

General overview

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

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

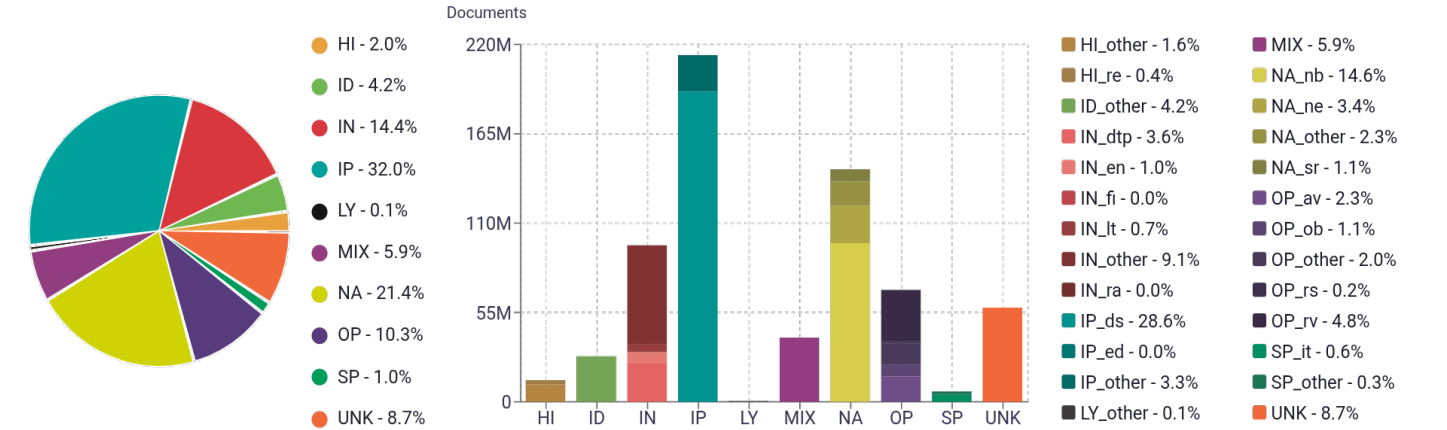
Top 10 domains

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

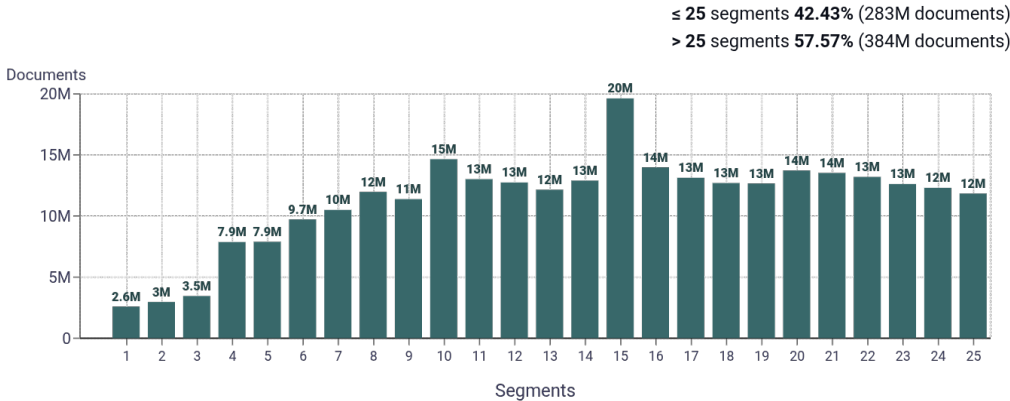
Domain	Docs	% of total
com	304M	45.58%
jp	146M	21.86%
net	50M	7.53%
co.jp	46M	6.83%
info	16M	2.45%
org	16M	2.33%
ne.jp	14M	2.03%
biz	6M	0.90%
or.jp	4.8M	0.72%
xyz	4.4M	0.66%

Register labels

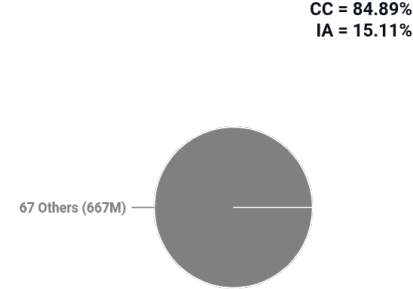


MT:3.3% | 22M Documents

Documents size (in segments) ⓘ

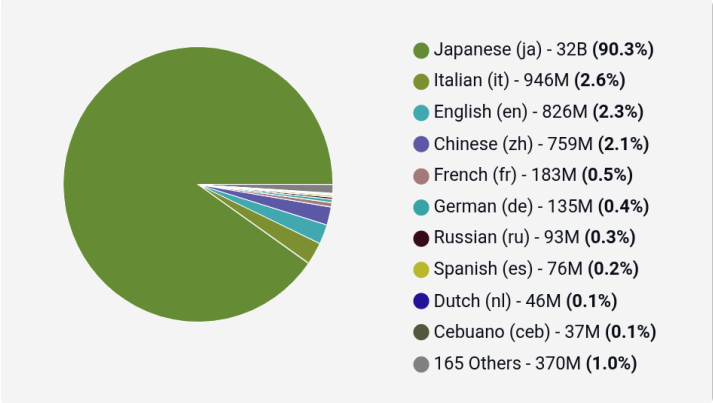


Document collections

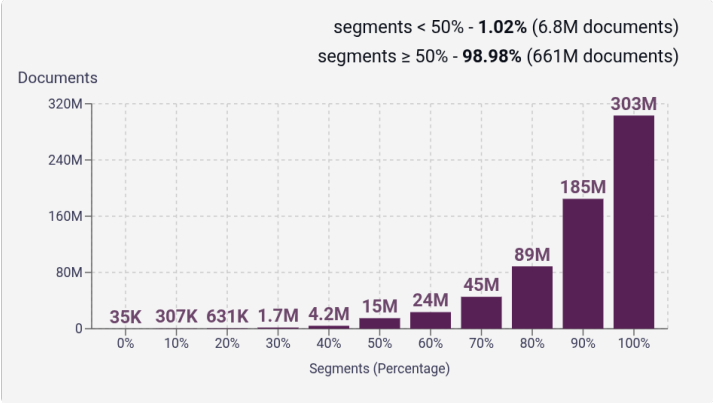


Language Distribution

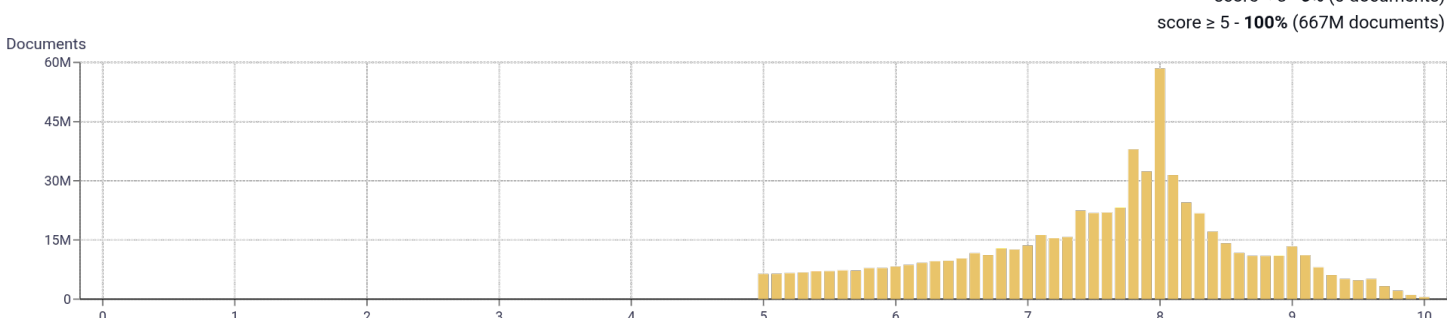
Number of segments in the Japanese (jpn) corpus



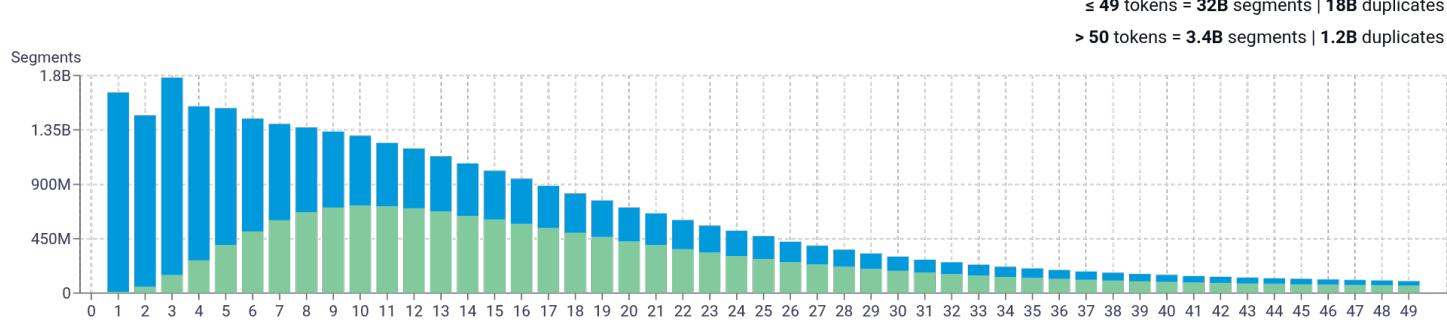
Percentage of segments in Japanese (jpn) inside documents



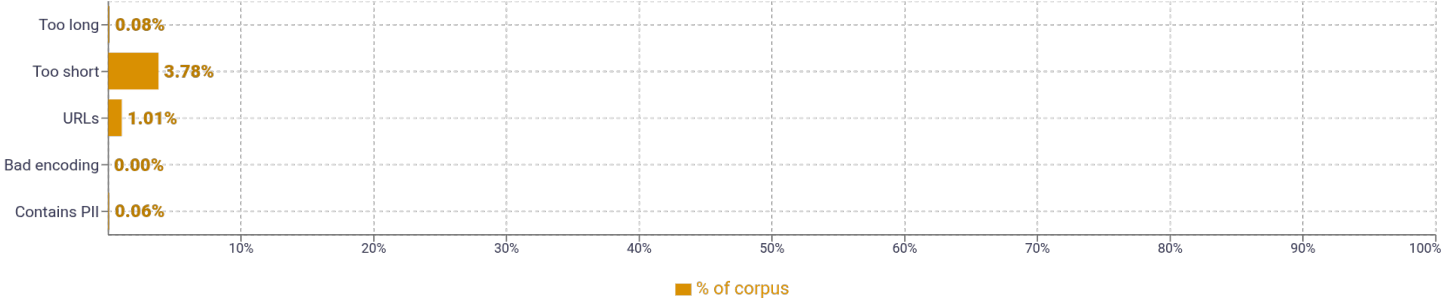
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	て 21,012,489 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>, , , 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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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>, , , 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	ntp
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				