HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-fin_Latn	9/18/2025	Finnish

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
49,558,089	1,365,997,637	754,820,208 (55.26 %)	31B	217,907,732,179	210.57 GB

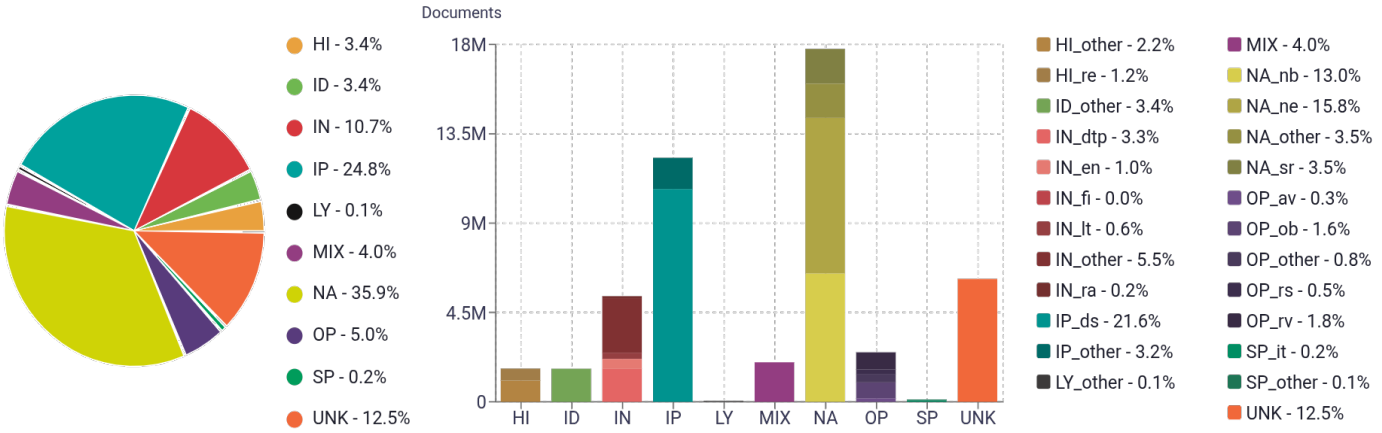
Top 10 domains

Domain	Docs	% of total
blogspot.com	3.2M	6.46%
blogspot.fi	1.2M	2.40%
mtv.fi	1.1M	2.13%
iltalehti.fi	941K	1.90%
docplayer.fi	733K	1.48%
yle.fi	477K	0.96%
vuodatus.net	413K	0.83%
hs.fi	405K	0.82%
wordpress.com	355K	0.72%
wikipedia.org	345K	0.70%

Top 10 TLDs

Domain	Docs	% of total
fi	30M	59.66%
com	13M	26.10%
net	2M	4.01%
eu	1.5M	3.12%
org	951K	1.92%
info	306K	0.62%
se	252K	0.51%
ru	223K	0.45%
nl	136K	0.27%
ee	94K	0.19%

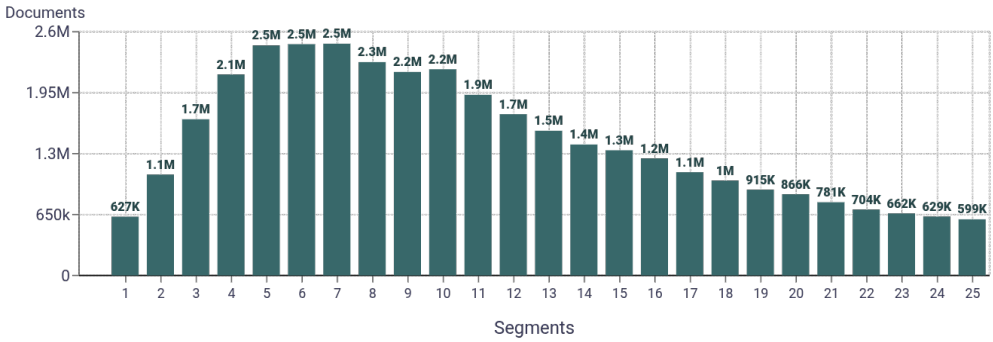
Register labels



MT:9.2% | 4.6M Documents

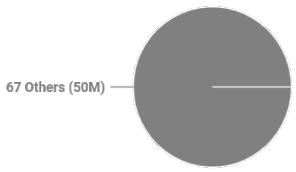
Documents size (in segments)

≤ 25 segments **72.62%** (36M documents)
> 25 segments **27.38%** (14M documents)



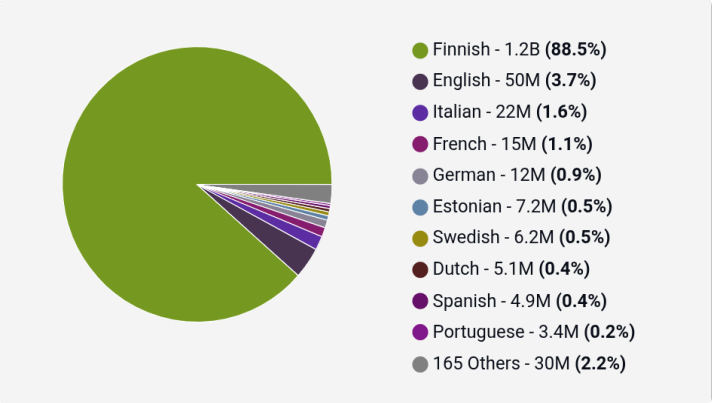
Document collections

CC = 89.88%
IA = 10.12%

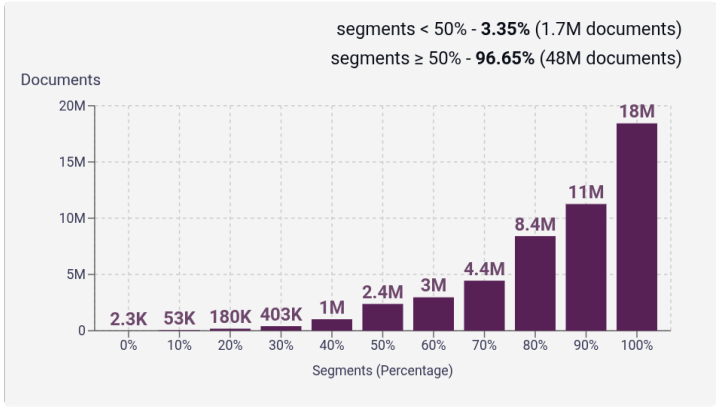


Language Distribution

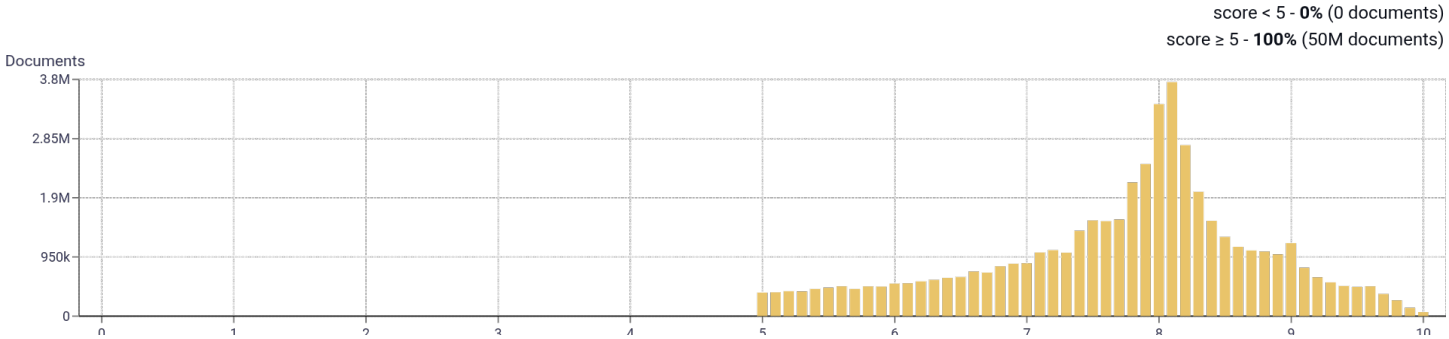
Number of segments in the Finnish corpus



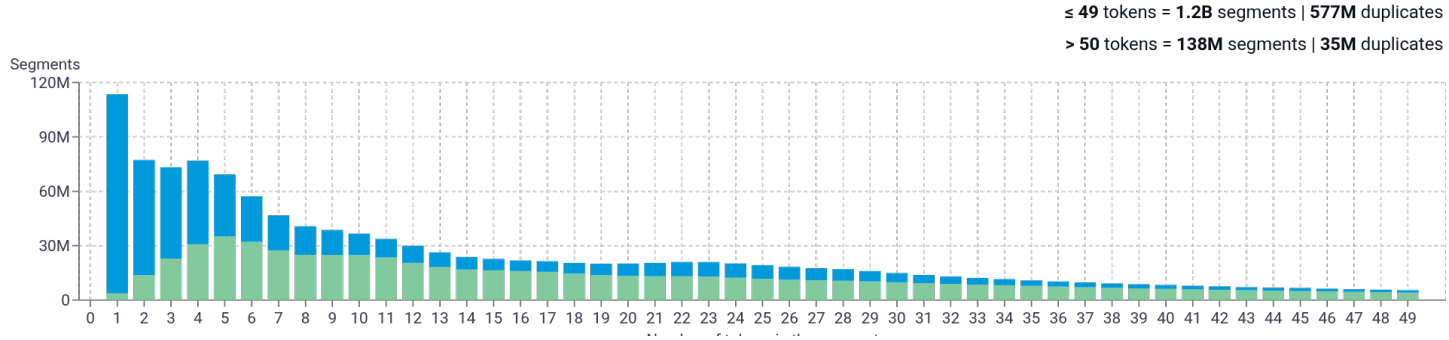
Percentage of segments in Finnish inside documents



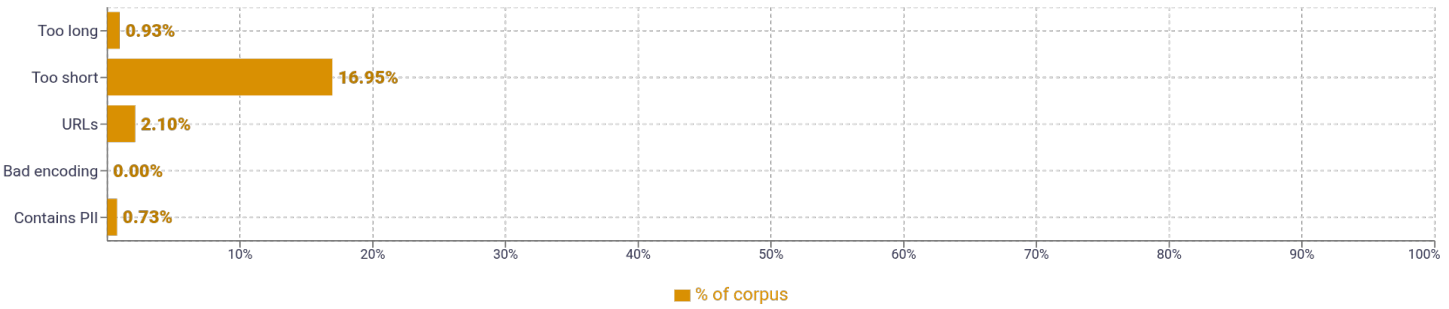
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	voi | 63,964,633 · porno | 54,777,826 · hieronta | 53,834,872 · seksi | 44,019,558 · sex | 40,452,135
2	thai hieronta | 22,613,073 · eroottinen hieronta | 7,933,914 · nainen etsii | 7,009,112 · muun muassa | 5,781,575 · tällä hetkellä | 4,773,984
3	nainen etsii miestä | 2,510,129 · thai hieronta helsinki | 1,180,648 · one night stand | 1,081,961 · ilmoita asiaton viesti | 999,096 · lasten ja nuorten | 918,424
4	nainen etsii nuorempaa miestä | 427,632 · nainen etsii nuorta miestä | 371,719 · nainen etsii miestä seksiä | 360,239 · seksitreffit nainen etsii miestä | 313,854 · nainen etsii miestä oulu | 308,405
5	body to body massage helsinki | 255,784 · tallink spa conference hotel kokemuksia | 193,387 · meriton grand conference spa hotel | 191,327 · grand conference spa hotel kokemuksia | 186,989 · nimi ja kotipaikka yhdistyksen nimi | 156,693

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.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	**How-to or instructions**	**HI**	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	**Informational persuasion**	**IP**	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	**Opinion**	**OP**
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	**Informational description**	**IN**	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				