The Norwegian Colossal Corpus: A Text Corpus for Training Large Norwegian Language Models

Per E Kummervold, Freddy Wetjen, Javier de la Rosa
National Library of Norway (NLN), Norway
per@capia.no, freddy.wetjen@nb.no, versae@nb.no

Abstract

Norwegian has been one of many languages lacking sufficient available text to train quality language models. In an attempt to bridge this gap, we introduce the Norwegian Colossal Corpus (NCC), which comprises 49GB of clean Norwegian textual data containing over 7B words. The NCC is composed of different and varied sources, ranging from books and newspapers to government documents and public reports, showcasing the various uses of the Norwegian language in society. The corpus contains mainly Norwegian Bokmål and Norwegian Nynorsk. Each document in the corpus is tagged with metadata that enables the creation of sub-corpora for specific needs. Its structure makes it easy to combine with large web archives that for licensing reasons could not be distributed together with the NCC. By releasing this corpus openly to the public, we hope to foster the creation of both better Norwegian language models and multilingual language models with support for Norwegian.

Keywords: natural language, text corpus, Norwegian

1. Introduction

Most modern approaches to natural language processing (NLP) require the training of language models on vast amounts of text. Ideally, these textual sources represent the language as it is used in a variety of settings. The power of unsupervised training on extremely large corpora was first demonstrated by Devlin et al. (2019) in their seminal work Bidirectional Encoder Representation from Transformers (BERT). Funded by Google, their model was trained on 16GB of uncompressed English text and comprised approximately 3.3B words. As the most popular language online, the size of the available corpora for English has steadily increased to over 800GB of text (Gao et al., 2021), a figure that is difficult to reach for lower-resource languages[1].

With only about 5 million native speakers, Norwegian is considered one of the lower-resource languages. The Norwegian language is a single spoken language with two official written languages: Bokmål and Nynorsk. Bokmål is the predominant written form and is the most widely used in everyday life in Norway. However, public entities are required by the Norwegian Language Act [Språklova] to produce at least a quarter of their publicly available documents in Nynorsk (Lovdata, 2022).

1.1 Norwegian Transformer Models

Until recently, the only way to utilize language models with languages such as Norwegian was via multilingual models. The multilingual version of BERT (mBERT) (Devlin et al., 2019) was trained solely on Wikipedia dumps constituting 104 different languages. Two of those languages were Norwegian Bokmål and Norwegian Nynorsk. While they did not specify the exact size of the Norwegian text in the corpus, our estimate suggests it was between 0.5GB and 1.0GB of text (70M and 140M words, respectively), where roughly 80% was Norwegian Bokmål. A similar estimate was made by Wu et al. (2020). Evaluated on a Named Entity Recognition (NER) task, the model achieved F1-scores of 83.8 and 85.6 for Bokmål and Nynorsk (Kummervold et al., 2021).

A Norwegian BERT-based model called NorBERT was recently released by Kutuzov et al. (2021). The model was trained on 3.7GB of text (540M words)[2] from the Norwegian Newspapers Corpus (Språkbanken, 2019), which was compiled and released in 2019 by the Norwegian Language Bank, a unit of the National Library of Norway (NLN). Evaluated on the same NER test set as the mBERT model above, the NorBERT model achieved 89.9 for Norwegian Bokmål and 86.1 for Norwegian Nynorsk.

Meanwhile, Kummervold et al. (2021) used a combination of data from NLN and external web resources to compile a corpus of 109GB (18B words) to train a Norwegian BERT-based model named NB-BERT. The model improved the state-of-the-art on the NER task to 91.2 for Norwegian Bokmål and 88.9 for Norwegian Nynorsk, showing that increasing the corpus size could improve Norwegian transformer models. Unfortunately, due primarily to legal reasons, the corpus used could not be shared publicly.

1.2 The National Library of Norway

The main role of the NLN is to preserve and provide access to all information published in Norway. The texts in their collections span hundreds of years and exhibit varied uses in society. A large amount of these texts are now digitally available, partly based on digitization and partly on born-digital documents obtained via the current legal deposit agreement. All kinds of historical written materials can be found in the collections, although we discovered that the most relevant resources for building an appropriate corpus for NLP are books, magazines, journals, and newspapers (see Table 1).

In this work, we introduce the Norwegian Colossal Corpus (NCC), a corpus made up of approximately 49GB (7B words) of text, mostly in Norwegian Bokmål and Norwegian Nynorsk. Our goal was to collect all publicly

[1] Over 1 billion people consider English their first or second language (Eberhard et al., 2022).

[2] In their paper, the authors stated that the online newspaper corpus was 4.7GB (1.2B words).

available and shareable text resources in Norway to make it possible for everyone to train large Norwegian language models, and to enable multilingual models to integrate a representative part of the Norwegian language.

2. Data Sources

The NCC is a collection of multiple heterogeneous data sources. All the work in the preparation of the dataset and all the software produced are licensed under the terms of a CC BY-SA 3.0 (2021) license. However, the individual corpora are under different licenses. Table 1 presents an overview of some of the main characteristics of the sub-corpora and their associated licenses. The NCC can be simplified as follows: 1) books and newspapers from the NLN that are out of copyright; 2) public documents (governmental or otherwise); 3) online newspapers; and 4) Wikipedia. These categories are also reflected in the licenses they are published under.

2.1 Library Books and Newspapers

The NLN has had a large and well-established digitization program in place since 2006. This includes all kinds of printed materials, such as books, newspapers, journals, and other small prints. Most of the books included in the NCC are out of copyright material, and released under the terms of a CC0 1.0 (2021) license. Newspapers are subject to a special agreement between the National Library and the publishers, and are released under the CC BY-NC 2.0 (2021) license. Together, text from these sources account for 6.2 and 14.0 GB (860M and 2B words), respectively.

2.2 Public Documents

The Norwegian Copyright Act, called *Åndsverkloven*, ensures the creators' rights to the distribution of their work. However, the law has an important exception: §14 states that any material that is adopted, issued or published by public authorities as part of their public task is not protected by the act. This is a comprehensive regulation that encompasses both central and regional authorities. It includes all reports, laws, regulations, and statements, as well as their official translations. Typically, this would even span contracted work when it is prepared as part of the body's public activities.

This makes it uncomplicated to include large amounts of public documents in the NCC. The Norwegian government has also created a special license, namely, the Norwegian License for Open Government Data 2.0 (Norwegian Digitalisation Agency, 2021), which is recommended for use regarding the sharing of such data. For most practical uses, this license is similar to a CC BY-SA license.

2.2.1 LovData CD

Lovdata is a Norwegian foundation that publishes the legal information of Norway. A collection of their CDs and DVDs, last published in 2005, is also included in the NCC. It contains a comprehensive set of legal resources, including laws, court verdicts, and public reports. While the data themselves also fall under The Copyright Act §14, the collection of these data was protected for 15 years by rights established in §24 in the same act until 2020.

For privacy purposes, we excluded documents that were likely to contain sensitive personal information. For instance, we excluded all court verdicts. The total volume of this sub-corpus is 0.4GB of text (55M words).

2.2.2 Government Reports

The NCC also includes almost all reports, propositions, and notes created by the Norwegian government from 1995 to 2021. The documents are available in a native XHTML format and can be downloaded directly from the government's official API. The internal format allows the fine-grained control of the output, which enables the filtering of headings, footers, tables, and other HTML elements, resulting in a high-quality corpus with very few errors.

The Government Report sub-corpus contains 1.1GB of text (1.3B words).

2.2.3 Parliament Collections

The archives contain the work done in the parliament and the parliamentary committees. Most of this text has gone through OCR processing. Fortunately, the materials exhibited a high print quality, and the OCR processing was conducted with a particular focus on the appropriate set of parameters to ensure the proper recognition of texts. This resulted in an average OCR quality generally higher than that of the other OCR-derived sources, which allowed us to keep a larger part of this corpus (see Section 3.3).

The Parliament Collections sub-corpus contains 8.0GB of text (1.3B words).

2.2.4 Public Reports

Government institutions in Norway publish reports and investigations from Norwegian society. To aid in the sharing of this information, the NLN created a portal where similar institutions can upload and share reports and knowledge.

Text extraction for these reports in PDF format has been performed in an ongoing project. The sub-corpus contains a total of 3,365 documents. These documents were published and collected between 2015–2020. The project extracted text from the PDF files using the tool pdf2txt (Ubuntu, 2022). The results are being used in the NCC.

There was considerable overlap between this corpus and other public corpora; however, after deduplication, the unique content was 0.5GB of text (80M words).

2.2.5 Målfrid

The Målfrid Corpus (2021), provided and collected by the Norwegian Language Bank, is the result of a focused web crawl from public institutions' websites. These websites store public sector information from a wide range of domains and institutions, ranging from healthcare to taxation.

Corpus	License	Size	Words	Documents	Avg Words/Doc
Government Reports	(NLOD 2.0, 2021)	1.1 GB	155,318,754	4,648	33,416
Library Books	(CC0 1.0, 2021)	6.2 GB	861,465,907	24,253	35,519
Library Newspapers	(CC BY-NC 2.0, 2021)	14.0 GB	2,019,172,625	10,096,424	199
LovData CD	(NLOD 2.0, 2021)	0.4 GB	54,923,432	51,920	1,057
Målfrid Collection	(NLOD 2.0, 2021)	14.0 GB	1,905,481,776	6,735,367	282
Newspapers Online	(CC BY-NC 2.0, 2021)	3.7 GB	541,481,947	3,695,943	146
Parliament Collections	(NLOD 2.0, 2021)	8.0 GB	1,301,766,124	9,528	136,625
Public Reports	(NLOD 2.0, 2021)	0.5 GB	80,064,396	3,365	23,793
Wikipedia	(CC BY-SA 3.0, 2021)	1.0 GB	140,992,663	681,973	206
Total		**48.9 GB**	**7,060,667,624**	**21,303,421**	**332**

Table 1: Sub-corpora in the Norwegian Colossal Corpus.

The main purpose of the crawl was to determine the language use distribution in various institutions. This work was conducted in close collaboration with the Language Council of Norway [Språkrådet]. The output from the crawl was Web ARChive (WARC, 2022) files, which can be processed and turned into differently derived language resources.

In the NCC, the subset containing all PDFs from the Målfrid Corpus was processed using a slightly modified version of MuPDF (Andersson, 2022). The software provided convenient Python bindings that were leveraged to build an ad-hoc extractor for our purposes. Using the stylistic and layout information contained in the PDFs, we focused the extraction on the main text body of each document, ignoring other elements such as tables, footnotes, and image captions. This approach produced very long and consistent documents with few errors and PDF-related artifacts.

We processed around 9.2M public PDF documents harvested from December 2020 to January 2021 belonging to 311 different institutions, yielding a training material of 14.0GB of text (1.9B words).

2.3 Online Newspapers

The category of Online Newspapers (Språkbanken, 2022) in the NCC is a newly processed version of the Norwegian Newspaper Corpus. The Norwegian Newspaper Corpus comprises an array of fairly large newspapers in Norway crawled from the web by the Norwegian Language Bank. Due to the increasing use of paywalls, the size for each year decreased for the corpus belonging in the period 1998 to 2019.

The Online Newspapers corpus is based on an agreement with the publishers allowing it to be distributed under the CC BY-NC 2.0 (2021) license. This is the non-commercial version of the Creative Commons license. This license is kept when included in the NCC.

2.4 Wikipedia

A dump from Wikipedia was downloaded on June 20, 2021. The text contains both Bokmål and Nynorsk This is marked in the corpus in the tag “doc_type”.

This corpus was previously distributed by Wikipedia under the CC BY-SA 3.0 license and accounts for 1GB of text (141M words).

2.5 Excluded Sources

Common Crawl (2022) is a non-profit organization that has been collecting data from the web and providing these archives to the public since 2011. Common Crawl-based datasets are popular for training transformer models and are the basis for the enormous 800GB Pile dataset (Gao, 2020), among others.

There are extracted Norwegian datasets that are also based on Common Crawl. The Open Super-large Crawled Aggregated coRpus (OSCAR) (Suárez et al., 2019) contains 4.7GB (800M words) of Norwegian Bokmål and 54MB (9M words) of Norwegian Nynorsk.

Using a cleaned version of Common Crawl, Google compiled a multilingual version of their English colossal corpus, called MC4 (2022), for training their mT5 model (Xue et al., 2020). The Norwegian part of that dataset is roughly 94GB (14B words).

Both OSCAR and the MC4 datasets have been made available on Hugging Face (2022). Unfortunately, their respective licenses do not allow for redistribution within the NCC. To overcome this limitation, we are releasing scripts for the preparation, cleaning, deduplication, and formatting of these datasets, so they can be interleaved

with the NCC. By combining NCC with OSCAR and MC4, it should be possible to create a deduplicated Norwegian corpus with over 100GB of text (15B words).

3. Processing the Dataset

As illustrated in Figure 1, the different sub-corpora were processed uniformly. We developed our data pipeline to process input source files and produce a dataset ready for training language models.

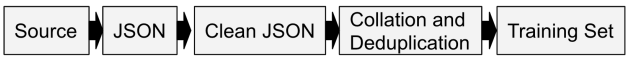

Figure 1: Processing pipeline

3.1 Source Files

The source files in the NCC are in multiple formats, ranging from the XML-based METS/ALTO formats to HTML, JSON, and various plain-text formats. In the first step of the processing pipeline, these files are unpacked when necessary but otherwise kept in their original formats.

For the OCR-derived sources there is a two-step pipeline. First, exact digital copies of the printed sources are created in JPEG 2000 format (2022). Then, these digital images undergo optical character recognition (OCR) and structure analyses. The resulting structure of the object and its actual textual content are encoded using the industry standard Metadata Encoding and Transmission Standard (METS)/Analyzed Layout and Text Object (ALTO) formats (Library of Congress, 2022).

Both OCR software and hardware have been subjected to general technical improvements over the years. Thus, the confidence values reported by these software applications also tended to change over time. After running a manual qualitative assessment, we discovered a significant improvement in the OCR quality of documents scanned after 2009. Therefore, the books that were scanned and processed before this period had to undergo a new OCR pass using Tesseract version 4.0 (Ubuntu, 2022). After this second pass, we further filtered out OCR errors by deleting any document or paragraph wherein the OCR software reported a confidence level below 90%.

3.2 JSON Lines Files

All sub-corpora were then converted to a common format to simplify the processing further while keeping all the relevant information from the sources. We chose JSON Lines (2022), a newline-delimited JSON format that enables processing both from the command line and programmatically. In this format, each document is a valid self-contained JSON object that may contain anything from a one-line notice in a newspaper to several hundred-page-long government reports.

While all JSON Lines objects share a few common keys, such as “id,” “doc_type,” and “paragraphs,” the rest are dependent on the metadata available in the original source. For instance, non-OCR documents do not need to have word confidence information. Since we strive to keep as much of the information from the original source document as possible, the exact JSON keys will vary from source to source. No documents are deleted on this level, even if we know that the quality might be too low to include in the final corpus. Appendix 1 presents some examples of the schema used within different JSON objects.

3.3 Clean JSON Lines Files

The most computationally intensive part of the pipeline was the standardization and cleaning of the different sub-corpora. We developed a set of parametric cleaning rules that could be applied in sequence to the JSON Lines files. The default parameters of these rules are very idiosyncratic of our source material and dependent on the very nature of each sub-corpora. For example, one rule might look for extremely long words and remove them since this is typically an unwanted OCR artifact. However, digitally born documents may contain long URLs or hashtags that meet the condition but must be kept in the document. Thus, it might make sense to filter out paragraphs containing extremely long words only for OCRed documents.

Moreover, for OCR-derived documents, we calculated the average word confidence at both the paragraph and document levels and filtered out those below 90%. Other cleaning rules relate to UTF-8 character encoding normalization, the minimum length necessary for a document to be kept, ensuring that sentences were not cut-off at the end of page breaks, or filtering out email addresses and personal identification numbers and replacing them with generic forms. More details on the various cleaning procedures are described in Appendix 2.

3.4 Collation and Deduplication

We also computed an MD5 hash for each paragraph and used that information for deduplicating the entire corpus at the paragraph level across all sub-corpora. At the same time, all redundant metadata in the JSON Lines files were deleted. After deduplication, the paragraphs were merged and analyzed with FastText (Joulin et al., 2016) to detect and annotate the main language in which they are written. That information was also added to the metadata of the document.

3.5 Training Dataset

The last step in the pipeline was converting the deduplicated and cleaned files into the distribution JSON Lines format. At this point, merged paragraphs are kept together as the textual value of a document, along with its other associated metadata. The corpus is distributed as a single massive JSON Lines file containing 21M documents. However, for practical reasons, it is also available as compressed 1GB shards that may be consumed as a streaming Hugging Face Dataset. The final format schema is shown in Table 2.

With JSON Lines as the base format, creating sub-corpora directly from the NCC is a trivial task. Appendix 3 illustrates a few examples on how to create a corpus that

only contains Norwegian Nynorsk, and another corpus only from sources within the 20th century.

Key	Type	Description
id	String	Source unique identifier
doc_type	String	Describing the type of media text extracted from books, newspapers, etc.
publish_year	Integer	Year the text was published
lang_fasttext	String	Language of text as identified by FastText
lang_fasttext_conf	String	Confidence calculated by FastText
text	String	The complete utf-8 document; if longer than 1M characters, it is split

Table 2: Final dataset format

4. Conclusion

Norwegian is only spoken by 5 million people, a small number when compared to more than 1 billion people who speak English. Nevertheless, the NCC is roughly three times the size of the English corpus used to train the original BERT.

The creation of such a large Norwegian corpus is underpinned by several notable premises. First, the legal deposit law and the ongoing process of digitizing large amounts of documents for long-term preservation and access, put the National Library of Norway in a practical position to actually be able to collect these text resources. The second premise is related to the distribution of the corpus. For out-of-copyright materials, this does not pose a problem. However, these sources are often in an older writing style. What makes the corpus colossal is that the Norwegian Copyright Act makes it easy to redistribute any text from a public body's activity, and an agreement with newspaper publishers allows the inclusion of large amounts of newspapers.

In addition to building a large corpus, we have also strived to make the corpus as heterogeneous as possible. While it can be challenging to measure to what extent corpus heterogeneity may impact large language models, it is our understanding that having texts with both different writing styles and different contexts will be beneficial. We have also included a significant amount of metadata. It is our hope that researchers will utilize this to investigate the effect of heterogeneity on language model quality.

The corpus contains both Norwegian Nynorsk and Norwegian Bokmål. Current research indicates that low-resource languages, such as Nynorsk, will benefit from being trained together with semantically similar higher-resource languages (Pires et al., 2019).

To encourage alternative uses of the corpus, we did include metadata like language, document type, and publishing year. This allows for the creation of, for instance, a Norwegian Nynorsk only corpus. It also allows for combining several of these metatags for creating even more specialized corpora. While the NCC was created with current transformer models in mind, we hope the corpus will be used for purposes beyond our expectations.

5. Future Work

While the state-of-the-art NB-BERT language model introduced by Kummervold et al. (2021) included in its training set an early version of the Norwegian Colossal Corpus presented in this work, we still lack a rigorous evaluation of this newer iteration of the corpus. Ablation studies with NCC combined with other Internet sources such as OSCAR or mC4 would also be welcomed. In the same sense, one possible interesting avenue for research is related to the evolution of the Norwegian language itself. The out-of-copyright materials included in NCC use an older writing style when compared to contemporary Norwegian. More research needs to be done to assess the effect on downstream tasks. Appendix 3 explains how to leverage the corpus for setting up such experiments. Moreover, since the sources used to build NCC are in constant development, we plan to release periodical updates of the corpus as new materials become available.

6. Acknowledgements

We are grateful to Andre Kåsen and Svein Arne Brygfjeld for providing valuable feedback on this paper during the final phases. Our thanks also go to Olaus Bergstrøm for useful feedback on the copyright and licensing aspects of this paper. We also wanted to thank the reviewers for their great feedback and comments.

7. Bibliographical References

Andersson, T. (2022). MuPDF documentation. <https://www.mupdf.com/docs/> [Accessed 10 January 2022].

Common Crawl. (2022). <https://commoncrawl.org> [Accessed 14 January 2022]

Creative Commons. (2021). *CC0 1.0 Universal (CCO1.0) Public Domain Dedication*. <https://creativecommons.org/publicdomain/zero/1.0> [Accessed 10 January 2022].

Creative Commons. (2021). *Attribution-NonCommercial 2.0 Generic (CC BY-NC 2.0)*. <https://creativecommons.org/licenses/by-nc/2.0/> [Accessed 10 January 2022].

Creative Commons. (2021). *Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0)*. <https://creativecommons.org/licenses/by-sa/3.0/> [Accessed 10 January 2022].

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In

Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Eberhard, D. M., Simons, G. F, & Fennig, C. D. (eds.). (2022). Ethnologue: Languages of the World. Twenty-fifth edition. Dallas, Texas: SIL International. <http://www.ethnologue.com> [Accessed 25 April 2022].

Gao, L., Biderman, S.R., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2021). *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*. ArXiv:abs/2101.00027.

Hugging Face. (2022). *Datasets*. <https://huggingface.co/datasets/> [Accessed 14 January 2022].

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). *Fasttext. zip: Compressing text classification models*. arXiv preprint arXiv:1612.03651.

JPEG 2000. (2022). <https://jpeg.org/jpeg2000/> [Accessed 14 January 2022].

JSON Lines. (2022). <https://jsonlines.org/> [Accessed 14 January 2022].

Kummervold, P. E., De la Rosa, J., Wetjen, F., & Brygfjeld, S. A. (2021). Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Kutuzov, A., Barnes, J., Velldal, E., Øvrelid, L., & Oepen, S. (2021). Large-Scale Contextualised Language Modelling for Norwegian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 30–40, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Library of Congress. (2022). *ALTO: Technical Metadata for Layout and Text Objects*. <https://www.loc.gov/standards/alto/> [Accessed 14 January 2022].

Library of Congress. (2022). *Metadata Encoding and Transmission Standard (METS)*. <http://www.loc.gov/standards/mets/> [Accessed 14 January 2022].

Lovdata. (2018). *The Copyright Act.* Åndsverksloven. <https://lovdata.no/dokument/NL/lov/2018-06-15-40> [Accessed 14 January 2022].

Lovdata. (2022). Norwegian *Language Act*. Språklova. <https://lovdata.no/dokument/NL/lov/2021-05-21-4> [Accessed 14 January 2022].

Målfrid Corpus. (2021). <https://www.nb.no/sprakbanken/en/resource-catalogue/ oai-nb-no-sbr-69/> [Accessed 14 January 2022].

National Library of Norway. (2020). *Digitising at the National Library*. <https://www.nb.no/en/digitizing-at-the-national-library/> [Accessed 14 January 2022].

Norwegian Digitalisation Agency. (2021). *Norwegian License for Open Government Data (NLOD) 2.0*. <https://data.norge.no/nlod/en/2.0/> [Accessed 14 January 2022].

Pires, T., Schlinger, E., & Garrette, D. (2019). How Multilingual is Multilingual BERT?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Språkbanken. (2022). Norsk aviskorpus <https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-4/> [Accessed 14 January 2022].

Suárez, P. J. O., Sagot, B., & Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *The 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

TensorFlow. (2022). MC4 TensorFlow Datasets. <https://www.tensorflow.org/datasets/catalog/c4> [Accessed 14 January 2022].

Ubuntu. (2022). Pdf2txt software package. Software Centre | Ubuntu. <https://ubuntu.com/blog/tag/ubuntu-software-center#developer> [Accessed 14 January 2022].

Ubuntu. (2022). Tesseract OCR version 4.0. Software Centre | Ubuntu. <https://ubuntu.com/blog/tag/ubuntu-software-center#developer> [Accessed 14 January 2022].

WARC (2022), Web ARChive file format. <https://www.loc.gov/preservation/digital/formats/fdd/fdd000236.shtml> [Accessed 14 January 2022].

Wu, S. & Dredze, M. (2020). Are All Languages Created Equal in Multilingual BERT?. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Baruna, A. & Raffel, C. (2020). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Appendix 1: JSON Lines Format

This JSON Lines format is used when the original sources are converted. These are two examples of the format used.

Library Books

Key		Example	Description
id		digibok_2006080900001	The original unique ID from the library.
doc_type		book	Document type.
ocr_date		20191224	Date for scanning in the format yyyy/mm/dd. Set to N/A if not in the mods post.
publish_date		20190101	Date for publication. For books, this is set to 0101 for the publication year. Set to N/A if not available.
language_detected		nob	Three-letter language code. *nob* for Bokmål and *nno* for Nynorsk. Only reported for books in METS/ALTO.
tesseract_version		4.1.1	Tesseract version if the program is used for OCR. If not, this field is dropped.
abbyy_version		8.1	Abbyy version, as reported in METS/ALTO.
document_word_confidence		0.9	Float 0-1. Average calculated from word confidence in META/ALTO while processing.
page		1	Integer. Page number (taken from METS/ALTO) if documents are divided into one document per page.
paragraphs			
	paragraphs_id	1	Integer. Starts with 0. Calculated during processing.
	block	1	Integer. Block number reported by METS/ALTO.
	confidence	0.90	Float. Paragraph word confidence.
	text	This field contains the actual text of the corpus	UTF-8 encoded text.

External Files

Key		Example	Description
id		wikipedia_11111	Unique reference to the exact document.
doc_type		wikipedia_download_nno	The document type.
language_detected		nob	Three-letter language code if available. Dropped if not.
paragraphs			
	paragraphs_id	1	Integer. Starts with 0. Calculated during processing.
	text	This field contains the actual text of the corpus	UTF-8 encoded text.

Appendix 2: Cleaning Rules

There are custom defaults for the various document types. These are the default settings.

Key	Default	Description
min_alphawords_paragraph	0	Sets a minimum number of words in a paragraph with only letters [a-Å]. Typically used in OCR text.
min_length_article	20	Minimum number of characters in an article.
min_words_paragraph	20	Minimum number of words in a paragraph.
max_word_length_paragraph	1000	Maximum word length in a paragraph. Typically used in OCR text, where extremely long words are sometimes encountered. Might also be triggered by URLs.
remove_control_characters	true	Removes control characters.
standardize_punctuation	true	Standardizes the punctuation.
replace_usernames_tweets	false	Replaces usernames in tweets: i.e., @twitteruser.
replace_urls	false	Replaces URLs.
replace_email_addresses	false	Replaces email addresses with placeholders.
fix_unicode	true	Fixes unicode errors.
normalise_unicode	true	Normalizes the unicode.
min_ocr_date	20090101	Minimum OCR date. No effect if “assume_late_missing_dates” is true and the value is not explicitly set.
min_publish_date	18140517	Minimum publish date. No effect if “assume_late_missing_dates” is true and the value is not explicitly set.
min_document_word_confidence	0.9	Minimum average word confidence in a document. If this is not from an OCR source, the default value is set to 1.0.
min_confidence_paragraph	0.9	Minimum average word confidence in a paragraph. If this is not from an OCR source, the default value is set to 1.0.
remove_non_terminated_paragraphs	true	Headings usually do not have punctuation marks at the end. This removes paragraphs without a punctuation at the end.
truncate_last_valid_sentence	true	In many books, it is hard to concatenate pages. This means some sentences are broken in half. This removes the last sentence.
minimise_jsonl	true	Minimizes the size of the JSON file: i.e., removes metadata that are not necessary later.
assume_late_missing_dates	true	If there are missing dates, assume that the missing dates are today. Valid for both publish_date and ocr_date.
drop_paragraphs_with_encoding_errors	true	Drops all paragraphs with encoding errors.
drop_paragraphs_with_curly_brackets	true	Drops paragraphs with curly {brackets}. This effectively removes JavaScript from a lot of web documents.

Appendix 3: Creating Sub-Corpora

Since the data format is based on JSON Lines that guarantees one single document on each line, it is possible to use standard command line tools to create sub corpora.

The simplest way is to use “cat” and “grep” for extracting for instance one of the document types. The line below will simply output all the lines with the document type “maalfrid_ssb”:

```
$ cat corpus.json | grep "\doc_type\": \"maalfrid_ssb\"" > sub_corpus.json
```

For doing more advanced filtering, it is possible to use “jq”. Here is an example of creating a corpus with documents published after 1970:

```
$ cat a.json | jq 'select(.publish_year >= 1970)' | jq -s|jq -c .[] > final.json
```

One example of extracting documents that has a FastText language confidence above 0.8. This confidence usually indicates that the text is relatively clean:

```
$ cat corpus.json | jq 'select(.lang_fasttext_conf|tonumber >= 0.8)' | jq -s|jq -c .[] > sub_corpus.json
```

Another example where only long documents are extracted:

```
$ cat corpus.json | jq 'select(.text|length >= 1000)' | jq -s|jq -c .[] > sub_corpus.json
```