HPLT Analytics report

General overview

Corpus	Date	Language
hplt-v3-sat_Olck	9/18/2025	Santali (sat)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
4,719	75,233	59,195 (78.68 %)	2.4M	11,543,667	27.14 MB

Top 10 domains

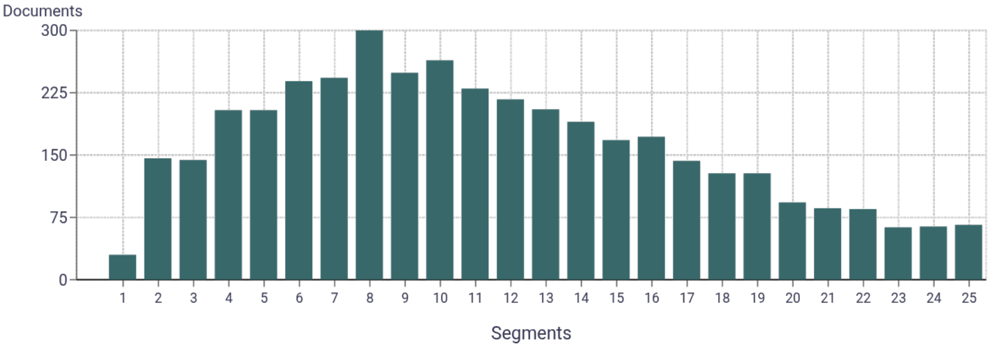
Domain	Docs	% of total
wikipedia.org	4K	84.95%
ayoarang.in	165	3.50%
vikaspedia.in	147	3.12%
santalinews.com	137	2.90%
tribetv.in	90	1.91%
parsipoha.com	32	0.68%
arshalgroup.com	31	0.66%
raharahla.com	20	0.42%
globalvoices.org	20	0.42%
wikimedia.org	14	0.30%

Top 10 TLDs

Domain	Docs	% of total
org	4.1K	85.89%
in	403	8.54%
com	240	5.09%
click	12	0.25%
cf	7	0.15%
ac.in	2	0.04%
भारोत	1	0.02%
net	1	0.02%

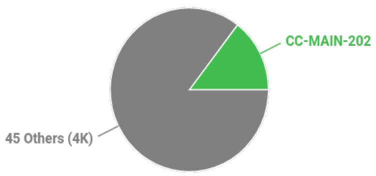
Documents size (in segments) ⓘ

≤ 25 segments **86.06%** (4.1K documents)
> 25 segments **13.94%** (658 documents)



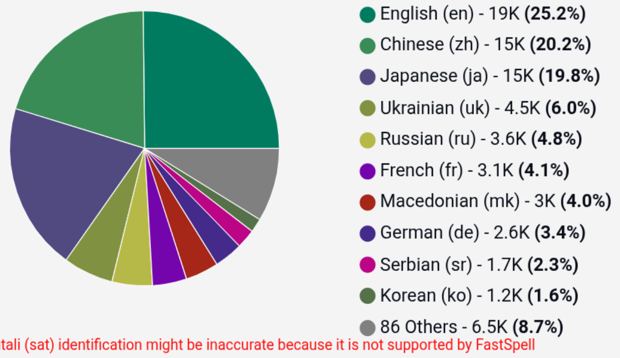
Document collections

CC = 99.98%
IA = 0.02%



Language Distribution

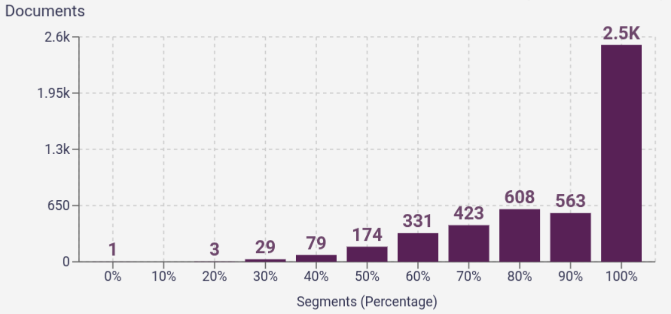
Number of segments in the Santali (sat) corpus



*Santali (sat) identification might be inaccurate because it is not supported by FastSpell

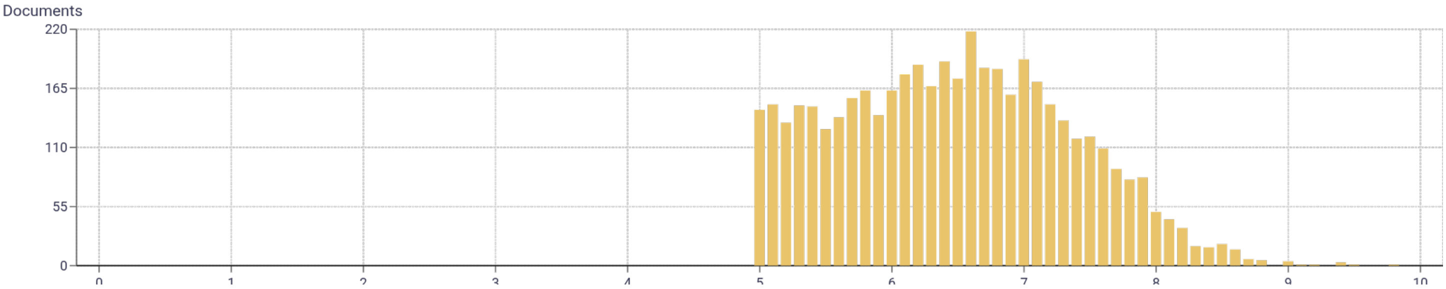
Percentage of segments in Santali (sat) inside documents

segments < 50% - **2.37%** (112 documents)
segments ≥ 50% - **97.63%** (4.6K documents)



Distribution of documents by document score

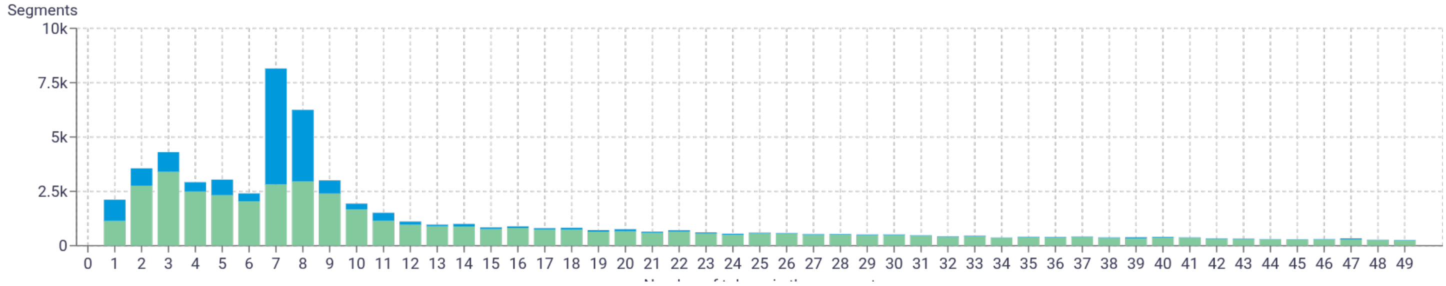
score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (4.7K documents)



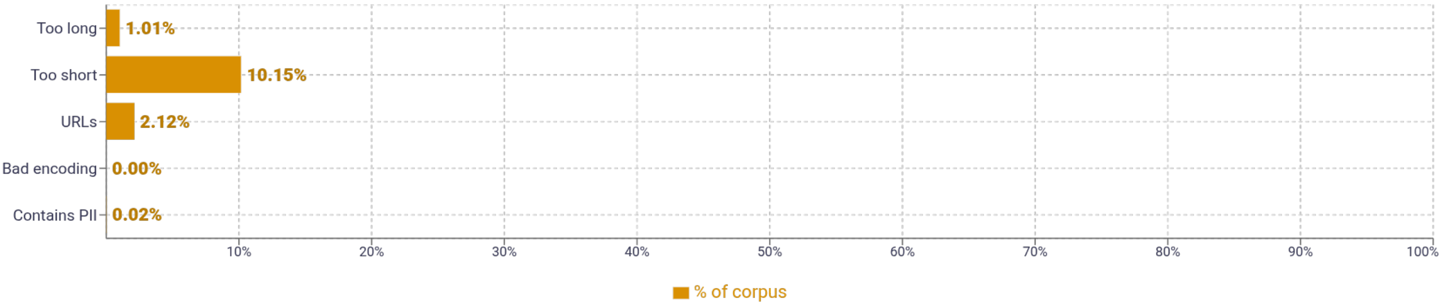
Segment length distribution by token

≤ 49 tokens = **60K** segments | **16K** duplicates

> 50 tokens = **15K** segments | **454** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	[illegible] | 29,460 [illegible] | 14,303 [illegible] | 11,604 [illegible] | 11,527 [illegible] | 10,466
2	[illegible] [illegible] | 14,158 [illegible] [illegible] | 1,688 [illegible] [illegible] | 1,613 from the | 1,502 the original | 1,360
3	from the original | 1,354 archived from the | 1,344 the original on | 1,199 [illegible] [illegible] [illegible] | 473 [illegible] [illegible] [illegible] | 431
4	archived from the original | 1,344 from the original on | 1,199 [illegible] [illegible] [illegible] [illegible] | 425 [illegible] [illegible] [illegible] [illegible] | 256 [illegible] [illegible] [illegible] [illegible] | 239
5	archived from the original on | 1,191 [illegible] [illegible] [illegible] [illegible] [illegible] | 360 [illegible] [illegible] [illegible] [illegible] [illegible] | 230 text was provided for refs | 223 no text was provided for | 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.	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				