HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-jpn_Jpan-SAMPLED	10/3/2025	Japanese

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
667,404,438	35,767,735,744	33,130,761 (0.09 %)	99.91%	818B	1,461,409,987,151	3.63 TB

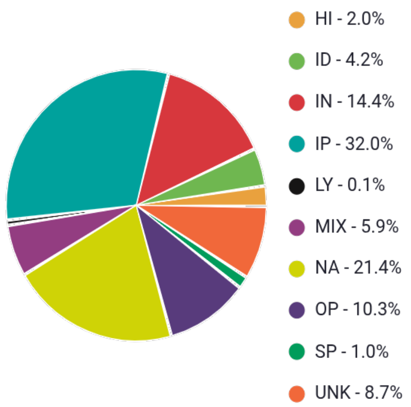
Top 10 domains

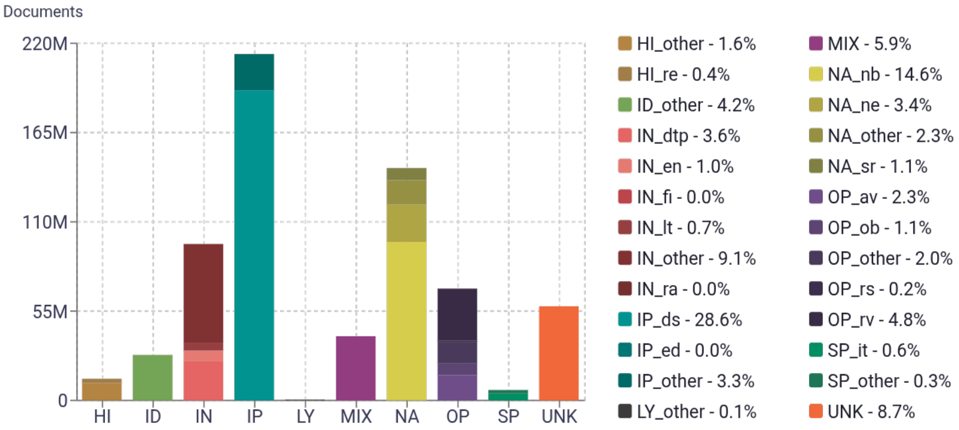
Domain	Docs	% of total
fc2.com	32M	4.76%
hatenablog.com	11M	1.62%
cocolog-nifty.com	10M	1.50%
exblog.jp	7.1M	1.07%
fortune-uranai....	5.4M	0.80%
yahoo.co.jp	4.4M	0.66%
seesaa.net	3.8M	0.57%
goo.ne.jp	3.6M	0.54%
tabelog.com	3.1M	0.47%
biglobe.ne.jp	2.6M	0.39%

Top 10 TLDs

Domain	Docs	% of total
com	304M	45.58%
jp	146M	21.86%
net	50M	7.53%
co.jp	46M	6.83%
info	16M	2.45%
org	16M	2.33%
ne.jp	14M	2.03%
biz	6M	0.90%
or.jp	4.8M	0.72%
xyz	4.4M	0.66%

Register labels



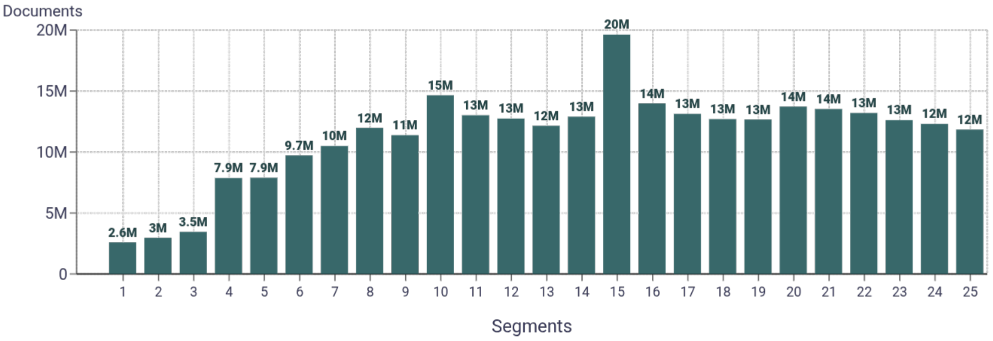


MT:3.3% | 22M Documents

Documents size (in segments) ⓘ

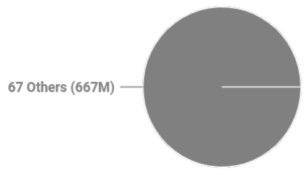
≤ 25 segments **42.43%** (283M documents)

> 25 segments **57.57%** (384M documents)



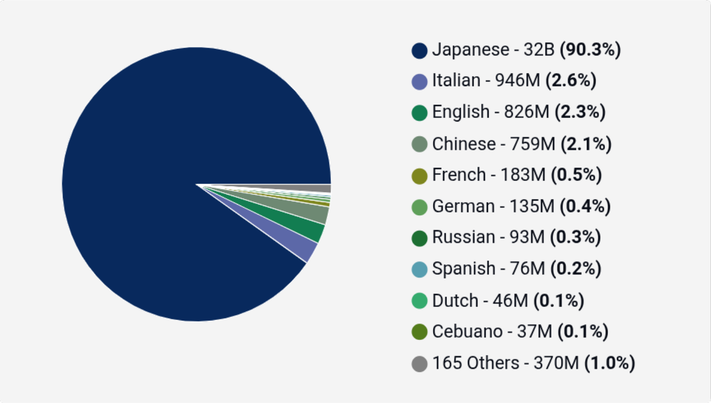
Document collections

CC = 84.89%
IA = 15.11%

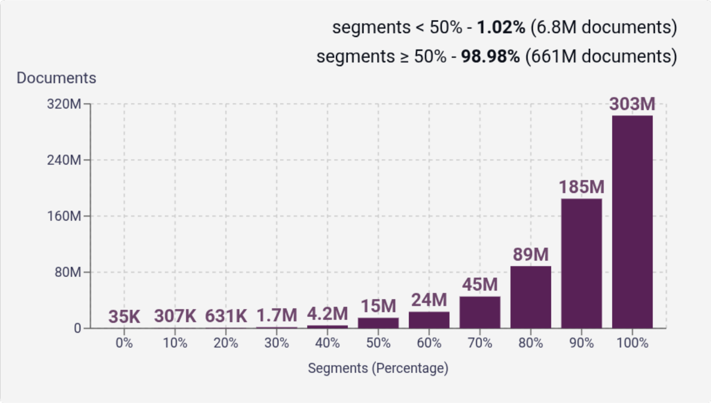


Language Distribution

Number of segments in the Japanese corpus



Percentage of segments in Japanese inside documents

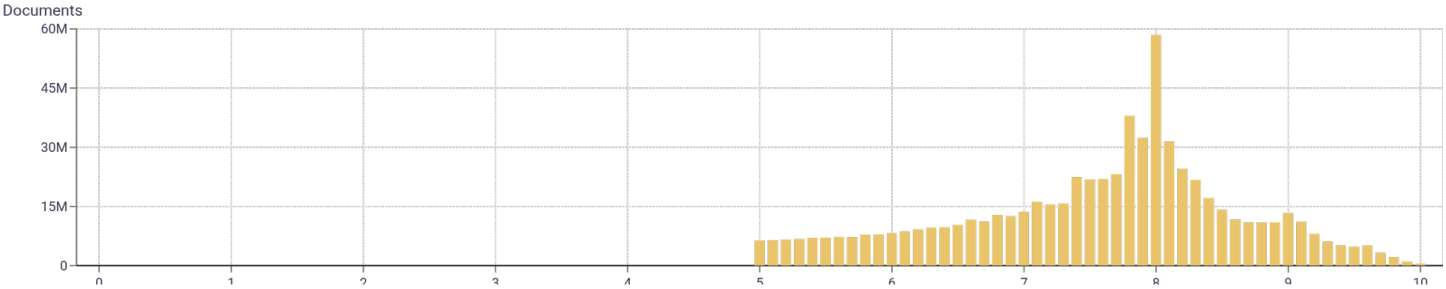

segments < 50% - **1.02%** (6.8M documents)

segments ≥ 50% - **98.98%** (661M documents)

Distribution of documents by document score

score < 5 - **0%** (0 documents)

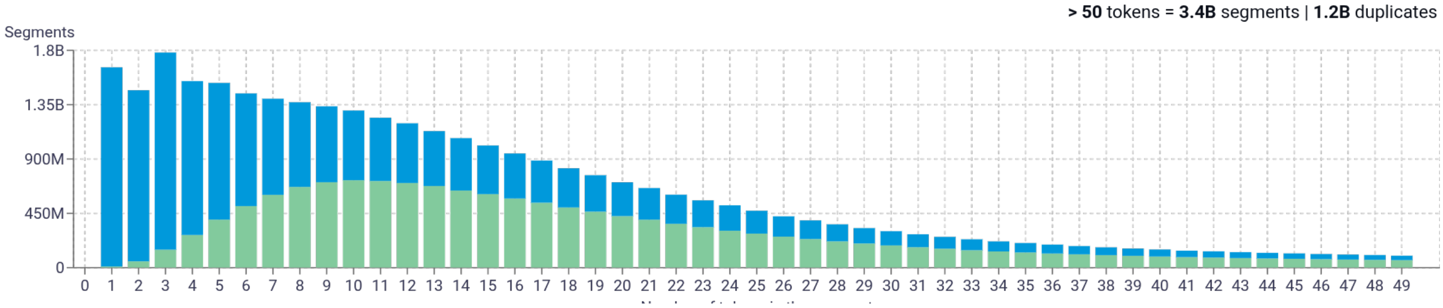
score ≥ 5 - **100%** (667M documents)



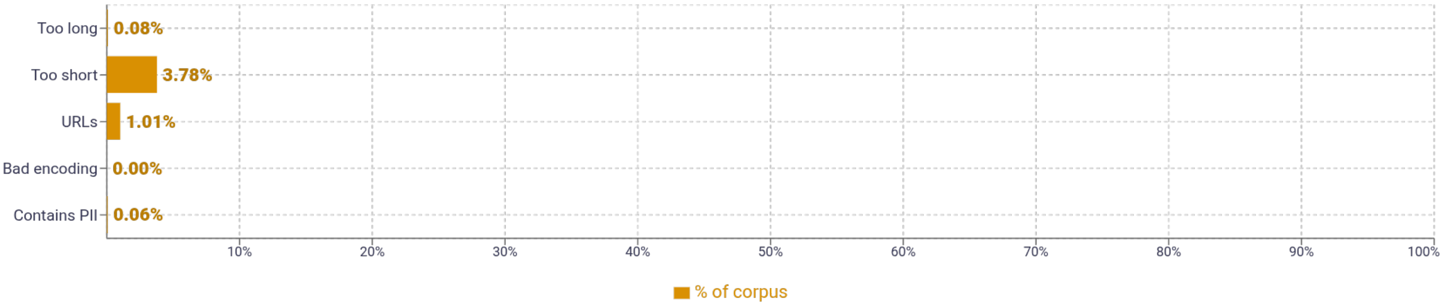
Segment length distribution by token

≤ **49** tokens = **32B** segments | **18B** duplicates

> **50** tokens = **3.4B** segments | **1.2B** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	て | 21,012,489 で | 16,793,720 た | 13,473,279 ます | 8,191,653 な | 6,729,083
2	て い | 2,862,824 て いる | 2,623,445 まし た | 2,427,244 い ます | 1,506,673 さ れ | 1,505,429
3	て い ます | 1,409,695 さ れ て | 642,107 て い た | 639,080 さ れ た | 417,165 れ て い | 409,248
4	さ れ て いる | 240,605 さ れ て い | 236,053 れ て い ます | 235,438 て い まし た | 191,964 こと が でき ます | 185,241
5	か も しれ ませ ん | 176,734 さ れ て い ます | 135,902 で は あり ませ ん | 105,411 する こと が でき ます | 63,342 で は ない でしょう か | 51,223

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