HPLT Analytics report

General overview

Corpus	Analytics date	Language
HPLT-v2-fin_Latn.tsv	9/16/2024	Finnish (fi)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
34,815,601	976,619,897			149.9 GB	154,736,868,874

Top 10 domains

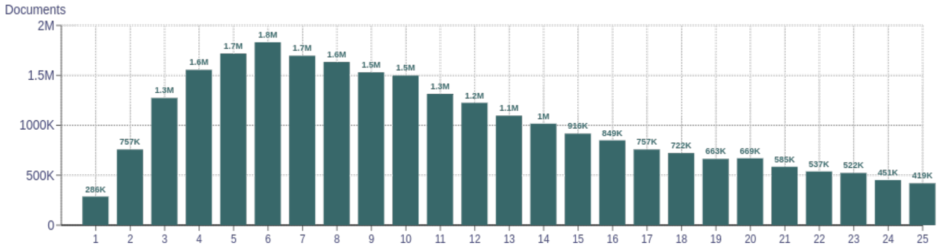
Domain	Docs	% of total
blogspot.com	3M	8.60
blogspot.fi	2.8M	8.09
wikipedia.org	926K	2.66
mtv.fi	661K	1.90
docplayer.fi	531K	1.53
vuodatus.net	446K	1.28
suomi24.fi	429K	1.23
lily.fi	299K	0.86
wordpress.com	277K	0.80
yle.fi	238K	0.68

Top 10 TLDs

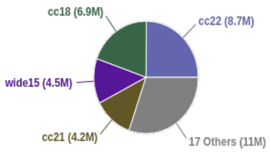
Domain	Docs	% of total
fi	21M	60.79
com	9M	25.76
net	1.6M	4.54
org	1.3M	3.86
eu	304K	0.87
info	232K	0.67
se	102K	0.29
de	97K	0.28
ru	76K	0.22
no	55K	0.16

Documents size (in segments)

<= 25 segments **73.31%** (26M documents)
> 25 segments **26.69%** (9.3M documents)

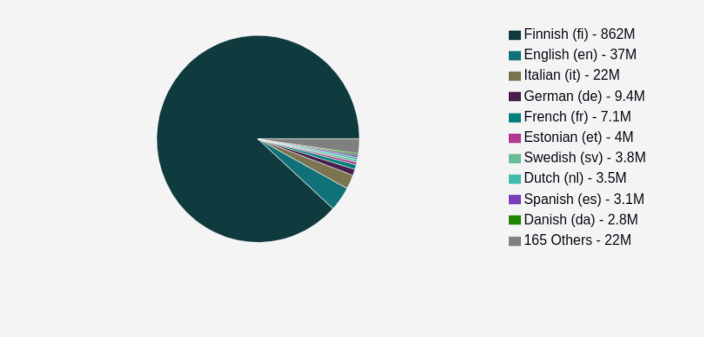


Documents by collection

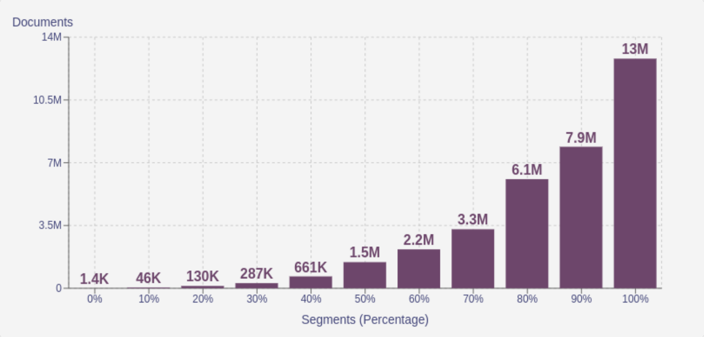


Language Distribution

Number of segments

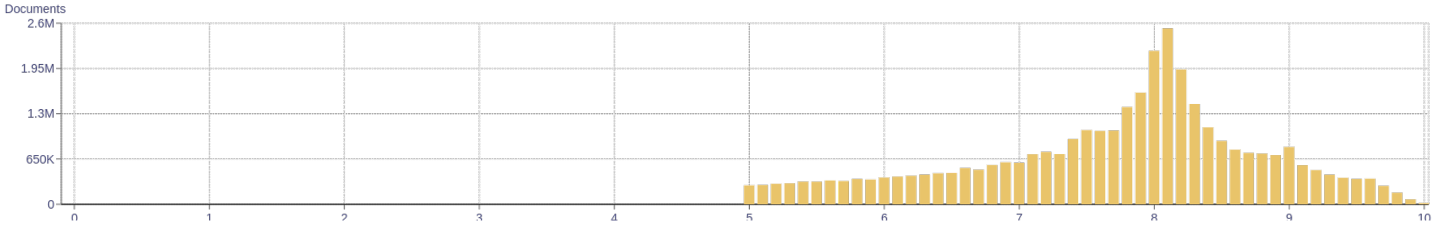


Percentage of segments in Finnish (fi) inside documents

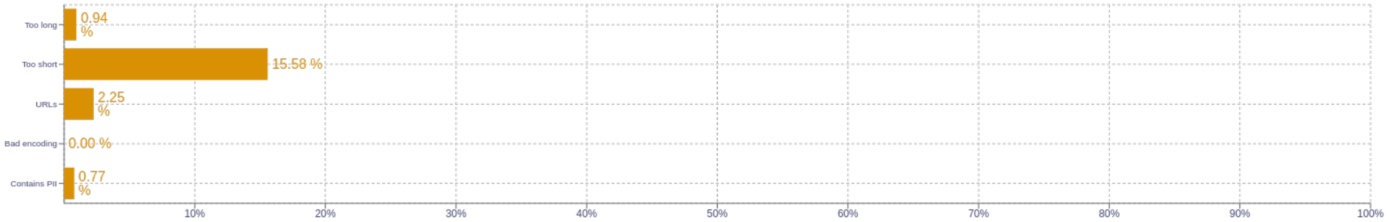


Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (35M documents)



Segment noise distribution



About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, <ul>, <ol>, etc.) replaced by newlines.

Volumes - Tokens

Tokenized with https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md

Type-Token Ratio

Lexical variety computed as *number or types (uniques)/number of tokens*, after removing punctuation (https://www.sltinfo.com/wp-content/uploads/2014/01/type-token-ratio.pdf).

Document size (in segments)

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, <ul>, <ol>, etc.) replaced by newlines.

Language distribution

Language identified with FastSpell (https://github.com/mbanon/fastspell).

Distribution of segments by fluency score

Obtained with Monocleaner (https://github.com/bitextor/monocleaner).

Distribution of documents by average fluency score

Obtained with Monocleaner (https://github.com/bitextor/monocleaner).

Distribution of documents by document score

Obtained with Web Docs Scorer (https://github.com/pablop16n/web-docs-scorer/).

Segment length distribution by token

Tokenized with https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md

Segment noise distribution

Obtained with Bicleaner Hardrules (https://github.com/bitextor/bicleaner-hardrules/).

Frequent n-grams

Tokenized with https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md, after removing n-grams starting or ending in a stopword. Stopwords from https://github.com/hplt-project/data-analytics-tool/blob/main/scripts/resources/README.txt