HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
fin_Latn.jsonl.tsv	6/18/2025	Finnish (fi)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
34,815,557	976,395,607	369,343,799 (37.83 %)	22B	154,736,868,874	149.9 GB

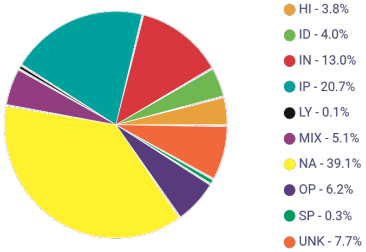
Top 10 domains

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%

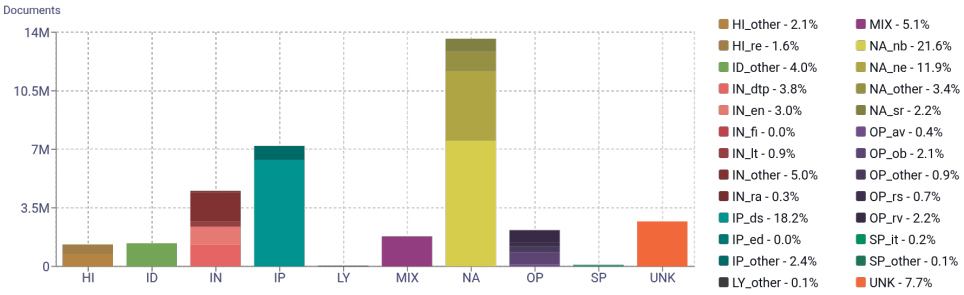
Top 10 TLDs

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%

Register labels

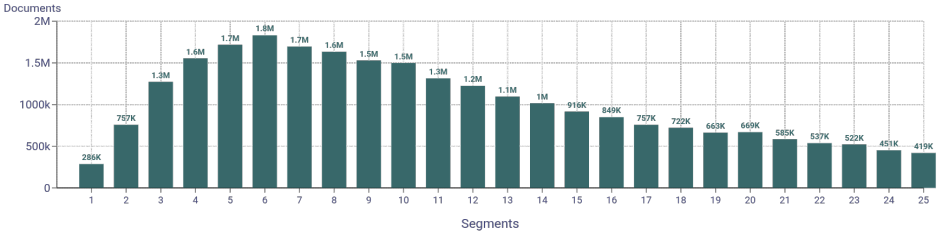


MT:4.9% | 1.7M Documents



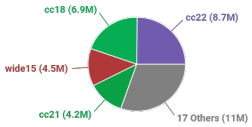
Documents size (in segments)

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



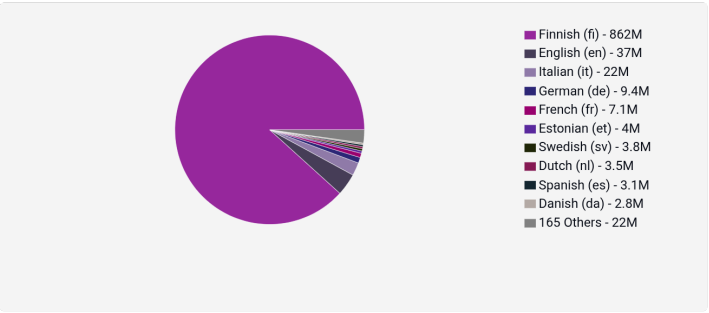
Documents by collection

CC = 67.47%
IA = 32.53%

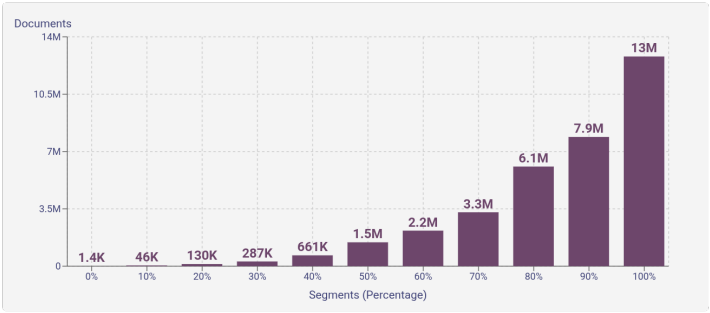


Language Distribution

Number of segments in the Finnish (fi) corpus

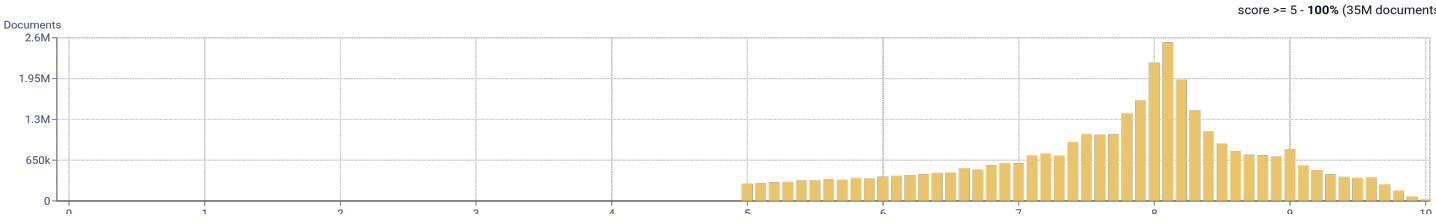


Percentage of segments in Finnish (fi) inside documents



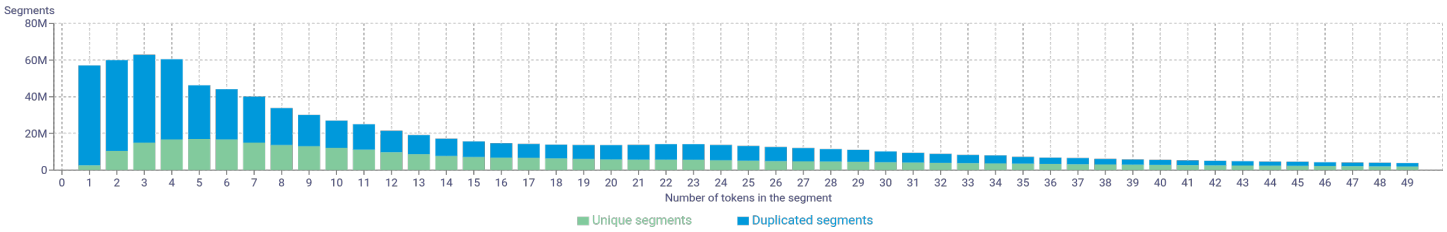
Distribution of documents by document score

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

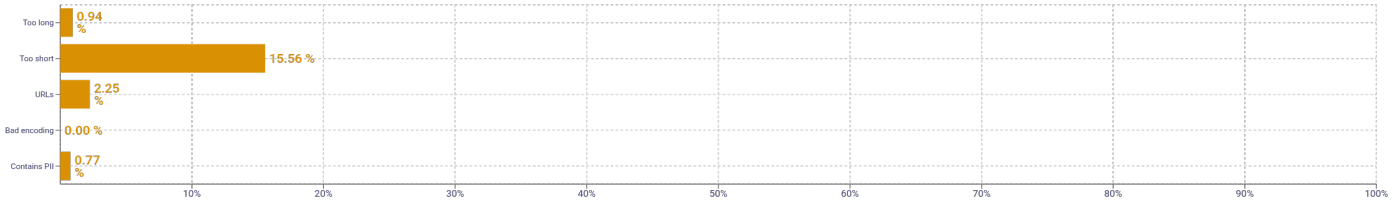


Segment length distribution by token

≤ 49 tokens = **314M** segments | **555M** duplicates
> 50 tokens = **107M** segments | **52M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	voi | 45786156 · jo | 32904042 · vain | 30085692 · sitten | 29553005 · vielä | 26025452
2	muun muassa | 4164197 · tällä hetkellä | 3484605 · muokkaa wikitekstiä | 3162248 · lue lisää | 2938806 · tällä kertaa | 2168233
3	lasten ja nuorten | 763765 · hotellia viimeisen tunnin | 436579 · ketjusta on poistettu | 396132 · a b c | 388571 · hallituksen esitys eduskunnalle | 374676
4	hotellia viimeisen tunnin sisällä | 436579 · tarkasteli tätä hotellia viimeisen | 355308 · henkilöä tarkasteli tätä hotellia | 355308 · hallituksen esitys eduskunnalle laiksi | 239730 · a b c d | 231039
5	henkilöä tarkasteli tätä hotellia viimeisen | 355308 · aineen tai seoksen ja yhtiön | 162106 · kokouksen laillisuuden ja päätösvaltaisuuden toteaminen | 149027 · a b c d e | 144817 · esityksen pääasiallinen sisältö esityksessä ehdotetaan | 142340

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

Register labels

Name	Abbr.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av