acc-a*)	Wrong syntax (*ling-sy*)	
from Mainl. Scand. (*term-fs*)	Omission (*acc-o*)	Other grammar errors (*ling-o*)	
Sensible neologism (*term-s*)		Punctuation (*ling-p*)	
Non-sensible neologism (*term-n*)		Spelling (*ling-sp*)	

Table 1: Main Error Categories and Subcategories in the ECS-D.

Evaluation task	Scale
Direct Assessment	0-5
"The translation uses wrong words" (repr. *Terminology*)	tick
"The translation is incomplete" (repr. *Accuracy*)	tick
"The translation has inflectional errors" (repr. *Linguistic*)	tick

Table 2: Simplified evaluation scheme for crowd: ECS-S. Only DA is required, others are optional.

allel data for evaluating MT systems for Faroese, they often fail to capture Faroese cultural contexts, dialectal variations, and sociolinguistic factors such as the formality gap (Jacobsen, 2021).

3 Method

3.1 Dataset and Models

To test the English-to-Faroese MT quality, we first compiled a small dataset[1] of 200 English sentences, sourced mainly from the English versions of a Faroese news outlet and from municipal documents. This selection ensures that we have English-language content that is relevant in Faroese settings. The dataset was translated into Faroese with three different models: GPT-Sw3 (1.3B, Ekgren et al. (2024)), a NLLB (1.3B, NLLB Team et al. (2022)) and Claude 3.5 Sonnet (October, 2024, (Anthropic, 2024)). In this work, we utilize an NLLB model fine-tuned for English-Faroese translation, introduced in (Scalvini et al., 2025b)[2]. We selected these models to represent the range of options for English-Faroese translation: a multilingual NMT system, an open source, language-family specific LLM, and one closed-source, commercial LLM. Both LLMs were few-shot prompted using five high-quality examples, selected by an expert from the Sprotin Corpus (Mikkelsen, 2021).

[1] [https://huggingface.co/datasets/ibennd/sentences_eng-lang_cont-fao]

[2] [https://huggingface.co/barbaroo/nllb_200_1.3B_en_fo]

4 Experimental Design

Initially, a small subset of the data was analyzed to identify typical translation errors, using an error categorization scheme derived from the Multidimensional Quality Metrics (MQM) (Burchardt, 2013). Insights from this preliminary analysis guided the development of a more tailored expert evaluation framework (Table 1). After full expert evaluation, the results informed a simplified framework for crowd evaluation. The main steps of the experimental design were as follows:

1. Initial Evaluation (Subset):
   - Evaluate a subset (50 sentences, randomly sampled from the sentences sourced from news) translated with all three models, using MQM-inspired categories.
   - Identify frequent, impactful error types.
   - Expand on and retain common error categories while simplifying or removing those with few or no observed instances.
   - Develop a revised, more targeted expert error scheme: *Error Categorization Scheme Detailed* (ECS-D).
2. Full Expert Evaluation (Full Dataset):
   - Translate all 200 sentences with three models.
   - Perform expert evaluation (one human expert) with ECS-D: assign Direct Assessment (DA) scores (0-5) and categorize errors into main and subcategories (Terminology, Accuracy, Linguistic Conventions, Miscellaneous) (Table 1).

- Analyze correlations between DA and error categories to identify which errors affect overall perceived quality and compare model performance.

3. Simplified Crowd Evaluation (Full Dataset):
   - Derive a simplified evaluation scheme based on ECS-D findings: *Error Categorization Scheme Simple* (ECS-S).
   - Use DA (0-5) plus three "tickable" boxes corresponding to the most frequent/impactful errors from ECS-D, phrased for non-experts (Table 2).
   - Have a group of 19 language users evaluate the 200 sentences (around 67 from each model; one set of 10 for each user) and compare crowd results with expert evaluation to assess alignment.

Recent works have shown benefits of using the ESA framework for evaluating MT (Kocmi et al., 2024; Scalvini et al., 2025b). The ESA is less detailed than the MQM, and could potentially fit both expert and crowd evaluators. However, ESA does not give us information on error types. In Faroese MT, identifying frequent error types helps target specific issues in training and evaluation.

Model	Expert DA	Crowd DA	Rank
GPT-SW3	2.74 ± 1.15	2.16 ± 1.60	3
NLLB	4.28 ± 0.68	3.56 ± 1.16	2
Claude	4.40 ± 0.64	4.35 ± 0.70	1

Table 3: Mean scores and standard deviation of expert and crowd DA for the three models and ranking.

Model	Expert r	Crowd r	Weight. Crowd r
GPT-SW3	-0.29	-0.37	-0.56
NLLB	-0.75	-0.71	-0.69
Claude 3.5	-0.80	-0.76	-0.75

Table 4: Pearson correlation scores between DA and number of errors for expert and crowd evaluation. Marked in yellow: $r > 0.25$. Marked in green: $r > 0.75$. All $p < 0.05$.

5 Results and Discussion

5.1 Expert evaluation analysis

The overall performance of the three systems is given in Table 3, based on DA. Claude achieved the highest translation quality, closely followed by NLLB. GPT-SW3's score reflects substantial issues with translation quality and consistency, which is expected given it is a small-sized LLM.

Expert Correlation Scores			
Model	**Terminology**	**Accuracy**	**Linguistic**
GPT-SW3	-0.09	-0.46	-0.050
NLLB	-0.63	-0.37	-0.35
Claude 3.5	-0.58	-0.19	-0.48
Crowd Correlation Scores			
Model	**Terminology**	**Accuracy**	**Linguistic**
GPT-SW3	0.08	-0.60	-0.18
NLLB	-0.49	-0.56	-0.32
Claude 3.5	-0.72	-0.44	-0.29
Correlation between Expert and Crowd			
Model	**Terminology**	**Accuracy**	**Linguistic**
All	0.35	0.50	0.45

Table 5: Pearson correlation scores between main error categories and DA. Marked in yellow: $r > 0.25$ and $p < 0.05$.

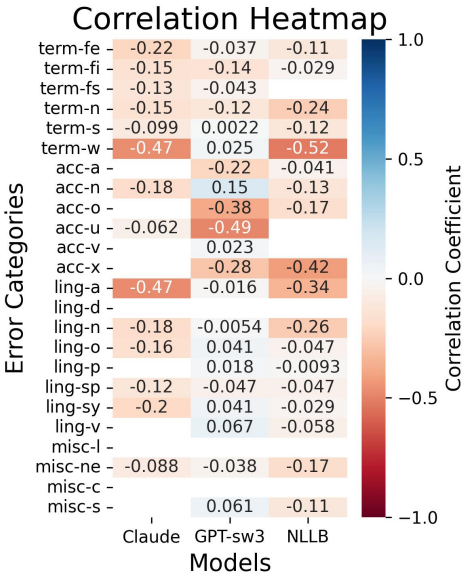

Figure 1: Heatmap of Pearson correlations between subcategorized errors and DA for all models.

Looking at the correlation between DA and number of errors for each sentence in Table 4, we see a high correlation for NLLB and Claude, but a lower correlation for GPT-SW3. This low correlation may stem from ignoring error severity. Given GPT-SW3's poor performance, a few significant errors could heavily impact translation quality. To determine which error types have the greatest impact on perceived translation quality, we correlated all error types with the DA score (Figure 1). Most impactful error types appeared to be model-specific. For GPT-SW3, 'Undertranslation' (*acc-u*) showed to significantly impact quality, while 'Omissions' and 'Major Mistranslations' also contributed. For NLLB, 'Wrong

term' (*term-w*) had the strongest negative correlation, followed by 'Major Mistranslations' and 'Adjective morphology' errors. In Claude, 'Adjective morphology' (*ling-a*) was most impactful, followed closely by 'Wrong term'. Though less frequent, 'Foreign words' still affected perceived quality. These findings informed the phrasing of error categorization for crowd users (Table 2), focusing on *wrong words* ('Wrong term', 'Foreign words'), *incomplete translations* ('Undertranslation', 'Omission', 'Major mistranslation'), and *inflectional errors* ('Adjective morphology', 'Noun morphology'). Table 5 shows similar patterns at the main categories: NLLB and Claude align with Terminology errors, while GPT-SW3 correlates moderately with Accuracy, reflecting its highest subcategory correlations with 'Undertranslation', 'Omission', and 'Major mistranslation'.

5.2 Crowd evaluation analysis

The overall scores from the crowd evaluation align with the expert evaluation, showing a correlation of r=0.78 (p=1.17e-42) between Expert and Crowd DA. The ranking of models is also preserved (Table 3).

In the expert evaluation, the number of errors is descriptive of the actual error count for each sentence and is in principle unlimited. However, in the simplified framework, the error count for each sentence has only four possible values (0-1-2-3), as each of the three error types is ticked as either present or not: 0 if no errors are present, 3 if all error types are present. This simplification was necessary to allow non-experts to annotate. Even though the information is less granular with respect to expert evaluation, we still calculated the correlation between error type presence and crowd DA. This was done in order to confirm that the same error types are perceived as most impactful by both crowd and expert annotators. The correlation between number of error types and crowd DA score can be seen in Table 4. Looking at error categories, the correlation scores between error categories and DA (Table 5) demonstrate a very similar pattern for expert and crowd. Although the magnitude of the correlation can differ, both crowd and expert annotators tend to agree on the ranking of most impactful mistakes, with the notable exception of NLLB's scores, where crowd perceives Accuracy as most impactful, as opposed to Terminology in expert annotation. Comparing expert and crowd by correlation, Table 5 (last row) shows that experts and the crowd agree most on Accuracy errors, which are often easily perceived by non-experts, and least on Terminology, which requires more in-depth knowledge of specialized language.

5.3 Hybridizing crowd and expert annotation for augmented evaluation

Looking at Tables 4 and 5, we notice that both expert and crowd annotation methods provide low correlation scores for GPT-SW3. This is probably because these frameworks do not consider error severity, an impactful parameter when model performance is overall low. In an attempt to provide a more informative quantifier for overall translation quality, we defined a weighted sum of the error categories in ECS-S. Specifically, we used correlations between expert DA and error count for the main error categories (Table 5) as weights for summing the number of errors:

$$N_{\mathrm{W}} = C_{\mathrm{T}} \cdot T + C_{\mathrm{A}} \cdot A + C_{\mathrm{L}} \cdot L \qquad (1)$$

where C_{T} is the model-specific expert correlation for the category Terminology, T represents the Terminology error value (1 or 0, present or not present), C_{A} and A the equivalent values for Accuracy and C_{L}, L those for Linguistic errors. Ideally, the expert correlation scores should inform us on how much each error category impacts overall quality. The rationale behind these weights is an attempt to augment crowd annotation with expert knowledge. A hybrid approach combining a small number of expert annotations and a larger pool of crowd evaluators could be a viable solution for resource-constrained settings. By applying this weighing, we observe an improvement in overall correlation between crowd error count and crowd assigned DA score for GPT-SW3 (Table 4), from -0.37 to -0.56. The models with higher correlations do not seem to benefit from this modification. This aligns with the observation that distinguishing error gravity is more crucial for weaker models, as top-performing models predominantly make minor mistakes.

5.4 Assessing Bias in Subset Reuse

A potential issue in the evaluation process arises from the fact that the subset for initial evaluation (50 sentences) — analyzed to identify frequent error types for developing the ECS-D — were also part of the full 200-sentence evaluation dataset. This approach could introduce bias,

as certain error categories might be overrepresented in the subset, potentially affecting both expert and crowd evaluations. To examine this, we conducted a post-hoc analysis of correlation scores across all subcategories, comparing the subset and the remaining 150 sentences separately. We focused on GPT-SW3, the most error-prone model, providing the most informative insights despite the limitations of analyzing only one model. The results indicate that overall correlation patterns remain consistent between the subset and the other sentences. While some subcategories exhibit stronger correlations within the subset, others display higher correlations in the remaining dataset. Many subcategories maintain similar correlation values across both sets, suggesting that the process of using the subset for identifying error types and subsequently incorporating it into the full evaluation does not significantly distort the results.

6 Conclusion and Future Work

This study underscores the importance of error analysis in identifying language- and model-specific challenges in low-resource MT evaluation. Our expert framework, ECS-D, effectively identified frequent and impactful error types, while the simplified crowd evaluation framework, ECS-S, demonstrated overall alignment with expert assessments. This alignment paves the way for expanding the annotator pool, collecting evaluation data for low-resource languages. This study represents preliminary work toward a full crowd evaluation framework, suitable for the creation of a Faroese-specific neural metric, and for the promotion of targeted data collection efforts to address common translation mistakes efficiently. Furthermore, the adaptability of this framework makes it a promising approach for other under-resourced languages, allowing for systematic error identification and tailored evaluation strategies.

7 Limitations

Established evaluation frameworks, such as MQM and ESA, typically account for error severity, which is then used to weight errors into a cumulative score. In our study, we conducted a first-order error analysis aimed at identifying the types of errors that most significantly impact perceived translation quality among Faroese speakers.

At this stage, we chose not to incorporate error severity, a decision that proved to be a limiting factor for the lowest-performing model, GPT-SW3. In this model, a few major errors could substantially affect the overall quality. In the final evaluation framework, we will include error severity, designed in a way that allows non-expert language users to annotate it effectively.

Despite this limitation, we believe that our first-order analysis provides valuable insights into which error types have the highest impact from a native speaker's perspective.

For example, a high-performing model like Claude primarily produces high-level linguistic mistakes (e.g., inflectional errors) that do not significantly hinder the effective comprehension of the translation. In contrast, a less effective model tends to generate highly impactful errors in translation accuracy, such as mistranslations and undertranslations. These categories may require different weights, in addition to considering whether each error is classified as major or minor within its respective category.

Another limitation of this study is the involvement of only one language expert and the evaluation of each sentence by only one crowd annotator, which may undermine the statistical power of the analysis. Although the preliminary results show encouraging agreement between the expert and crowd annotations, it would be ideal to include multiple expert annotators in the development of the final evaluation framework. This shortcoming could be mitigated by calculating z-scores from the DA. However, the impact of that avenue may be limited, as we are primarily examining correlation, which is insensitive to average values.

References

Anthropic. 2024. Claude 3.5 sonnet. `https://www.anthropic.com`. Proprietary software, closed-source.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation

quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Iben Nyholm Debess, Annika Simonsen, and Hafsteinn Einarsson. 2024. Good or bad news? Exploring GPT-4 for sentiment analysis for Faroese on a public news corpora. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7814–7824, Torino, Italia. ELRA and ICCL.

Ariel Ekgren, Amaru Cuba Gyllensten, Felix Stollenwerk, Joey Öhman, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, Judit Casademont, and Magnus Sahlgren. 2024. GPT-SW3: An autoregressive language model for the Scandinavian languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7886–7900, Torino, Italia. ELRA and ICCL.

Jógvan í Lon Jacobsen. 2021. *Føroysk Purisma*. Fróðskapur, Faroe University Press.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. *arXiv preprint arXiv:2406.11580*.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset.

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

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.

Barbara Scalvini and Iben Nyholm Debess. 2024. Evaluating the potential of language-family-specific generative models for low-resource data augmentation: A Faroese case study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6496–6503, Torino, Italia. ELRA and ICCL.

Barbara Scalvini, Iben Nyholm Debess, Annika Simonsen, and Hafsteinn Einarsson. 2025a. Prompt engineering enhances Faroese MT, but only humans can tell. In *Proceedings of the 25th Nordic Conference on Computational Linguistics (NoDaLiDa)*, Talinn, Estonia. Forthcoming; accepted for publication.

Barbara Scalvini, Annika Simonsen, Iben Nyholm Debess, and Hafsteinn Einarsson. 2025b. Rethinking Low-Resource MT: The Surprising Effectiveness of Fine-Tuned Multilingual Models in the LLM Age. In *Proceedings of the 25th Nordic Conference on Computational Linguistics (NoDaLiDa)*, Talinn, Estonia. Forthcoming; accepted for publication.

Annika Simonsen. 2024. Improving Machine Translation for Faroese using ChatGPT-Generated Parallel Data. Master's thesis, University of Iceland, Reykjavík.

Annika Simonsen and Hafsteinn Einarsson. 2024. A Human Perspective on GPT-4 Translations: Analysing Faroese to English News and Blog Text Translations. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 24–36, Sheffield, UK. European Association for Machine Translation (EAMT).