Finnish Internet Parsebank

Juhani Luotolahti (✉ mjluot@utu.fi)
University of Turku

Jenna Kanerva
University of Turku

Jouni Luoma
University of Turku

Valtteri Skantsi
University of Oulu

Sampo Pyysalo
University of Turku

Veronika Laippala
University of Turku

Filip Ginter
University of Turku

Research Article

Keywords: web corpora, NLP, Finnish, syntax, NER, text register, Common Crawl, Internet crawl

Posted Date: July 14th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3138153/v1

Additional Declarations: No competing interests reported.

Finnish Internet Parsebank

Juhani Luotolahti[1*], Jenna Kanerva[1], Jouni Luoma[1], Valtteri Skantsi[3], Sampo Pyysalo[1], Veronika Laippala[2], Filip Ginter[1]

[1]TurkuNLP, Department of Computing, University of Turku, Finland.
[2]TurkuNLP, School of Languages and Translation Studies, University of Turku, Finland.
[3]TurkuNLP, Faculty of Humanities, University of Oulu, Finland.

*Corresponding author(s). E-mail(s): mjluot@utu.fi;
Contributing authors: jmnybl@utu.fi; jouni.a.luoma@utu.fi; valtteri.skantsi@oulu.fi; sampo.pyysalo@gmail.com; mavela@utu.fi; figint@utu.fi;

Abstract

We present a Finnish web corpus with multiple text sources and rich additional annotations. The corpus is based in large parts on a dedicated Internet crawl, supplementing data from the Common Crawl initiative and the Finnish Wikipedia. The size of the corpus is 6.2 billion tokens from 9.5 million source documents. The text is enriched with morphological analyses, word lemmas, dependency trees, named entities and text register (genre) identification. Paragraph-level scores of an n-gram language model, as well as paragraph duplication rate in each document are provided, allowing for further filtering of the dataset by the end user. Thanks to changes in the 2023 Finnish copyright legislation, the corpus is openly available for research purposes, and can also be accessed through the NoSketchEngine concordance tool and the dep_search dependency tree query tool, all at https://turkunlp.org/finnish_nlp.html.

Keywords: web corpora, NLP, Finnish, syntax, NER, text register, Common Crawl, Internet crawl

1 Introduction

Large text corpora play an important role in natural language processing and computational linguistics. As NLP methodology recently came to rely on pre-trained language models, the importance of large high-quality textual corpora became even more pronounced. It can be argued, that a text corpus of sufficient quality is nowadays a necessary prerequisite for the development of state-of-the-art NLP tools and applications in any given language. In addition to language model pre-training and tool development, large text corpora have an important place in linguistic research in numerous analytical tasks that require corpus-based statistics of various kinds.

The world wide web provides a rich and readily available source of text, and has therefore been utilized by many to create multilingual web corpora. Text extracted from the Internet is varied and can be obtained in large amounts, nevertheless, the quality of raw text crawled from the Internet naturally presents a major challenge. Among other issues, Internet text for instance contains non-textual noise like HTML tags and encoding errors, machine generated material, as well as boilerplate and irrelevant text from the web pages, all of which affect the final corpus quality. Since the use of text corpora is broad and statistical methods heavily depend on the source material, its quality is naturally important.

Especially recently, web corpora are often collected with a single task in mind: large language model pre-training. Without additional annotations and quality control, the value of such raw text corpora for other applications and linguistic research is not fully materialized. Such additional annotations may include, for instance, morphosyntactic analysis, named entity annotation, a language model -based text quality estimate, text registers (genres), and other similar metadata. These annotations can provide additional context, disambiguation, and semantic information that would otherwise be difficult to infer from the raw text alone.

While recent efforts have resulted in massive corpora for languages with a large number of speakers and therefore a major web presence, languages such as Finnish constitute only a small proportion of text in these massive web-based resources and need a more targeted effort. In addition to being underrepresented in large multilingual web corpora, small languages can also be represented in a way that is biased or partial.

In this paper, we work towards addressing these issues for Finnish, by presenting a Finnish web corpus which is based on a custom crawl, extending publicly available crawl resources and focusing specifically on the Finnish Internet. Further, the corpus has several layers of automatically produced linguistic annotations, making it applicable also to tasks other than language model pre-training.

2 Related work

There is a large body of literature on web corpora based on a custom crawl from the Internet. The most notable examples include the COW (Jakubíček, Kilgarriff, Kovář, Rychlỳ, & Suchomel, 2013), Wacky (Schäfer, 2015), and TenTen (Baroni, Bernardini, Ferraresi, & Zanchetta, 2009) web corpus families. In our work, we build especially upon the TenTen corpora, in terms of using similar web crawling procedures, but we also enrich our corpus with additional annotation layers, as discussed in the introduction.

Recently, rather than executing own, dedicated web-crawl, many web corpora source their data solely from the Common Crawl resource[1], a very large, openly accessible web crawl maintained by the Common Crawl foundation. The Oscar corpus (Abadji, Suárez, Romary, & Sagot, 2021) is a Common Crawl -derived multilingual collection of web corpora, including a dataset of Finnish as well. Similarly, the Common Crawl data has been used to create the mc4 multilingual corpus (Goyal, Du, Ott, Anantharaman, & Conneau, 2021), popular especially for language modelling purposes. Finally, the Nordic Pile (Öhman et al., 2023) is a corpus of Nordic languages, created using parts of the mc4 corpus and other data sources like Wikipedia. It, however, does not include Finnish, in addition to not being publicly available for legal reasons.

In terms of additional annotations of the raw texts, the Oscar corpus offers language modelling-based annotation on content harmfulness and language detection scores, but no other annotation. Dependency parses are included in all COW corpora and many, but not all of the Wacky corpora. Both the COW and TenTen corpora provide POS-tags, as well as paragraph and document boundaries based on the source web pages. We summarize the various corpora with respect to their sizes and annotation layers in Table 1, including for comparison also the Finnish Parsebank introduced in this work.

In terms of data cleanup, the TenTen corpora perform a boilerplate removal and near duplicate removal on the collected text, but no further filtering steps are applied. The Wacky corpora in addition to boiler-plate removal and deduplication perform keyword-based content cleanup which removes text with low function word count as well as pornographic or otherwise undesirable text. The mc4 corpus uses cleanup based on character heuristics, as well as keyword based pornography text removal. The COW corpora employ cleanup based on character heuristics. Finally, the Nordic Pile corpus employs also character and sentence features for text cleanup. Oscar uses a heuristic method for cleaning the plain text.

Corpus	**Size**	**Size (Fin.)**	**Dep.**	**POS**	**Morph.**	**NER**	**Reg.**	**Lemma**
mc4	27.0TB	104.0GB	No	No	No	No	No	No
Oscar	6.4TB	41.0GB	No	No	No	No	Yes	No
Nordic-Pile	1.2TB	-	No	No	No	No	No	No
ukWac	13.0GB	-	No	Yes	No	No	No	Yes
enTenTen	13.0GB	-	No	Yes	No	No	Yes	Yes
Parsebank	44.8GB	44.8GB	Yes	Yes	Yes	Yes	Yes	Yes

Table 1 Comparison of the corpora discussed in Section 2 w.r.t. the total size, size of Finnish section if any, and additional annotation of dependency syntax (Dep.), POS, morphology tags (Morph.), named entities (NER), text registers (Reg.) and word lemmas.

[1] https://commoncrawl.org

3 Finnish Internet Parsebank

Next, we describe the corpus in terms of its sources, text cleanup process, and the annotation layers.

3.1 Data sources

Our corpus is based on three primary data sources: Finnish Wikipedia, Common Crawl, and a custom web-crawl.

We obtained a database dump of the Finnish Wikipedia and used the mwlib[2] tool to extract plain text from it. This yielded approximately 1.5 GB of high-quality plain text.

The Common Crawl dataset includes both plain text and raw HTML files, at the time without language metadata. We employed a language detection step using CLD3 as the language detector and MapReduce to download only the Finnish-language plain text from the Amazon cloud service that hosts Common Crawl. As shown in Table 2, this resulted in only a moderate amount of new data (3.2GB deduplicated text) on top of Wikipedia (1.5GB deduplicated text).

Consequently, we conducted a dedicated web crawl using the SpiderLing web crawler (Suchomel & Pomikálek, 2012). This web crawler is specifically designed for collecting monolingual plaintext web corpora. It comprises a web crawling engine, a trigram-based language detector, and a boilerplate remover called Justext, which is responsible for extracting plain text. Moreover, the crawler is lightweight and easy to run. The crawl was seeded with the list of all domain names in the `.fi` top-level domain, as well as the URLs of all Finnish text pages we gathered from Common Crawl in the previous step. The crawl was carried out between 2014 and 2016.

The final sizes of text obtained from the three sources are summarized in Table 2, which shows that the dedicated webcrawl constitutes by far the largest portion of the final corpus. Note that in the newer versions of Common Crawl, a considerably stronger emphasis is placed on multilingual coverage, and the benefit of a dedicated webcrawl might be smaller but very unlikely to vanish entirely.

3.2 Deduplication and Processing

We combined the three sources of plain text and performed a first round of coarse deduplication, removing full, exact duplicates from the corpus. This step reduced the amount of material to approximately 40% of the original. This step was implemented by a simple text hashing of the source documents. While some studies remove duplicated material entirely, we chose to keep one instance of each document.

For the next stage of deduplication, we used Onion (Pomikálek, 2011), a dedicated tool for fuzzy, paragraph-level deduplication. Since the corpus was too large for a single run of the software, we split the text into multiple parts and performed deduplication on pairs until the entire corpus was processed. The output of this process is a corpus with paragraphs marked as being either unique or duplicate-material. After identifying duplicates, we discarded all documents with more than 75% duplicate paragraphs and

[2] https://github.com/pediapress/mwlib/

divided the remaining text into buckets based on their maximum duplication rate. The buckets will be referred to as D-25, D-50, and D-75, where the number denotes the maximum allowed duplication rate in the document. Here duplication rate means the amount of text in the document that has a duplicate somewhere else in the corpus. The bucket D-25 has documents with less than 25% duplicate material, and represents the part of the corpus with the least amount of duplication, while the bucket D-50 has documents which contain more than 25% duplicate text, but less than 50%. The last bucket contains documents with more than 50% duplicate text, but less than 75%. The last bucket contains most of the forum posts in the corpus due to quotations in the forum threads introducing text duplication. This division of the corpus is inspired by a prior study of Baroni and Kilgarriff (2006) suggesting that web text with higher duplication rate is more likely to be problematic and of limited linguistic interest, such as warnings and copyright messages.

In addition to deduplication, we improved the quality of the corpus by addressing encoding errors, which are common in web-based text. We used the Python library `ftfy` (fix text for you) for this purpose, which is capable of correcting many types of commonly met encoding errors.

Table 2 shows the material sizes before and after deduplication, and Table 3 details the frequency of top-level domains in the corpus. The most interesting observation here is that focusing solely on the national `.fi` domain would lead to a substantial loss of coverage.

Source Material	Original size	Deduplicated size	Retained
Crawled Text	280.0 GB	55.8 GB	20%
Finnish Wikipedia	1.5 GB	1.3 GB	87%
CommonCrawl	5.0 GB	3.2 GB	65%

Table 2 Material sizes before and after deduplication.

TLD	Frequency
fi	61%
com	21%
net	7%
org	2%
others	9%

Table 3 Frequency of top-level domains in the corpus.

3.3 Cleanup Process

After deduplication and the crawler-based boilerplate text removal, we performed a heuristic cleanup process and also removed machine-translated or generated text from the corpus.

3.3.1 Character-based Heuristics

To filter out undesirable text, we utilized a simple heuristic based on the Unicode categories of the characters in a document. This filtering was applied at the document level, and documents were removed from the corpus if they did not meet the criteria.

The heuristic required that at least 65% of the characters in a document be Latin lowercase characters, with no more than 10% punctuation or numerals, and uppercase characters accounting for at most 15% of the characters. Additionally, no more than 30% of the characters could be non-Latin. We implemented this heuristic using the Python unicodedata library.

This cleanup stage resulted in the removal of approximately 24 million tokens from the corpus.

3.3.2 Removal of Machine-translated Content

Removal of machine-translated and machine-generated content is a typical step in construction of web-based corpora. To this end, we trained a dedicated classifier using the FinCORE dataset (Skantsi & Laippala, 2023) discussed in greater detail in Section 3.4.4. This step removed the most material out of all the filtering steps, resulting in the removal of more than a billion tokens, as many low-quality noisy documents were identified in this step, in addition to genuine machine translated content.

3.4 Annotations

Each document has a number of additional annotations and metadata. These include text-level metadata for document boundaries, headings, paragraphs[3] and identification of duplicate text blocks. Each document and each paragraph has associated language model perplexity scores for possible further filtering. Finally, each document is given a full dependency parse tree, including also word lemmas, POS-tags and morphological tags for each token, a named entity annotation and a register-label which is included in the comments.

3.4.1 Dependency Parsing

The corpus was parsed with the 2022 updated version of the Turku Neural Dependency Parser pipeline (Kanerva, Ginter, Miekka, Leino, & Salakoski, 2018; Kanerva, Ginter, & Salakoski, 2020), a state-of-the-art full dependency parser for Finnish. The pipeline carries out sentence segmentation and tokenization, part-of-speech and morphological tagging, dependency parsing in the Universal Dependencies scheme, and word lemmatization. The evaluation of the pipeline on UD treebank test set as well as a sample of the parsebank data is reported in Section 4.4.

[3] Paragraph boundary information is not available for the about 5% of documents originating directly from Common Crawl, since the WET files did not contain enough information to recover paragraphs reliably. It is available for the remaining 95% of the data.

3.4.2 Named Entity Recognition

Named entity recognition was performed using a FinBERT model (Virtanen et al., 2019) trained on the corpus introduced by Luoma, Chang, Ginter, and Pyysalo (2021). The corpus is annotated for named entities using Ontonotes (Hovy, Marcus, Palmer, Ramshaw, & Weischedel, 2006) conventions, and thus marks 18 different name and numeric entity types. The specific tagging approach was adapted from Luoma and Pyysalo (2020), where each input sample for a prediction contains a sentence from original data and as much of following sentences that fit in to the sample maximum length. If no more subsequent sentences are available, we use documentwise wrapping to fill the samples: after the end of the document, we fill with tokens from the beginning of the same document. In this approach, the same sentence will appear as part of multiple input samples with different amounts of context before and after it. We aggregate the predictions of each original sentence from different samples to get the final prediction for that sentence. We use the concatenation of the four last transformer layer outputs as input to the dense classification layer instead of just using the last transformer layer output as in the original method. In addition to assigning labels for tokens by taking the maximum of mean label probabilities, we use the Viterbi algorithm for correcting the output label sequences.

3.4.3 Language Model scoring

We score every document and every paragraph with two KenLM (Heafield, 2011) n-gram language models, a delexicalized model using POS-tags instead of words and a standard lexicalized language model, using words. The motivation of the delexicalized language model is to give meaningful perplexities to grammatical text regardless of the frequency of the tokens themselves, allowing the text to depart from the topics and word distribution of the model training data. The training data for these language models was the Finnish Wikipedia for the delexicalized model and the Finnish discussion board Suomi24 for the lexicalized model.

In a small-scale manual evaluation, we verified that higher perplexity for delexicalized models is generally assigned to colloquial language, while grammatically correct language receives lower perplexity, as would be expected. The perplexities of the lexicalized model in our tests generally correlates with how good or typical Finnish language a paragraph or a document is. The lexicalized model score is thus more useful for further filtering of the dataset, nevertheless the delexicalized model scores are included with the dataset as an additional available feature as well.

3.4.4 Registers

Register (genre) information was added using a register identification model trained on the FinCORE corpus (Skantsi & Laippala, 2023). This corpus was originally compiled from a random sample of the Parsebank documents that have been manually annotated for text register. The register taxonomy consists of eight main classes, illustrated in Table 4. In addition, the FinCORE dataset has the *Other* category consisting mainly of machine-translated / generated text and other texts not written by humans.

We used the FinBERT model (Virtanen et al., 2019) and the nine register categories in a multi-label setting, since a small proportion of the FinCORE documents are labeled with more than one main register category. The model fine-tuning was done following the best setting reported by Skantsi and Laippala (2023). Evaluation of the model on the FinCORE test data shows that the classifier achieves an F1-score of 82%. All documents in the *Other* register category were discarded from the Parsebank. The distribution of text w.r.t. registers in the corpus is detailed in Table 5.

Main register	Sub-register examples
Narrative	News report, news blog, sports report, personal blog
Informational persuasion	Description with intent to sell, News-opinion blog or editorial
Opinion	Opinion blog, review, Regligious blog or sermon
Informational description	Job description, description of a thing, encyclopedia article
How-to / instructions	Recipe
Interactive discussion	Discussion forum, question-answer forum
Lyrical	Poem, song
Spoke	Interview, formal speech
Others	Machine-translated or generated texts

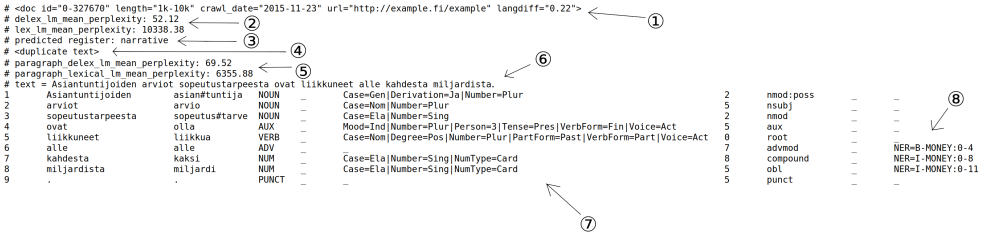
Table 4 Main text register classification in the FinCORE dataset together with examples of sub-registers for each main register. Note that the sub-register classification is not used in the Parsebank and is listed only for illustration.

Register	Frequency [tokens]	Frequency [documents]	Proportion [tokens]	Proportion [documents]
Narrative	2,662,035,312	4,513,080	42.9%	47.3%
Interactive discussion	1,581,173,892	1,796,202	25.5%	18.8%
Informational description	798,297,630	1,336,444	12.9%	14.0%
Opinion	724,785,194	765,754	11.7%	8.0%
Informational persuasion general	272,865,644	793,388	4.4%	8.3%
How-to/instructions	127,199,229	303,554	2.0%	3.2%
Spoken	30,703,143	32,175	0.5%	0.3%
Lyrical	2,735,291	6872	0.04%	0.1%

Table 5 Text register counts and proportion in the corpus.

3.5 Format of the corpus

The Parsebank is distributed in the CoNLL-U format with document- and paragraph-level metadata included as comments in the conllu files. The conllu format is a commonly used format for text with dependency parses and is widely supported by many tools. It is primarily designed to encode dependency trees, but has provisions also for both span-level and word-level metadata and annotations.



```
# <doc id="0-327670" length="1k-10k" crawl_date="2015-11-23" url="http://example.fi/example" langdiff="0.22">
# delex_lm_mean_perplexity: 52.12
# lex_lm_mean_perplexity: 10338.38
# predicted register: narrative
# <duplicate text>
# paragraph_delex_lm_mean_perplexity: 69.52
# paragraph_lexical_lm_mean_perplexity: 6355.88
# text = Asiantuntijoiden arviot sopeutustarpeesta ovat liikkuneet alle kahdesta miljardista.
1       Asiantuntijoiden        asian#tuntija   NOUN    _       Case=Gen|Derivation=Ja|Number=Plur                                      2       nmod:poss       _       _
2       arviot                  arvio           NOUN    _       Case=Nom|Number=Plur                                                    5       nsubj           _       _
3       sopeutustarpeesta       sopeutus#tarve  NOUN    _       Case=Ela|Number=Sing                                                    2       nmod            _       _
4       ovat                    olla            AUX     _       Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act         5       aux             _       _
5       liikkuneet              liikkua         VERB    _       Case=Nom|Degree=Pos|Number=Plur|PartForm=Past|VerbForm=Part|Voice=Act   0       root            _       _
6       alle                    alle            ADV     _       _                                                                       7       advmod          _       NER=B-MONEY:0-4
7       kahdesta                kaksi           NUM     _       Case=Ela|Number=Sing|NumType=Card                                       8       compound        _       NER=I-MONEY:0-8
8       miljardista             miljardi        NUM     _       Case=Ela|Number=Sing|NumType=Card                                       5       obl             _       NER=I-MONEY:0-11
9       .                       .               PUNCT   _       _                                                                       5       punct           _       _
```

Fig. 1 Format of the corpus. 1: The document tag, 2: Language Modelling scores, 3: Predicted Register, 4: Duplicate text and paragraph markers, 5: Language modelling scores for a paragraph, 6: The plaintext, 7: CoNLL-U dependency tree with lemmas, POS and morphological tags, 8: NER-labels

The format is illustrated in Figure 1. In the figure we see document borders being marked with a comment with a doc – tag. The tag contains the crawl date, if available, the URL and language detector score. Information contained in the tag differs between different sources of text. The language modelling scores and the register classification of the document are encoded as conllu comments. Paragraphs of text duplicated somewhere else in the corpus are marked with duplicate text - tags. For each paragraph, language modelling scores are included as well. The standard 10-column conllu data follows with the tokens, lemmas, POS tags, morphological tags, and dependency relations. The NER annotation is included as part of the conllu in the final column. The untokenized plain text of each sentence is provided as well.

4 Evaluation

To evaluate the resulting corpus, we investigate the statistics of the corpus, perform a manual error analysis on randomly sampled sentences and try to evaluate the reach of the resource by comparing it to Google search result statistics. We also evaluate the corpus annotation layers.

4.1 General Statistics

The size of the corpus is 6.2 billion tokens in 9.5 million documents, the distribution across the duplication rate buckets is detailed in Table 6. The sizes per source of the text are presented in table 7. We can see most of the plaintext comes from the web-crawl, followed by the Wikipedia and then the CommonCrawl.

Table 3 shows the most common Internet top-level-domains in the material. More than half of the material is found within the `.fi` - top-level-domain, the domain reserved for Finnish websites followed by more generic domains.

4.2 Manual Evaluation of Text Quality

To get an understanding of the quality of the text in the corpus, we performed a manual evaluation of general text quality. To this end, we sampled 400 random sentences from each of the three duplication buckets, totaling 1200 sentences of the corpus. These

Bucket	Tokens	Documents
D-25	3,868,557,423	6,493,510
D-50	1,319,882,022	1,842,626
D-75	1,011,355,890	1,211,495
All	6,199,795,335	9,547,999

Table 6 Token counts in the duplication rate buckets of the corpus. D-NN means that at most NN% of each document is formed by duplicated paragraphs. Smaller number indicates better quality of the documents in the respective bucket.

Bucket	Crawl Tokens	CC Tokens	Wikipedia Tokens
D-25	3,516,550,725	264,439,473	87,567,225
D-50	1,317,520,398	61,034	2,300,590
D-75	1,010,524,203	38,239	793,448
All	5,844,595,326	264,538,746	90,661,263

Table 7 Token counts in the duplication rate buckets of the corpus per data-source. D-NN means that at most NN% of each document is formed by duplicated paragraphs. Smaller number indicates better quality of the documents in the respective bucket. *CC* refers to Common Crawl.

sentences were evaluated to be either a complete, meaningful sentence in the Finnish language or not. The results of the evaluation are shown in Table 8. Overall more than 94% of the sampled sentences were good Finnish text across all three buckets.

We also inspected and categorized the sentences judged erroneous into five types of errors (in order of prevalence): leftover formatting, tokenization or sentence splitting errors, non-Finnish text, encoding errors and lists.

Tokenization and sentence splitting errors were the most common error making up 38 % of the found errors, they usually consisted of a sentence which was terminated too early. Leftover formatting was common in forum posts and usually was formatting used to identify quotations and style, this was the next most common type of error found in the sample making up 35 % of the errors found. Non-Finnish text made up 21 % of the errors. Encoding errors were found in 3% of the sampled sentences. Lists, sentences which were lists of some sort being interpreted as sentences by the process, for example a menu with its items concatenated is an example of this type of an error, made up 3 % of the sampled errors.

In a separare evaluation prior to removal of machine translated content from the corpus, roughly 10% of the sentences were deemed machine translated, but none were found after the removal. The lack of machine translated content and the rarity of encoding errors reflects our use of machine translated content removal and encoding error fixing. These results are summarized in Table 9.

Bucket	Proportion
D-25	95.0%
D-50	94.5%
D-75	93.3%

Table 8 Proportion of sentences without any issue in the test sample, per duplication bucket of the corpus.

Issue	Proportion	Count
Token-sentence-split	38%	26/68
Leftover formatting	35%	24/68
Language	21%	14/68
Encoding	3%	2/68
Lists	3%	2/68

Table 9 Error distribution among the text samples with an identified issue.

4.3 Coverage Estimation

We performed an estimate of the coverage of our corpus compared to the Google search index using the heuristic method proposed by Kilgarriff (2007). In this estimate, the counts of words in a de-duplicated corpus are compared to their Google hit counts, corrected for a possible difference in duplication rates between the corpus and the google index. The former rate is known, the latter is not known, which leads to a range, rather than a single estimate.

Following Kilgarriff (2007), we sampled a number of lowercase mid-frequency tokens (in our case 100) from the corpus and calculated their hits in the Google search engine. On average Google search returned 45 times the amount of hits per search word compared to their frequency in our deduplicated corpus. It could then be inferred that our corpus would have a coverage of approximately 2.2% of the Finnish Google search index, if our rate of text deduplication is identical to that of the deduplication in Google's search index. It could be also argued that since our deduplication preserves 20% of the crawled data (Table 2), our corpus consists of approximately 11% of the Google search index, if it has no deduplication.

Since we cannot know how exactly the data is indexed and deduplicated by the Google search engine, we can only conclude that our corpus coverage is most likely in the units of percent of the Google index. This is comparable to what has been reported in relation to other web-based corpora (Kilgarriff, 2007).

4.4 Evaluating the quality of morpho-syntactic analyses

In order to estimate the quality of morpho-syntactic analyses produced by the Turku neural parser on this dataset, we randomly sample 30 documents from the Parsebank and manually annotate these documents for token and sentence segmentation, part-of-speech and morphological tags, lemmas as well as dependency relations. This manually annotated sample is available through Universal Dependencies v.2.7 data release[4], and described in detail in Kanerva and Ginter (2022).

During the manual annotation, 5 documents were skipped due to them being manually determined as machine translated[5], and new documents were sampled to replace these. Each document was truncated after 25 sentences to avoid overly long documents biasing the evaluation towards particular topics.

The evaluation results are shown in Table 10, where the parsing accuracy on the Parsebank sample is compared to two other Finnish treebanks, Finnish-TDT and Finnish-PUD, both including manual UD annotation consistent with the Parsebank sample annotation. Finnish-TDT (Haverinen et al., 2014; Pyysalo, Kanerva, Missilä, Laippala, & Ginter, 2015) is a broad coverage, general Finnish treebank with 15,000 sentences in different genres such as blogs, fiction, grammar examples, legal text, news and Wikipedia articles, while Finnish-PUD (Zeman et al., 2017) is an external Finnish test set from the collection of parallel UD treebanks each including the same 1,000 sentences collected from Wikipedia and news articles. These sentences were first translated into Finnish and afterwards annotated according the UD annotation schema. The parsing model used in these experiments was trained on the Finnish-TDT corpus, the only one of these including also training and development sections.

As seen in Table 10, the scores are somewhat lower for the Parsebank than for the TDT and PUD treebanks, which are clean text and reflect the training data distribution of the parser. However, in all metrics, the results show good quality of the analyses, with morphological analyses (UPOS, UFeats, Lemmas columns in the table) being in the mid-to-high 90's range in terms of accuracy, and the parse trees (UAS, LAS) crossing 86% accuracy in terms of LAS. Token and word segmentation (Tokens, Words) is nearly 100% accurrate, but interestingly sentence segmentation shows a clear degradation compared to the standard-language UD treebanks. It is worth noting that the studies discussed in the following section used an earlier (pre-transformer) version of the parser pipeline, notably lower in accuracy compared to the analyses described here.

Treebank	Tokens	Sent.	Words	UPOS	UFeats	Lemmas	UAS	LAS
TDT	99.6	87.2	99.6	97.9	96.7	95.8	93.0	91.0
PUD	99.6	91.3	99.6	98.0	97.1	95.3	94.0	92.1
Parsebank	99.3	80.3	99.3	96.3	95.2	94.3	89.1	86.4

Table 10 Evaluation of the parsing accuracy on a sample of parsebank documents, and compared to two publicly available Finnish treebanks. All reported scores are percentage of accuracy of the appropriate parser output. UAS and LAS stand for unlabeled and labeled attachment score.

[4] Finnish-OOD (https://github.com/UniversalDependencies/UD_Finnish-OOD)

[5] Note that machine translation prediction and removal was skipped during document sampling to prevent the possibility of accidentally discarding extremely difficult documents falsely identified as machine translated or generated.

Label	Precision	Recall	FB1	Support
CARDINAL	90.48%	92.68%	91.57	42
DATE	92.59%	87.72%	90.09	108
EVENT	40.00%	80.00%	53.33	10
FAC	40.00%	66.67%	50.00	10
GPE	98.57%	94.52%	96.50	70
LANGUAGE	100.0%	83.33%	90.91	5
LOC	33.33%	100.0%	50.00	3
MONEY	66.67%	57.14%	61.54	12
NORP	83.33%	83.33%	83.33	12
ORDINAL	90.00%	94.74%	92.31	20
ORG	92.94%	91.86%	92.40	85
PERCENT	100.0%	100.0%	100.0	8
PERSON	92.92%	98.13%	95.45	113
PRODUCT	50.00%	20.00%	28.57	4
QUANTITY	50.00%	50.00%	50.00	2
TIME	70.00%	66.67%	68.29	20
WORK_OF_ART	90.91%	95.24%	93.02	22
ALL	89.01%	89.01%	89.01	546

Table 11 NER evaluation results

4.5 Evaluating the quality of named entity annotation

To assess the quality of the named entity annotation in the data, we manually annotated mentions of named entities in the set of documents used in the evaluation of the morpho-syntactic analyses (Section 4.4 above) using the same Ontonotes types and guidelines applied when annotating the training corpus for the tagger.[6]

The evaluation results are summarized in Table 11. While the performance of the tagger on this data is lower than the 93% F-score reported on its original in-domain test set (Luoma et al., 2021), we find that the automatic annotation achieves a respectable 89% F-score with balanced precision and recall on the over 500 mentions in the sample. For the important and comparatively frequently mentioned person, organization and GPE (geo-political entity) types, the quality of the annotation exceeds 90% F-score, while performance is notably lower for rare types such as event and facility. These results are broadly in line with expectation from previous work on this and similarly annotated corpora and confirm that even though the dataset represents a domain shift compared to the corpus on which the tagger was trained, the named entity annotation of the corpus is of reasonably high quality.

5 Prior use of the dataset

The corpus has been used to enable a number of studies, some of which we list here to illustrate the use cases for the dataset.

[6] One document had been removed from the dataset prior to this analysis, so only 29 of the 30 documents were used in this evaluation.

Language model training

The corpus was among the primary datasets used to train a series of Finnish language models, starting from the Finnish word2vec models[7], through the Finnish BERT model *FinBERT* (Virtanen et al., 2019) and very recently the Finnish GPT-3 model *FinGPT-3*[8]. As reported by Virtanen et al. (2019), the FinBERT model pushed the state of the art on many NLP tasks. Interestingly, the word2vec models induced on the data were used in a brain imaging study on the connection of fMRI patterns and word vectors (Kivisaari et al., 2019). This study is an example of less typical studies directly enabled by corpus building work such as the Finnish Parsebank.

Linguistic research

The dataset was used in numerous linguistic studies. Of especial interest is that of Huumo et al. (2017) which used an early version of the dependency trees in the Parsebank to find cases of an exceptionally rare Finnish syntactic phenomenon. Other linguistic studies which make use of the dependency syntax structures include a study on discourse connectives (Laippala, Kyröläinen, Kanerva, & Ginter, 2018) and on emoticons (Laippala, Kyröläinen, Kanerva, Luotolahti, & Ginter, 2017). The corpus also served as the source data for the work of Skantsi and Laippala (2023) on Finnish text register classification, used to provide the text register metadata described in Section 3.4.4.

Other

The NoSketchEngine word concordance tool using the dataset has been reportedly extensively used in teaching of Finnish as the second language.[9]

6 Distribution

The current European Union copyright law has a research exemption which allows for publication of otherwise copyrighted text for research purposes. It was introduced in European Union's Directive on Copyright in the Digital Single Market, which more specifically has exemption for data mining called "The Exception for Text and Data Mining (TDM)" or more informally "the right to mine". This exemption was incorporated into the Finnish legislation in April 2023, making it possible to distribute the corpus for research purposes upon request. Further, concordances from the corpus can be explored using the NoSketchEngine tool, and the dependency trees of a sample of the data can be searched using the `dep_search` tool.

7 Conclusions

We presented the Finnish Internet Parsebank, a web-based corpus of the Finnish language. The corpus is set apart from other recent web-corpus work by having several additional layers of annotation, which aim to make the corpus applicable not only as a

[7] https://turkunlp.org/finnish_nlp.html
[8] https://turkunlp.org/gpt3-finnish
[9] Source: private communication with the users

language model training data, but also as a data source for linguistic and other inquiry. While the majority use will likely remain in language model training, we feel it is important to cater also for other, even if rarer use cases. In this vein, we listed several studies which drew their data from this corpus and its annotation layers. Naturally, the corpus also served as an important data source in training a series of Finnish language models, from word2vec embeddings, to the recently completed FinGPT-3 model. In our evaluations of both the underlying textual data as well as the additional annotations, we found the corpus of being of what we believe to be a sufficient quality for various applications. Additionally, the metadata includes language model scores as well as duplication rate scores, which allow further filtering of the data to a higher quality subset. Thanks to the recent changes in Finnish legislation, the dataset is openly available for research purposes.[10]

8 Acknowledgements

This work was generously supported by the KONE Foundation, and further funding was provided by the Research Council of Finland. Computational resources were provided by CSC - IT Center for Science, Finland.

9 Funding

Juhani Luotolahti and Jenna Kanerva were funded by KONE foundation and Research Council of Finland and the work of Jouni Luoma, Valtteri Skantsi, Sampo Pyysalo, Veronika Laippala and Filip Ginter was funded by the Research Council of Finland. The authors declare no conflicts of interest.

10 Author Contributions

Juhani Luotolahti carried out the crawls and data processing as well as led the manuscript writing. Filip Ginter contributed to the dataset cleanup code, many of the analysis pipelines, and helped editing the manuscript to its final form. Jenna Kanerva provided morphosyntactic analyses and their evaluation, and wrote the respective sections of the paper. Jouni Luoma and Sampo Pyysalo provided the NER predictions and their evaluation, and wrote the respective sections of the paper. Valtteri Skantsi and Veronika Laippala provided the register predictions and their evaluation, and wrote the respective sections of the paper. Filip Ginter and Veronika Laippala conceived the overall study, obtained the funding and served as the PIs of the project.

References

Abadji, J., Suárez, P.J.O., Romary, L., Sagot, B. (2021). Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. H. Lüngen, M. Kupietz, P. Bański, A. Barbaresi, S. Clematide, & I. Pisetta

[10] https://turkunlp.org/finnish_nlp.html

(Eds.), (pp. 1 – 9). Mannheim: Leibniz-Institut für Deutsche Sprache. Retrieved from https://nbn-resolving.org/urn:nbn:de:bsz:mh39-104688

Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, *43*(3), 209–226,

Baroni, M., & Kilgarriff, A. (2006). Large linguistically-processed web corpora for multiple languages. *Eacl'06: Proceedings of the eleventh conference of the european chapter of the association for computational linguistics: Posters & demonstrations; 2006 apr 5-6; trento, italy. stroudsburg (pa): Association for computational linguistics; 2006. p. 87-90.*

Goyal, N., Du, J., Ott, M., Anantharaman, G., Conneau, A. (2021). Larger-scale transformers for multilingual masked language modeling. *arXiv preprint arXiv:2105.00572*, ,

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., ... Ginter, F. (2014). Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, *48*, 493-531, https://doi.org/10.1007/s10579-013-9244-1 Retrieved from http://dx.doi.org/10.1007/s10579-013-9244-1 (Open access)

Heafield, K. (2011, July). KenLM: Faster and smaller language model queries. *Proceedings of the sixth workshop on statistical machine translation* (pp. 187–197). Edinburgh, Scotland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W11-2123

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., Weischedel, R. (2006). Ontonotes: the 90% solution. *Proceedings of the human language technology conference of the naacl, companion volume: Short papers* (pp. 57–60).

Huumo, T., Kyröläinen, A.-J., Kanerva, J., Luotolahti, J., Salakoski, T., Ginter, F., Laippala, V. (2017). Distributional semantics of the partitive A argument construction in Finnish. In M. Luodonpää-Manni, E. Penttilä, & J. Viimaranta (Eds.), *Empirical approaches to cognitive linguistics: Analysing real-life data.* Cambridge Scholars Publishing.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., Suchomel, V. (2013). The tenten corpus family. *7th international corpus linguistics conference cl* (pp. 125–127).

Kanerva, J., & Ginter, F. (2022). Out-of-domain evaluation of finnish dependency parsing. *Proceedings of the 13th international conference on language resources* *and evaluation (lrec'22)* (p. 1114--1124). Retrieved from http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.120.pdf

Kanerva, J., Ginter, F., Miekka, N., Leino, A., Salakoski, T. (2018). Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. *Proceedings of the conll 2018 shared task: Multilingual parsing from raw text to universal dependencies.* Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/K18-2013

Kanerva, J., Ginter, F., Salakoski, T. (2020). Universal lemmatizer: A sequence to sequence model for lemmatizing universal dependencies treebanks. *Natural Language Engineering*, 1–30, https://doi.org/10.1017/S1351324920000224 Retrieved from http://dx.doi.org/10.1017/S1351324920000224

Kilgarriff, A. (2007). Googleology is bad science. *Computational linguistics*, *33*(1), 147–151,

Kivisaari, S.L., van Vliet, M., Hultén, A., Lindh-Knuutila, T., Faisal, A., Salmelin, R. (2019, Feb 25). Reconstructing meaning from bits of information. *Nature Communications*, *10*(1), 927, https://doi.org/10.1038/s41467-019-08848-0 Retrieved from https://doi.org/10.1038/s41467-019-08848-0

Laippala, V., Kyröläinen, A.-J., Kanerva, J., Ginter, F. (2018). Dependency profiles in the large-scale analysis of discourse connectives. *Corpus linguistics and linguistic theory*, ,

Laippala, V., Kyröläinen, A.-J., Kanerva, J., Luotolahti, J., Ginter, F. (2017). Dependency profiles as a tool for big data analysis of linguistic constructions: A case study of emoticons. *Journal of Estonian and Finno-Ugric Linguistics. Grammar in Use: Approaches to Baltic Finnic.*, *8*, 127–153,

Luoma, J., Chang, L.-H., Ginter, F., Pyysalo, S. (2021). Fine-grained named entity annotation for finnish. *Proceedings of the 23rd nordic conference on computational linguistics (nodalida)* (pp. 135–144).

Luoma, J., & Pyysalo, S. (2020, December). Exploring cross-sentence contexts for named entity recognition with BERT. *Proceedings of the 28th international conference on computational linguistics* (pp. 904–914). Barcelona, Spain (Online): International Committee on Computational Linguistics. Retrieved from https://aclanthology.org/2020.coling-main.78

Öhman, J., Verlinden, S., Ekgren, A., Gyllensten, A.C., Isbister, T., Gogoulou, E., … Sahlgren, M. (2023). The nordic pile: A 1.2 tb nordic dataset for language modeling. *arXiv preprint arXiv:2303.17183*, ,

Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora* (Unpublished doctoral dissertation). Masaryk university, Faculty of informatics, Brno, Czech Republic.

Pyysalo, S., Kanerva, J., Missilä, A., Laippala, V., Ginter, F. (2015). Universal Dependencies for Finnish. *Proceedings of nodalida 2015* (pp. 163–172). NEALT. Retrieved from https://aclweb.org/anthology/W/W15/W15-1821.pdf

Schäfer, R. (2015). Processing and querying large web corpora with the COW14 architecture. P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, & A. Witt (Eds.), *Proceedings of challenges in the management of large corpora 3 (cmlc-3).* Lancaster: IDS.

Skantsi, V., & Laippala, V. (2023). Analyzing the unrestricted web: The finnish corpus of online registers. *Nordic Journal of Linguistics*, 1–31, https://doi.org/10.1017/S0332586523000021

Suchomel, V., & Pomikálek, J. (2012). Efficient web crawling for large text corpora. *Proceedings of the seventh web as corpus workshop (wac7)* (pp. 39–43).

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., … Pyysalo, S. (2019). *Multilingual is not enough: Bert for finnish.*

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., … Li, J. (2017). CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.* Association for Computational Linguistics.