HPLT Analytics report

HPLT Analytics

General overview

Corpus	Date	Language
hplt-v3-eng_Latn-SAMPLED	9/24/2025	English

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
3,216,000	78,157,600	65,761,988 (84.14 %)	2.5B	12,879,848,191	12.08 GB

Top 10 domains

Domain	Docs	% of total
blogspot.com	92K	2.86%
wordpress.com	41K	1.26%
typepad.com	4.4K	0.14%
tumblr.com	3.4K	0.11%
stackexchange.com	3.4K	0.10%
wikipedia.org	3K	0.09%
yahoo.com	2.8K	0.09%
cbslocal.com	2.6K	0.08%
google.com	2.5K	0.08%
blogspot.co.uk	2.5K	0.08%

Top 10 TLDs

Domain	Docs	% of total
com	2.1M	66.85%
org	269K	8.38%
co.uk	118K	3.66%
net	104K	3.23%
edu	54K	1.69%
com.au	49K	1.52%
ca	41K	1.29%
info	21K	0.66%
in	21K	0.64%
gov	18K	0.57%

Register labels

- HI - 3.0%
- ID - 2.8%
- IN - 32.9%
- IP - 10.0%
- LY - 0.2%
- MIX - 7.0%
- NA - 24.9%
- OP - 8.4%
- SP - 0.5%
- UNK - 10.3%

- HI_other - 2.1%
- HI_re - 0.9%
- ID_other - 2.8%
- IN_dtp - 14.6%
- IN_en - 0.5%
- IN_fi - 0.0%
- IN_lt - 0.5%
- IN_other - 16.8%
- IN_ra - 0.6%
- IP_ds - 7.9%
- IP_other - 2.0%
- LY_other - 0.2%
- MIX - 7.0%
- NA_nb - 4.8%
- NA_ne - 13.9%
- NA_other - 3.2%
- NA_sr - 2.9%
- OP_av - 0.2%
- OP_ob - 1.3%
- OP_other - 3.6%
- OP_rs - 1.0%
- OP_rv - 2.3%
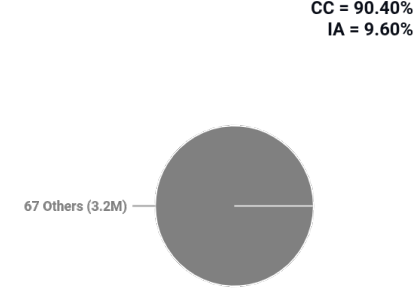- SP_it - 0.3%
- SP_other - 0.2%
- UNK - 10.3%
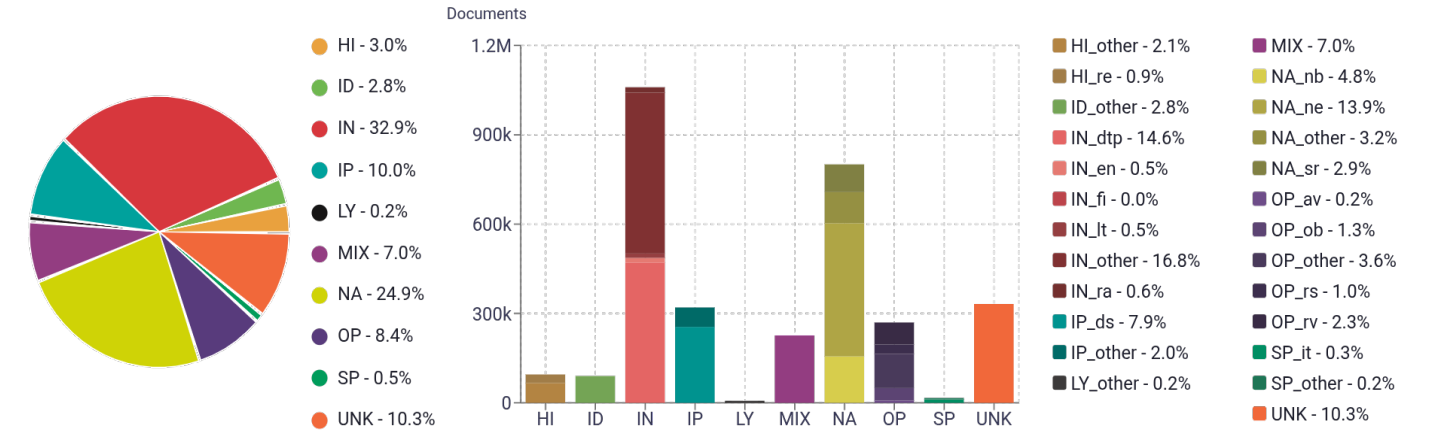
MT:1.9% | 61K Documents

Documents size (in segments) ⓘ
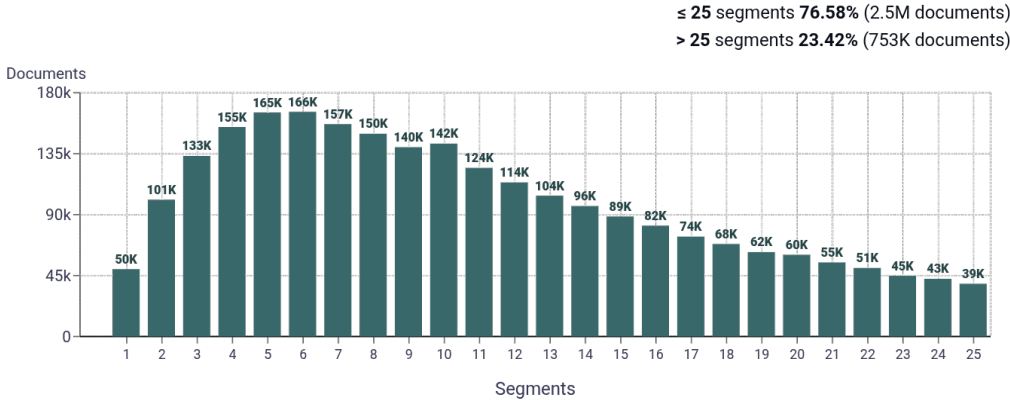
≤ 25 segments **76.58%** (2.5M documents)
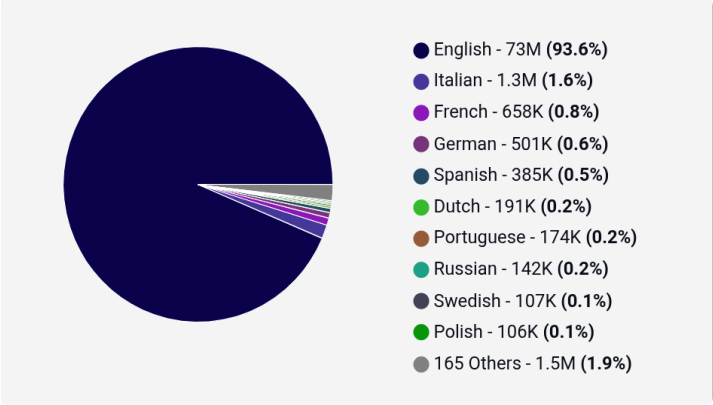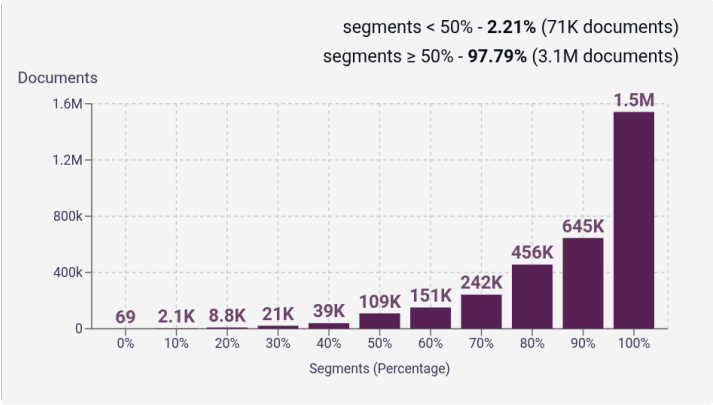> 25 segments **23.42%** (753K documents)
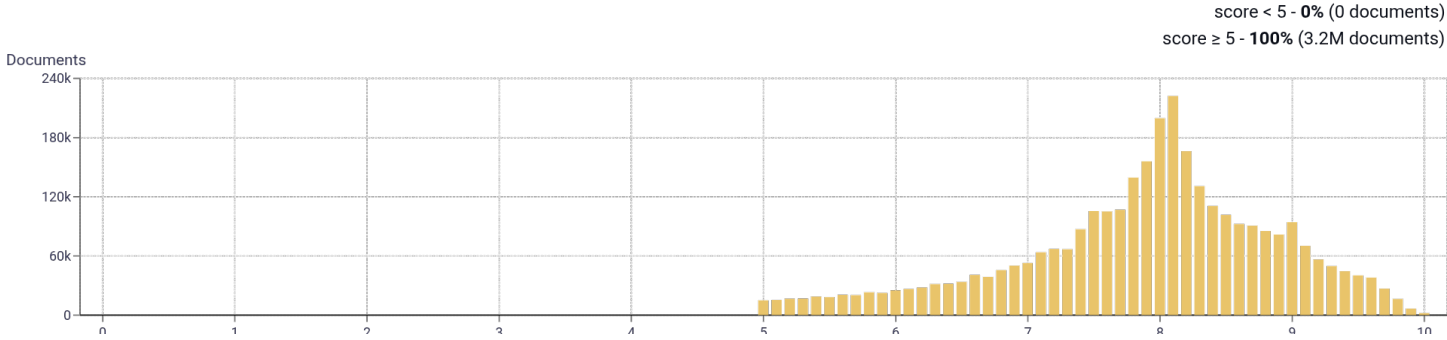

Document collections

CC = 90.40%
IA = 9.60%

67 Others (3.2M)

Language Distribution

Number of segments in the English corpus

- English - 73M **(93.6%)**
- Italian - 1.3M **(1.6%)**
- French - 658K **(0.8%)**
- German - 501K **(0.6%)**
- Spanish - 385K **(0.5%)**
- Dutch - 191K **(0.2%)**
- Portuguese - 174K **(0.2%)**
- Russian - 142K **(0.2%)**
- Swedish - 107K **(0.1%)**
- Polish - 106K **(0.1%)**
- 165 Others - 1.5M **(1.9%)**

Percentage of segments in English inside documents

segments < 50% - **2.21%** (71K documents)
segments ≥ 50% - **97.79%** (3.1M documents)

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (3.2M documents)

Segment length distribution by token

≤ **49** tokens = **62M** segments | **12M** duplicates
> **50** tokens = **16M** segments | **370K** duplicates

Segment noise distribution

- Too long — **0.94%**
- Too short — **13.79%**
- URLs — **2.39%**
- Bad encoding — **0.02%**
- Contains PII — **0.58%**

■ % of corpus
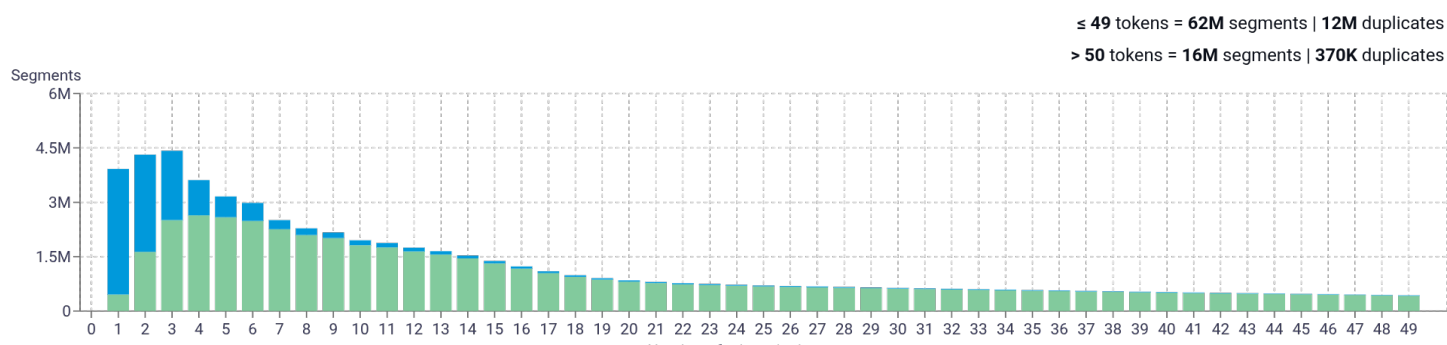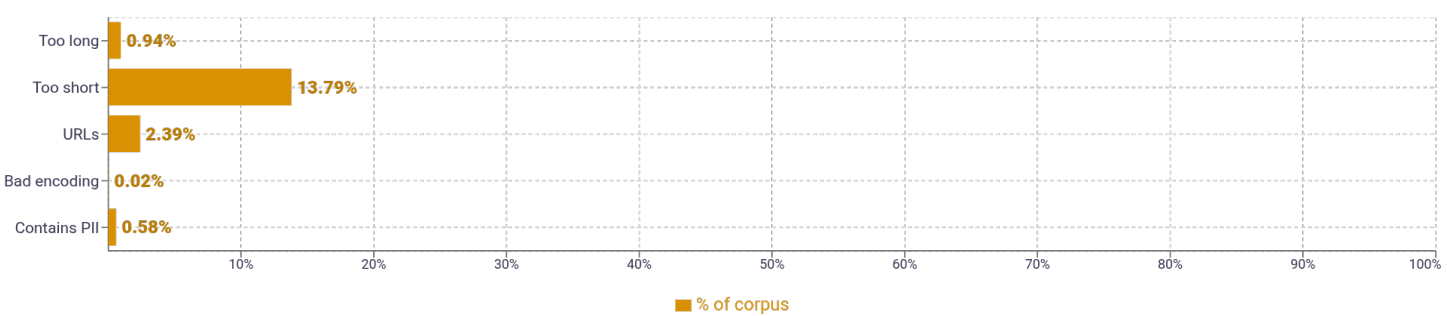
Frequent n-grams

SIZE	N-GRAMS	
1	time | 3,815,893 new | 3,725,847 like | 3,595,653 get | 3,053,016 would | 2,976,135	⧉
2	new york | 338,649 united states | 332,800 make sure | 301,633 years ago | 202,407 social media | 199,329	⧉
3	around the world | 100,436 post a comment | 98,412 take a look | 59,210 new york city | 53,619 need to know | 49,777	⧉
4	thank you so much | 30,834 end of the day | 19,321 give us a call | 15,316 due to the fact | 13,812 thank you for sharing | 13,045	⧉
5	everything you need to know | 7,595 president of the united states | 6,057 keep up the good work | 5,583 please feel free to contact | 5,141 embodiment of the present invention | 5,108	⧉

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				