Building a One-Million-Pair Bokmål–Nynorsk Translation Corpus: A Quality-First Harvesting and Cleaning Pipeline

Per Egil Kummervold*, **Thea Tollersrud***, **Angelina Zanardi***
National Library of Norway
per.kummervold@nb.no, thea.tollersrud@nb.no, angelina.zanardi@nb.no

Abstract

We present a high-quality parallel corpus for translation between Norwegian Bokmål (nb) and Nynorsk (nn), two closely related written standards of Norwegian. The corpus was assembled from two complementary sources: Nasjonal digital læringsarena (NDLA), an educational platform, and Nynorsk pressekontor (NPK), a newswire service. Our methodology prioritizes precision over volume, employing a multi-stage filtering pipeline designed to address the specific challenges of aligning near-neighbor languages. This pipeline combines paragraph-level alignment, deduplication, multilingual semantic similarity scoring, language identification confidence checks, structural consistency tests, and strict bidirectional adjudication by a Large Language Model (LLM). To address the common problem of untranslated or placeholder "pending" copies, we apply a rule that flags pairs with zero semantic distance when the Nynorsk side shows weak evidence of being distinctively Nynorsk. After filtering, we retained 191,695 pairs from NDLA and 809,164 pairs from NPK, resulting in a merged corpus of 1,000,859 parallel paragraphs. This resource demonstrates that a precision-oriented pipeline can produce data better suited for training robust machine translation systems and instruction-tuned models than larger but noisier alternatives.

Keywords: Norwegian Bokmål, Nynorsk, parallel corpus, data cleaning, semantic similarity, language identification, LLM adjudication

1. Introduction

Norwegian Bokmål and Nynorsk are the two official written standards of Norwegian. While their lexical and syntactic proximity makes large-scale parallel corpus construction appear straightforward, it in fact introduces unique alignment challenges. Near-identical surface forms often conceal subtle orthographic or stylistic differences, and automatic alignment systems frequently misclassify identical or near-identical paragraphs as valid translations when one side is merely an untranslated placeholder.

Despite their linguistic proximity, high-quality Bokmål–Nynorsk parallel resources remain limited. Such resources are important for tasks such as machine translation and evaluation of multilingual language models trained on Norwegian data.

This paper details the making of a one-million-pair parallel corpus with an explicit focus on precision. Our objective was to develop and document a staged, interpretable filtering pipeline that maximizes translational quality while maintaining broad topical coverage. By systematically addressing each error class with a dedicated processing stage and transparent thresholds, we produce a reliable resource for downstream NLP tasks such as Machine Translation. The key contribution is a detailed, reproducible methodology for creating high-precision corpora for closely-related language varieties, where traditional alignment methods often fall short.

We release the dataset under **CC-BY**.

2. Related work

Most modern bitext-mining approaches make use of sentence embeddings to automatically extract parallel sentence pairs from large collections. For example, the FISKMO project (Tiedemann et al., 2020) showed that multilingual sentence embeddings can align Finnish and Swedish texts: they embed sentences from each language into a shared vector space, then retrieve nearest neighbors as translation candidates by comparing sentence embeddings using cosine similarity.

A large-scale extension of this idea, CCMatrix (Schwenk et al., 2021), mined billions of sentence pairs from Common Crawl snapshots, discovering for instance about 17.7 million Danish–Norwegian bitexts (they do not specify nn vs. nb) by aligning Danish and Norwegian web sentences with multilingual embeddings. Such systems typically translate or index documents and then match similar sentences, leveraging the fact that related languages (like Danish vs. Norwegian) return many correct alignments on the open web.

We build on these embedding-based alignment methods but adapt them to the closely related written standards Norwegian Bokmål and Norwegian Nynorsk, integrating them into a multi-stage filtering pipeline. In addition to semantic similarity scoring, we incorporate language identification confidence and structural consistency checks. Our approach retains the advantages of CCMatrix-style methods

*All authors contributed equally.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8551–8555
11–16 May 2026.

for detecting near-equivalent translation pairs, while leveraging a LLM to identify and filter out cases where noise, omissions, or other inconsistencies have been introduced in one of the texts.

3. Sources and Licensing

The corpus is built from two sources chosen for their distinct registers, which improves the stylistic diversity of the final dataset while maintaining consistent orthographic norms.

Nasjonal digital læringsarena (NDLA)[2]**:** A public educational resource providing articles across a wide range of academic subjects for Norwegian high school students. They offer learning resources for 143 high school subjects, such as Norwegian, history, social sciences, math and science. The learning resources are written in two forms, Norwegian Bokmål and Nynorsk. We harvested articles that explicitly list parallel Bokmål and Nynorsk variants. Creator and license metadata are preserved for each paragraph pair.

The data is released under **CC-BY**.

Nynorsk pressekontor (NPK) [3]**:** A national newswire agency that works to increase the use of Nynorsk in the Norwegian press. We processed a collection of parallel news articles curated and manually inspected by the Norwegian Language Bank [4] and normalized them into a common schema. The subset used for this work is licensed under **CC0**.

4. Quality-First Design Principles

The development of our cleaning pipeline was guided by four core principles:

- **Precision Over Volume:** For near-neighbor language pairs like nb–nn, false positives and noisy alignments can disproportionately degrade model performance. Our pipeline is intentionally designed to trade recall for higher reliability.
- **Explainable Gates:** Each filtering step and its associated threshold (e.g., semantic distance, language ID confidence) is mapped to a specific, detectable error class. This makes the process auditable and allows for clear analysis of attrition at each stage.
- **Redundant Signals:** Rather than relying on a single heuristic, we combine multiple signals: semantic embeddings, language identification models, structural heuristics, and LLM judgments, in order to form a more robust and reliable filtering cascade.
- **Provenance Preservation:** Every paragraph pair that passes the filtering process retains essential metadata, including its source, original IDs or URLs (for NDLA), creator information, and license. This supports responsible and informed use of the data.

5. The Cleaning and Filtering Pipeline

Our process is structured as a sequence of stages, each targeting a specific type of noise. The pipeline is applied independently to the NDLA and NPK sources before the final merge.

Paragraphs serve as coherent semantic units, offering a more robust basis for alignment than sentences, which can have minor segmentation differences due to morphological or stylistic variation. This choice follows similar reasoning to large-scale parallel corpus projects such as ParaCrawl, where paragraph-level alignment has proven more reliable than sentence-level heuristics for noisy or heterogeneous data sources (Esplà-Gomis et al., 2019). For NDLA, articles are split into paragraphs, and pairs are formed by index. Articles with an unequal paragraph count between language versions are discarded to prevent misalignment. NPK data was already paragraph-aligned.

To foster topical diversity and prevent domain over-representation, we remove exact duplicates based on the Bokmål text. This step is performed early to reduce the computational cost of subsequent, more intensive stages.

5.1. Semantic Similarity Filtering

We use the multilingual embedding model `BAAI/bge-m3` to generate vector representations for each Bokmål and Nynorsk paragraph (Xiao et al., 2024). We chose this model because at the time we conducted the experiments, `BAAI/bge-m3` was one of the highest performing embedding models on bitext mining tasks for Scandinavian languages (including nn and nb) according to the Massive Text Embedding Benchmark (MTEB) Muennighoff et al. (2023). The semantic distance is calculated as $1 - \text{cosine\_similarity}(\mathbf{nb}, \mathbf{nn})$. Through manual inspection of borderline cases, we established a threshold of ≤ 0.15. We set this threshold in consensus by manually inspecting the output, and were conservative in our approach; aiming to reduce the number of observed errors as much as possible. Pairs exceeding this distance, which often indicate topic drift or gross misalignment, are removed.

[2] https://ndla.no/

[3] https://www.npk.no/

[4] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-80/

5.2. Language ID Confidence and the Zero-Distance Rule

A major source of noise comes from identical nb–nn pairs, which typically represent untranslated placeholders rather than legitimate equivalence. These pairs exhibit a semantic distance of exactly 0. However, removing all such cases would erroneously penalize naturally identical "radical" variants; these are forms that are valid in both Bokmål and Nynorsk. To resolve this, we combine semantic distance with language identification confidence. Using the `cis-lmu/glotlid` model (Kargaran et al., 2023), we inspect all pairs with zero distance and reject those where the predicted Nynorsk confidence falls below 0.10. This hybrid rule effectively removes placeholder duplicates while preserving genuine shared forms.

5.3. Structural Consistency Checks

Simple structural and pattern-based checks serve as reliable proxies for misalignment or incomplete edits. We filter out pairs with:

- Mismatched end-of-paragraph punctuation (e.g., one ends with a period, the other a question mark).
- Discrepancies in the numbers present in each text.
- Malformed or inconsistent newswire prefixes for NPK data (e.g., "(NPK-NTB)").
- Compares counts of structural characters (-, –, —, /, «», quotes, parentheses, etc.) between nb and nn and discards mismatches.

While the consistency checks filter out problematic pairs, we acknowledge that we are also dropping otherwise good pairs due to typographic or stylistic variation. For instance, the pair *"Jentene spelte ein særs svak kamp og fall med 32–33 for Tyskland."* and *"Jentene spilte en meget svak kamp og falt med 32-33 for Tyskland."* (English: *"The girls played a particularly weak match and lost by 32-33 against Germany."*) was filtered out because the dashes between the numbers `32` and `33` were different.

5.4. Bidirectional LLM Adjudication

As the final filter, we used DeepSeek-V3-0324 (DeepSeek-AI, 2024) to adjudicate the translational quality of remaining pairs. To ensure reproducibility, we use a consistent scoring rubric and deterministic decoding parameters. Each pair is evaluated bidirectionally (nb→nn and nn→nb) to detect asymmetric divergences in register or terminology. The model outputs a structured JSON rating across five dimensions: *Adequacy*, *Fluency*, *Terminology*, *Style*, and *Surface Accuracy*:

- **Adequacy:** Is the meaning preserved?
- **Fluency:** Is the target text well-formed?
- **Terminology:** Are technical terms translated correctly?
- **Formality/Style:** Is the register consistent?
- **Surface Accuracy:** Are there grammatical or spelling errors?

The model is asked to score each pair on a 1-5 Likert scale across each dimension. In line with our quality-first approach, only pairs receiving a perfect 5/5 in both directions are retained (Kocmi and Federmann, 2023). The model is also required to produce a short textual justification (1–2 sentences) for each score. We designed the prompt in this way because previous work has shown that requiring models to justify their ratings leads to more consistent and deliberative evaluations than purely numeric scoring (Chiang and Lee, 2023). We initially evaluated a small subset of pairs and manually inspected the model's outputs, adjusting the instructions until the model's judgments aligned with our human judgements.

The exact prompt is released in the accompanying Github repository to ensure reproducibility.

5.5. Conservative Normalization

The final text undergoes minimal normalization using `ftfy` to correct Unicode encoding artifacts. This is done carefully to avoid erasing meaningful orthographic distinctions between Bokmål and Nynorsk.

6. Corpus Size and Attrition

The multi-stage pipeline was applied to both datasets, resulting in the retention of high-quality pairs. Table 1 details the number of pairs removed at key filtering stages.

The merged corpus totals 1,000,859 high-quality paragraph pairs, achieving our target size while adhering to our quality-first principles.

7. Suitability for Translation and Instruction Tuning

The resulting corpus exhibits high semantic fidelity and stable stylistic correspondences between Bokmål and Nynorsk, making it well suited for supervised machine translation and for the creation of instruction-style training data. Previous work has shown that translation quality depends strongly on

Table 1: Filtering attrition by stage.

Filtering Stage	NDLA	NPK	Merged
Initial Pairs	272,821	1,069,439	1,342,260
Deduplication	32,624	117,360	149,984
(4.1) Semantic Distance (> 0.15)	5,722	23,788	29,510
(4.2) Zero-Distance + Low `nn_nn_conf`	20,408	11,908	32,316
(4.3) Structural Mismatches	2,188	36,978	39,166
(4.4) LLM Adjudication (Score $< 5/5$)	26,539	93,550	120,089
Final Pairs Kept	**191,695**	**809,164**	**1,000,859**

Note: Attrition counts per stage are not mutually exclusive. A single paragraph pair may trigger multiple filters, meaning the sum of removed pairs across individual stages exceeds the total number of purely discarded pairs.

data quality, and that cleaner corpora can outperform larger but noisier alternatives (Kocmi and Federmann, 2023; Kreutzer et al., 2022). Our pipeline follows this quality-first perspective and is designed specifically for a low-distance language pair, where seemingly small inconsistencies can introduce substantial noise.

In addition to conventional MT training, the paragraph pairs are also suitable for the construction of instruction-tuning datasets. For example, they can be reformatted as prompt–response pairs such as "Translate the following text from Bokmål to Nynorsk: ...". The accompanying code repository includes scripts for converting the corpus into such formats.

8. Limitations and Future Work

Register Balance: The corpus is dominated by educational texts and newswire articles. Other registers, such as fiction, conversational language, and social media, are underrepresented.

Threshold Sensitivity: The thresholds for semantic distance (0.15) and language identification confidence (0.10 in the zero-distance rule) were chosen through empirical inspection and may not transfer optimally to other domains or alignment scenarios.

LLM Evaluator Bias: The LLM adjudication step inherits potential biases and blind spots from the evaluator model. Different evaluator models or prompts may yield different retention decisions.

Paragraph Granularity: The corpus is aligned at the paragraph level. Although this is often more robust than sentence-level alignment for these data sources, sentence boundaries within aligned paragraphs do not always correspond one-to-one.

Finally, although the use of LLMs as automatic evaluators is promising, it raises open questions about cross-lingual calibration, consistency, and robustness (Bavaresco et al., 2024; Wang et al., 2024). Future work could explore ensemble-based or reference-based adjudication strategies, as well as downstream experiments that quantify the effect of this corpus on translation and instruction-tuning performance.

9. Ethics Statement

This work follows ethical best practices by preserving provenance information and respecting source licensing terms. For the NDLA portion of the corpus, creator and license metadata are retained at the paragraph-pair level. The NDLA data is distributed under a **CC-BY** license, while the NPK subset used in this work is distributed under a **CC0** license. We also document the data sources and filtering decisions in detail, in line with the spirit of Data Statements (Bender and Friedman, 2018). By emphasizing transparency and data quality, we aim to provide a reliable resource for Norwegian NLP research.

10. Availability

The complete parallel corpus, containing 1,000,859 paragraph pairs, is available on the Hugging Face Hub.[5] The code used for harvesting, normalization, and filtering is also openly available.[6] The repository includes the full processing pipeline, configuration settings, and the prompts used for LLM-based adjudication.

11. Acknowledgements

We thank Tita Enstad for contributions to the corpus creation process, and Google's TPU Research Cloud (TRC) for computational support.

[5] https://huggingface.co/datasets/NbAiLab/merged_npk_ndla_parallel_paragraphs

[6] https://github.com/NationalLibraryOfNorway/nob-nno-translation-corpus

12. Bibliographical References

Andrea Bavaresco, Raffaella Bernardi, Luca Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Enas Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Paul Mondorf, Victor Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Anish Kumar Surikuchi, Ece Takmaz, and Andrea Testoni. 2024. Llms instead of human judges? a large-scale empirical study across 20 nlp evaluation tasks. *arXiv*.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek-v3 technical report.

Miquel Esplà-Gomis, M. Lluis Forcada, and Gema Ramírez-Sánchez. 2019. Paracrawl: Web-scale parallel corpora for the crawl era. In *Proceedings of MT Summit XVII*, pages 118–128.

Amir Hossein Kargaran, Amir Imani, François Yvon, and Hinrich Schütze. 2023. Glotlid: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the Eighth Conference on Machine Translation (WMT23)*, pages 220–233. Association for Computational Linguistics.

Julia Kreutzer et al. 2022. Quality at scale: Revisiting the impact of data quality on neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. 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 6490–6500, Online. Association for Computational Linguistics.

Jörg Tiedemann, Tommi Nieminen, Mikko Aulamo, Jenna Kanerva, Akseli Leino, Filip Ginter, and Niko Papula. 2020. The FISKMÖ project: Resources and tools for Finnish-Swedish machine translation and cross-linguistic research. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3808–3815, Marseille, France. European Language Resources Association.

Yiran Wang, Yiming Xu, Jonathan H. Clark, and Philipp Koehn. 2024. Can large language models serve as reliable automatic evaluators? In *Findings of the Association for Computational Linguistics: ACL 2024*.

Shaohan Xiao et al. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv*.