Optimized for AI: Curating the Icelandic Gigaword Corpus for Stable LLM Training

Jón Friðrik Daðason, Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
{jon.fridrik.dadason, steinthor.steingrimsson}@arnastofnun.is

Abstract

The Icelandic Gigaword Corpus (IGC) is a primary resource for Icelandic NLP, with its current version containing 2.7 billion words of curated text. The IGC is traditionally distributed in a TEI-XML format, a hierarchical structure that allows for rich linguistic annotation and metadata. However, this format introduces significant friction for modern machine learning workflows. Even high-quality curated corpora have been found to contain "unwanted" text sequences—such as fragmented lists or repetitive boilerplate that may trigger instabilities during training of large language models. In this paper, we present a new processing pipeline designed to optimize the IGC for AI development. We describe a filtering approach focusing on training stability, including fuzzy deduplication to reduce the risk of data leakage, with the aim to provide high-quality data for stable model convergence. Furthermore, we introduce a new JSONL distribution format that bridges the gap between TEI-XML and machine-actionable data, facilitating easier access and safer training for models aiming to work with Icelandic.

Keywords: LLM training data, Large corpora, Filtering

1. Introduction

Pre-training corpora for modern language models increasingly rely on web-crawled data. Although abundant, such data often contains substantial amounts of low-quality text and duplicate content. A manual evaluation of five multilingual web-crawled corpora found that 15 out of 205 language-specific subsets did not contain a single usable sentence, and that the proportion of usable text was below 50% in 87 subsets (Kreutzer et al., 2022). As a result, web-crawled corpora are typically filtered and deduplicated before pre-training. At the same time, Transformer-based language models have proven remarkably robust to noisy pre-training data. Filtering and deduplication often yield only modest average gains on downstream tasks, and the benefit of deduplication in particular is inconsistent across corpora and model scales (Raffel et al., 2020; Muennighoff et al., 2023).

The Icelandic Gigaword Corpus (IGC; (Steingrímsson et al., 2018; Barkarson et al., 2022)) is a curated, high-quality monolingual corpus for Icelandic, currently containing approximately 2.7 billion running words across a wide range of genres. In recent years, it has become the primary resource for pre-training and fine-tuning language models for Icelandic (Snæbjarnarson et al., 2022; Daðason and Loftsson, 2022; Daðason, 2025). Given the modest downstream impact often reported for text quality filtering and deduplication, it is reasonable to ask whether such preprocessing is necessary for a curated corpus such as the IGC. However, downstream performance is not the only consideration.

Recent work shows that even a small number of low-quality examples within a single training batch (e.g., highly repetitive n-gram sequences) can trigger immediate pre-training instability, potentially causing the model to plateau at worse performance or diverge during training (Walsh et al., 2025). Our own experience pre-training on the IGC suggests a similar pattern, where examples consisting primarily of non-running text (e.g., tabular content or lists of names, dates, or monetary amounts) or text in a foreign language appear to disproportionately contribute to training instability, despite representing only a small fraction of the corpus.

In this paper, we present a filtering and deduplication pipeline for the IGC, and we present a new JSON Lines (JSONL) distribution format for the corpus, optimized for LLM pre-training and fine-tuning, to be published alongside the TEI format.

2. Related Work

Text quality filtering is often performed using rule-based methods in which numerical features are extracted from text and compared against predetermined thresholds. Text is discarded if any feature falls outside an acceptable range. A major benefit of rule-based methods is that they are highly explainable, making them well-suited to settings where minimizing false positives is a priority. That said, there are no standard rulesets for text quality filtering, and the choice of rules and thresholds varies considerably between works.

The web-crawled C4 corpus (Raffel et al., 2020) used a comparatively simple heuristic filtering pipeline with both line-level and document-level rules. Lines were removed if they were very short,

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 49–56. Palma, Mallorca. May 11, 2026.

lacked a terminal punctuation mark, or matched boilerplate patterns (e.g., “terms of use”, “privacy policy”, or “cookie policy”). Documents were discarded if they contained certain strings indicating quality issues (e.g., “lorem ipsum”, “Javascript”, or “{”), or if a language classifier did not label them as English with high confidence.

The MassiveWeb corpus (Rae et al., 2022) applied document-level filtering based on a range of features. Documents were discarded if their word count or mean word length fell outside an acceptable range, or if they contained a high proportion of lines beginning with a bullet point, a high ratio of hash symbols or ellipses to words, a low proportion of words containing at least one alphabetic character, too few unique stop words, or a high proportion of repeated lines, paragraphs, or n-grams.

FineWeb (Penedo et al., 2024) reused many of the rules applied in C4 and MassiveWeb. In addition, documents were discarded if they had a high ratio of short lines, a high proportion of characters in duplicated lines, or a high proportion of lines ending without a terminal punctuation mark.

Despite the diversity of features and large rulesets applied in prior work, Daðason and Loftsson (2024) found that relatively few features had the greatest impact on text quality classification. Evaluating 13 commonly used features on a manually labeled dataset, they found that an unsupervised classifier achieved its highest F_1 score using only three: perplexity, stop word ratio, and mean subword length. The filtering pipeline described in this paper takes a similarly targeted approach, with rules derived directly from inspection of low-quality content found in the IGC rather than from prior work on web-crawled data.

The Text Encoding Initiative (TEI) guidelines (TEI Consortium, 2026) remain the standard for the long-term preservation and linguistic annotation of national corpora. They allow for nested structures and granular metadata with detailed information on everything from part-of-speech tags to licensing, provenance, and how the text was sourced and cleaned. TEI is used by national corpora such as the British National Corpus (Burnage and Dunlop, 1992), the Bulgarian National Corpus (Koeva et al., 2010), and the National Corpus of Polish (Przepiórkowski et al., 2008).

Although rich in metadata, the hierarchical complexity of TEI-XML introduces significant overhead for machine learning workflows. Consequently, recent projects have prioritized machine-actionable formats such as JSONL to facilitate high-throughput, stream-based data loading. The Norwegian Colossal Corpus (Kummervold et al., 2022) and the Danish Gigaword Corpus (Strømberg-Derczynski et al., 2021) are notable examples of this trend. Furthermore, other projects employ slightly different approaches to obtain the same goals. For example, the ParlaMint project (Erjavec et al., 2023), containing 17 parliamentary corpora in 16 main languages, is published both in the TEI-based ParlaMint format and a TSV-format that can easily be converted to JSONL or another format suitable for machine learning (ML) or NLP. While TEI will remain the master format for the IGC, we introduce a JSONL distribution alongside it to support language model pre-training and fine-tuning.

3. The IGC

Since first being released in 2018, the Icelandic Gigaword Corpus (IGC, Steingrímsson et al., 2018) has undergone significant expansion, with updated versions published annually or biannually. The corpus has grown from an initial 1.2 billion running words to its current iteration of approximately 2.7 billion words (Barkarson et al., 2022; Barkarson and Steingrímsson, 2024). The IGC is a tagged and lemmatized corpus. While news media constitutes the largest share of the data, the corpus is diverse, as seen in Table 1

The corpus is distributed under a dual-licensing scheme that balances open-access goals with copyright restrictions from various content providers. Approximately 63% of the corpus is available under a permissive license, CC BY 4.0, allowing for unrestricted use and redistribution. The largest categories under this license are *social media and internet forums* (∼30%), specific news outlets (∼17%) and parliamentary records (∼10%). The remaining 37% falls under the IGC Custom License, a more restrictive license which permits research use and training of language models but prohibits republication of raw texts. All downloads are centralized through the Árni Magnússon Institute for Icelandic Studies.

The IGC’s adoption has been broad, spanning corpus linguistics as well as modern NLP. On the linguistic side, it has been used to track language change, frequency distributions, and usage patterns, including as a resource underlying the maintenance of the Database of Modern Icelandic Inflection (Bjarnadóttir, 2012; Bjarnadóttir et al., 2019). In NLP, it has, e.g. served as a source for back-translation to generate synthetic parallel data for machine translation (Símonarson et al., 2021; Jasonarson and Steingrimsson, 2025), and as the core pretraining corpus for Icelandic encoder models (Snæbjarnarson et al., 2022; Daðason and Loftsson, 2022). Most recently, it provided training, validation, and test data for a language modeling task at GKÍ2026, the Icelandic AI competition[1].

[1]See: https://github.com/gervikeppnin/GKI2026/tree/main/golden_plate_on_thingvellir_NLP

Category	Running Words	(%)
Adjudications	79,625,568	2.95%
Law, bills and resolutions	60,623,312	2.24%
Published books	13,824,783	0.51%
Scientific/academic journals	20,894,101	0.77%
News media	1,442,810,126	53.43%
Parliamentary proceedings	266,115,169	9.85%
Social media	806,949,613	29.88%
Wikipedia	9,718,240	0.36%
Total	**2,700,560,912**	**100.00%**

Table 1: Distribution of running words and percentages across different text categories in the 2024 version of the IGC.

Despite this wide adoption, the IGC's primary distribution format presents a practical obstacle for LLM training. The corpus is encoded in TEI (Text Encoding Initiative) XML (TEI Consortium, 2026), a standard in corpus linguistics that provides rich structural and linguistic annotation. For language model training, however, this richness becomes friction: the deep nesting of XML tags means that raw text must be extracted through custom preprocessing pipelines, typically converting the corpus to plain text or JSONL while carefully preserving document boundaries and discarding markup. An unannotated TEI version introduced in 2022 reduces some of this overhead by providing cleaner sentence-level nodes, but the fundamental conversion step remains unavoidable, and doing it correctly is non-trivial. This paper addresses that gap directly, presenting a version of the IGC that has been processed and formatted specifically for LLM training.

4. The Processing Pipeline

In this section, we describe the processing pipeline used to filter and deduplicate the IGC. As the corpus contains a very low proportion of truly low-quality documents, we opted for a rule-based approach to text quality filtering in order to minimize false positives (i.e., erroneously removing high-quality text). Documents were first normalized, and boilerplate text was removed where possible. Documents containing issues that were not easily corrected, or that indicated deeper quality problems, were discarded entirely. Finally, although empirical evidence for the benefits of deduplication on downstream performance remains inconclusive, we applied fuzzy deduplication to reduce the risk of data leakage between training and validation splits.

4.1. Boilerplate Text

Many documents in the IGC contain strings that are irrelevant to their main content. This includes navigational elements, keywords, categories, lists of related articles, advertisements, social media links, and metadata. Additionally, we removed certain references to elements not present in the plain-text version of the documents, such as embedded videos or audio. We also removed frequently used signatures and author bylines, as their presence might bias models towards generating the same signatures or names in their output. We constructed a separate ruleset for each subcorpus, with each rule consisting of a regular expression pattern or a specific substring to be removed.

4.2. Escaped Elements

Some documents in the IGC contain escaped HTML or XML elements, such as > and & (representing the characters > and &, respectively). We unescaped all documents containing such elements. This sometimes required multiple unescape operations (e.g., first converting a doubly-escaped ampersand from & to &, and then to &). Some documents used custom escaped elements, which we unescaped by either inferring the appropriate form from the name of the element or by reviewing correctly rendered versions of the affected documents.

4.3. Character Normalization

Some documents contain nonprintable or otherwise undesirable characters, primarily private use area Unicode characters (code points reserved for private use) and certain ASCII control characters. We removed or normalized these characters as appropriate. Additionally, due to their rarity, we normalized all Unicode whitespace characters (e.g., no-break space and thin space) to either a literal space or a newline character.

4.4. Whitespace Normalization

Documents in the IGC have generally been preprocessed by collapsing multiple adjacent whitespace characters into one, stripping leading and trailing spaces from each line, and removing empty

lines. However, a small number of documents do not fully conform to this, and earlier steps in our pipeline may also have introduced or eliminated whitespace or left certain lines empty. We therefore repeated this whitespace normalization step to ensure a consistent input format.

4.5. Short Documents

It is common practice to remove short documents from pre-training corpora, although there is no standard approach to doing so (Raffel et al., 2020; Rae et al., 2022; Ettinger et al., 2025). Following Rae et al. (2022) and Penedo et al. (2024), we removed documents containing fewer than 50 words.

4.6. Stop Word Ratio

We calculated the ratio of stop words to all alphanumeric tokens within each document. Documents with a very low stop word ratio tend to consist primarily of non-running text (e.g., tabular data, lists, and bullet points), or foreign-language or non-linguistic content. We discarded documents whose stop word ratio fell below a minimum threshold of 22%.

4.7. Internal Duplication

We discarded documents in which at least 20% of sentences are duplicated. This can occur in short news articles that open with a brief preview, which is then repeated verbatim in the body of the article. For short articles, this preview can account for a significant portion of the total document length. Beyond this pattern, some documents contain unexpected instances of duplicate text, either present in the source file from which the text was extracted or mistakenly introduced during the extraction process itself. As it is difficult to determine which duplicate segments to remove, if any, we chose to discard documents with a high degree of internal duplication.

4.8. Code

Some documents contain unintended code, such as HTML, JavaScript, or XML, typically introduced when text is extracted from malformed source files. These are generally short and incomplete snippets that can be difficult to remove cleanly and may indicate deeper quality issues within the document. For this reason, we discarded any document containing code. Documents that intentionally contained code, such as Wikipedia articles on programming languages or computer science topics, were excluded from this filter.

4.9. Character Encoding Errors

Character encoding errors can be introduced when documents are decoded using an incorrect character encoding. For example, decoding a UTF-8-encoded file as Latin 1 will produce garbled output, such as rendering “ö” as “Ã¶”. Certain characters are strongly associated with such errors, and we discarded any document in which they were present.

4.10. Phrase Filtering

We identified and removed several classes of unwanted documents using targeted string matching. These include messages requesting that the reader log in, create an account, or subscribe in order to view the full contents. Documents containing such strings are often cut off mid-sentence, featuring only a heavily truncated version of the full text. We also removed documents containing warnings that JavaScript must be enabled to view the content, as these may indicate deeper quality issues. Finally, we discarded documents containing phrases closely associated with template-generated content, such as Wikipedia disambiguation pages.

4.11. Optical Character Recognition Errors

Some subcorpora in the IGC contain documents digitized using optical character recognition (OCR) software. On rare occasions, this process yields heavily garbled results, with a significant proportion of non-alphanumeric symbols that rarely occur in high-quality text. As correcting such errors is difficult, time-consuming, and risks introducing additional errors, we discarded documents where such symbols are present.

4.12. Old Documents

Documents published prior to 1930 often contain non-standard spelling and were thus discarded.

4.13. Fuzzy Deduplication

We performed fuzzy deduplication using MinHash and LSH with 20 bands of size 13. For each cluster of near-duplicate documents, we retained the longest document from the subcorpus whose documents had the highest overall pass rate on the preceding filtering steps, using this as a proxy for subcorpus quality.

5. The JSONL Format and Publication

To facilitate high-throughput training, we provide the processed IGC in JSONL format, as shown in

```
{
  "tei_archive": "IGC-News2-22.10.TEI.zip",
  "tei_path": "IGC-News2-22.10.TEI/fotbolti/2005/02/IGC-News2-fotbolti_1260328.xml",
  "source": "fotbolti.is",
  "altered": true,
  "text": "..."
}
```


Figure 1: Example of the new JSONL distribution format for the IGC. The inclusion of `tei_path` ensures full traceability to the original TEI-XML source.

Figure 1. Each line in the JSONL distribution represents a document in the corpus. It contains the raw text alongside essential provenance metadata, including whether it has been altered by the processing pipeline, as indicated in the "altered" field. This allows researchers to track any document back to its original TEI-XML source if the rich linguistic annotations (such as POS tags or lemmas) are required.

The IGC's dual-licensing model necessitates a split distribution strategy:

Open Access (CC BY 4.0): The portion of the corpus under permissive licensing is made available both in the CLARIN-IS repository[2] and on Hugging Face[3], which allows for immediate integration into standard ML data loaders. This part of the corpus is ≈795M running words after filtering.

Custom License: The other part of the corpus, published with the custom license which allows for research and training but restricts raw text redistribution, is only available through the CLARIN-IS repository[4]. This part of the corpus is ≈895M running words after filtering.

Users wishing to train on the full running 1.6 billion words of filtered data can easily merge the two JSONL streams. To support reproducible research, we also provide pre-defined training and validation splits for both portions of the corpus.

6. Discussion

Table 2 shows the number of documents and tokens discarded during the filtering and deduplication process. These statistics do not include documents from published books, journals, or social media, which are distributed as shuffled paragraphs or sentences for licensing and copyright reasons, making them unsuitable for inclusion in a pre-training corpus. The remaining text consists of 1.9 billion tokens, of which 1.6 billion are words, across 4.8 million documents. The most commonly discarded category by document count consists of documents containing fewer than 50 words, although these account for a negligible proportion of tokens in the IGC. The largest category by number of tokens is near-duplicates, totaling approximately 84 million tokens. The most common type of near-duplicates were news articles published both online and in printed newspapers. The stop word ratio filter flagged a substantial number of documents with a high proportion of non-running text. Although such documents represent a small proportion of the corpus, we believe their removal should result in measurably improved training stability. Overall, while a substantial number of documents were removed, most were simply unsuitable for pre-training rather than truly low-quality or noisy text.

Table 3 shows the number of documents that were normalized or altered, excluding those discarded entirely. Aside from boilerplate removal, these categories proved to be quite rare. Nevertheless, since we prioritized high accuracy with minimal false positives, we expect that even these infrequent corrections will help prevent the model from wasting capacity on malformed or noisy text.

These results strongly suggest that text quality in the IGC is generally very high, as expected for a curated corpus, but also confirm that a number of documents contain pathological text sequences that can negatively impact training stability. In practice, we observed that training examples consisting primarily of foreign-language text or non-running text sometimes caused spikes in gradient norms during pre-training, leading to increased training and validation losses, as described by Walsh et al. (2025). Although there is a subjective element to decisions about what text should be normalized or discarded, monitoring training stability can provide empirical grounding for such choices.

That said, filtering certain types of text in favor of training stability can come at a cost. For a monolingual model, it would be unrealistic to expect any benefits from cross-lingual transfer by retaining the small amount of foreign-language text present in

[2] https://repository.clarin.is/repository/xmlui/handle/20.500.12537/381

[3] https://huggingface.co/datasets/arnastofnun/IGC-2024-filtered-1

[4] https://repository.clarin.is/repository/xmlui/handle/20.500.12537/382

Category	Documents	Tokens
Short documents	348,841	12,777,006
Near-duplicates	215,553	83,956,512
Stop word ratio	175,923	38,791,290
Internal duplication	41,737	27,445,929
Code	18,382	7,552,647
Character encoding errors	7,047	1,321,170
Phrase filtering	7,039	1,142,784
OCR errors	6,799	3,745,532
Old documents	2,240	17,772,286
Total (unique)	**752,784**	**183,021,842**

Table 2: Number of documents and tokens for each category discarded in the filtering pipeline. Near-duplicates do not include the canonical document in each cluster.

Category	Documents
Boilerplate text	429,108
Escaped elements	10,996
Character normalization	3,212
Whitespace normalization	480
Total (unique)	**443,298**

Table 3: Number of documents normalized or otherwise altered by category, not including discarded documents.

the IGC. However, filtering out non-running text risks discarding meaningful information. For example, tables convey a great deal of structured content, but have historically been difficult to represent in plain text format. An alternative would be to encode such content in a format better suited for subword tokenizers and language models, such as JSON or Markdown. However, this would need to be performed at the text extraction stage, prior to filtering, and we therefore leave it to future work. Until then, discarding such text in favor of improved training stability remains the most practical option.

While a great deal of foreign-language text can be filtered using simple heuristics like stop word ratios, language classifiers offer a much more robust approach in principle. Documents that are primarily in other languages are excluded from the IGC, but foreign text often appears within Icelandic documents, sometimes as lengthy excerpts. For future versions of the pipeline, we plan to experiment with paragraph-level language identification. Fedorova et al. (2026) compare GlotLID (Kargaran et al., 2023) and two versions of OpenLID (Burchell et al., 2023), finding F_1 scores in the high 0.90s on hard evaluation sets and very close to 1 for other evaluation sets when disambiguating between Scandinavian languages. While they do not evaluate Icelandic, Burchell et al. (2023) report an F_1 score of 1.0 for Icelandic in their results for the first version of OpenLID. Incorporating a language identifier based on one of these recent tools could therefore serve as a valuable addition to the pipeline.

7. Conclusions and Future Work

We have presented a new JSONL distribution of the IGC, in a format specifically optimized for LLM training. We also described our filtering and deduplication pipeline and demonstrated that even high-quality corpora can contain segments that are flawed and unwanted for most NLP tasks. This included repetitive templates, code snippets, and tabular data or other non-running text that can trigger instabilities when training LLMs.

To safeguard against such issues we used targeted, mostly rule-based filtering. By removing approximately 752,000 problematic or redundant documents, we provide a more robust foundation for stable training without sacrificing the richness and linguistic diversity that makes the IGC unique among Icelandic text corpora. Furthermore, by distributing the corpus in a machine-actionable JSONL format with clear training and validation splits, we lower the barrier to entry for researchers and developers working on Icelandic AI.

Future iterations of this work will focus on two primary directions. First, rather than discarding non-running text such as tables or lists, we aim to develop heuristics to reformat these structures where possible into more useful representations (e.g., Markdown or structured JSON). This would preserve information that is currently lost during text extraction. Second, we plan to integrate state-of-the-art language identifiers, such as OpenLID, at the paragraph level to better handle code-switching and foreign-language excerpts within Icelandic documents. Through these continued refinements, we aim to ensure the IGC remains the definitive resource for the next generation of Icelandic language technology.

Acknowledgements

This work was funded by the Icelandic Strategic Research and Development Programme for Language Technology 2025, grant no. 250251-5301.

8. Bibliographical References

Starkaður Barkarson, Steinþór Steingímsson, and Hildur Hafsteinsdóttir. 2022. Evolving Large Text Corpora: Four Versions of the Icelandic Gigaword Corpus. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2371–2381, Marseille, France. European Language Resources Association.

Kristín Bjarnadóttir. 2012. The Database of Modern Icelandic Inflection. In *LREC 2012 Proceedings: Proceedings of "Language Technology for Normalization of Less-Resourced Languages", SaLTMiL 8 – AfLaT*, pages 67–72.

Kristín Bjarnadóttir, Kristín Ingibjörg Hlynsdóttir, and Steinþór Steingrímsson. 2019. DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 146–154, Turku, Finland. Linköping University Electronic Press.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An Open Dataset and Model for Language Identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Gavin Burnage and Dominic Dunlop. 1992. Encoding the British National Corpus. In Jan Aarts, Pieter de Haan, and Nelleke Oostdijk, editors, *English Language Corpora: Design, Analysis and Exploitation*, volume 10 of *Language and Computers*, pages 79–95. Rodopi, Amsterdam.

Jón Friðrik Daðason. 2025. *Language Representation Models for Low- and Medium-Resource Languages*. Ph.D. thesis, Reykjavik University.

Jón Friðrik Daðason and Hrafn Loftsson. 2022. Pretraining and Evaluating Transformer-based Language Models for Icelandic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7386–7391, Marseille, France. European Language Resources Association.

Jón Friðrik Daðason and Hrafn Loftsson. 2024. Unsupervised Outlier Detection for Language-Independent Text Quality Filtering. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 383–393, Torino, Italia. ELRA and ICCL.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michal Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2023. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57:415–448.

Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, et al. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

Mariia Fedorova, Nikolay Arefyev, Maja Buljan, Jindřich Helcl, Stephan Oepen, Egil Rønningstad, and Yves Scherrer. 2026. OpenLID-v3: Improving the Precision of Closely Related Language Identification – An Experience Report.

Atli Jasonarson and Steinthor Steingrimsson. 2025. AMI at WMT25 General Translation Task: How Low Can We Go? Finetuning Lightweight Llama Models for Low Resource Machine Translation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 695–704, Suzhou, China. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language Identification for Low-Resource Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Svetla Koeva, Diana Blagoeva, and Siya Kolkovska. 2010. Bulgarian National Corpus Project. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani,

Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Per Kummervold, Freddy Wetjen, and Javier de la Rosa. 2022. The Norwegian Colossal Corpus: A Text Corpus for Training Large Norwegian Language Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3852–3860, Marseille, France. European Language Resources Association.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling Data-Constrained Language Models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. *arXiv preprint arXiv:2406.17557*.

Adam Przepiórkowski, Rafał L. Górski, Barbara Lewandowska-Tomaszyk, and Marek Łaziński. 2008. Towards the National Corpus of Polish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2022. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Pétur Orri Ragnarson, Haukur Jónsson, and Vilhjalmur Thorsteinsson. 2021. Miðeind's WMT 2021 Submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 136–139, Online. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur Thorsteinsson, and Hafsteinn Einarsson. 2022. A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

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.

Leon Strømberg-Derczynski, Manuel Ciosici, Rebekah Baglini, Morten H. Christiansen, Jacob Aarup Dalsgaard, Riccardo Fusaroli, Peter Juel Henrichsen, Rasmus Hvingelby, Andreas Kirkedal, Alex Speed Kjeldsen, Claus Ladefoged, Finn Årup Nielsen, Jens Madsen, Malte Lau Petersen, Jonathan Hvithamar Rystrøm, and Daniel Varab. 2021. The Danish Gigaword Corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 413–421, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

TEI Consortium. 2026. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, 4.11.0 edition. Oxford, Providence, Charlottesville, Nancy. Last updated on 18th February 2026. Accessed February 2026.

Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, et al. 2025. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656*.

9. Language Resource References

Starkaður Barkarson and Steinþór Steingrímsson. 2024. Icelandic gigaword corpus (IGC-2024ext) - unannotated version. CLARIN-IS.

Starkaður Barkarson, Steinþór Steingrímsson, Þórdís Dröfn Andrésdóttir, Hildur Hafsteinsdóttir, Finnur Ágúst Ingimundarson, and Árni Davíð Magnússon. 2022. Icelandic gigaword corpus (IGC-2022) - unannotated version. CLARIN-IS.