

General overview

Corpus	Date	Language
hplt-v3-pan_Guru	9/18/2025	Panjabi

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,518,454	22,164,763	16,691,920 (75.31 %)	939M	4,258,101,914	9.97 GB

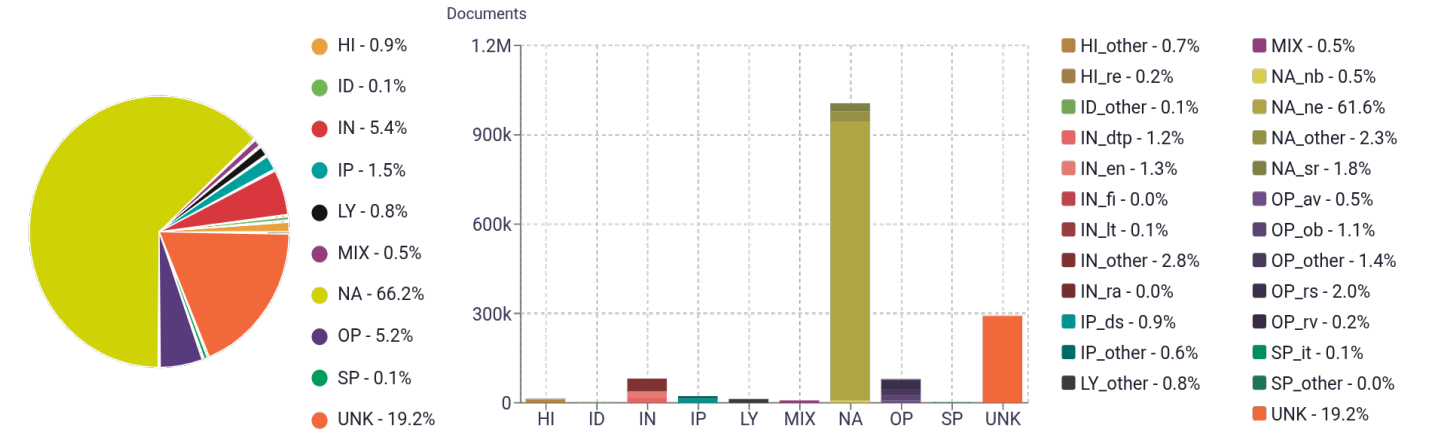
Top 10 domains

Domain	Docs	% of total
<a href="#">punjabkesari.in</a>	62K	4.11%
<a href="#">dailypost.in</a>	50K	3.26%
<a href="#">punjabijagran.com</a>	45K	2.95%
<a href="#">news18.com</a>	39K	2.57%
<a href="#">punjabitribuneo...</a>	37K	2.42%
<a href="#">ajitjalandhar.com</a>	29K	1.91%
<a href="#">ptcpunjabi.co.in</a>	27K	1.80%
<a href="#">hindustantimes.com</a>	26K	1.72%
<a href="#">ptcnews.tv</a>	23K	1.53%
<a href="#">punjabmailusa.com</a>	22K	1.48%

Top 10 TLDs

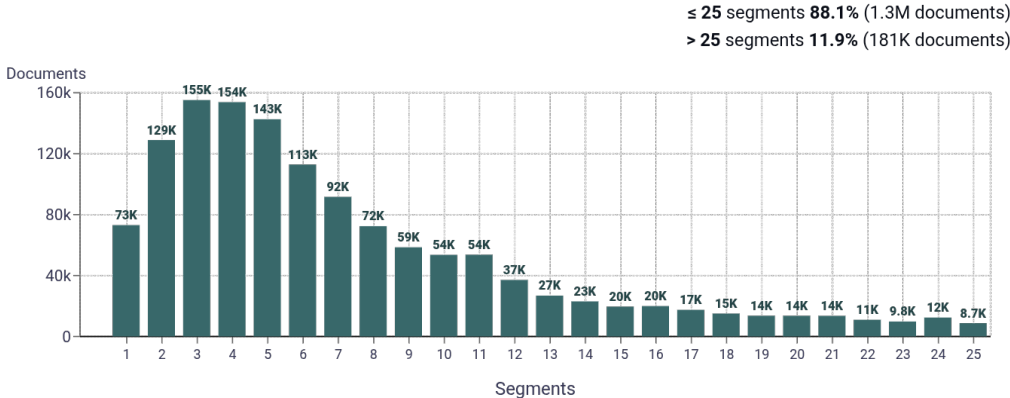
Domain	Docs	% of total
com	976K	64.26%
in	244K	16.05%
org	94K	6.22%
ca	34K	2.23%
co.in	30K	1.98%
tv	26K	1.70%
net	24K	1.58%
info	11K	0.70%
news	8K	0.53%
com.au	7.3K	0.48%

Register labels

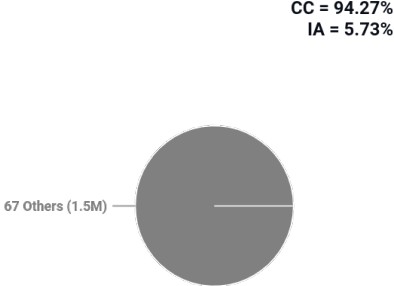


MT:15.1% | 229K Documents

Documents size (in segments) ⓘ

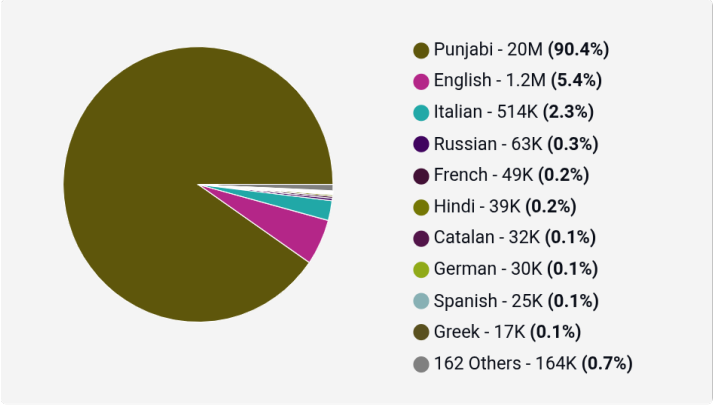


Document collections

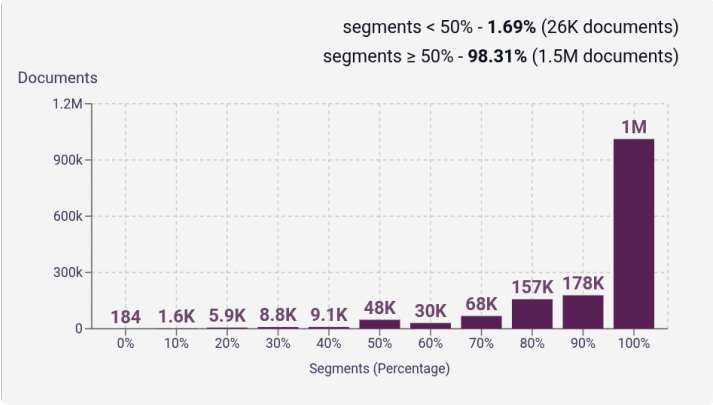


Language Distribution

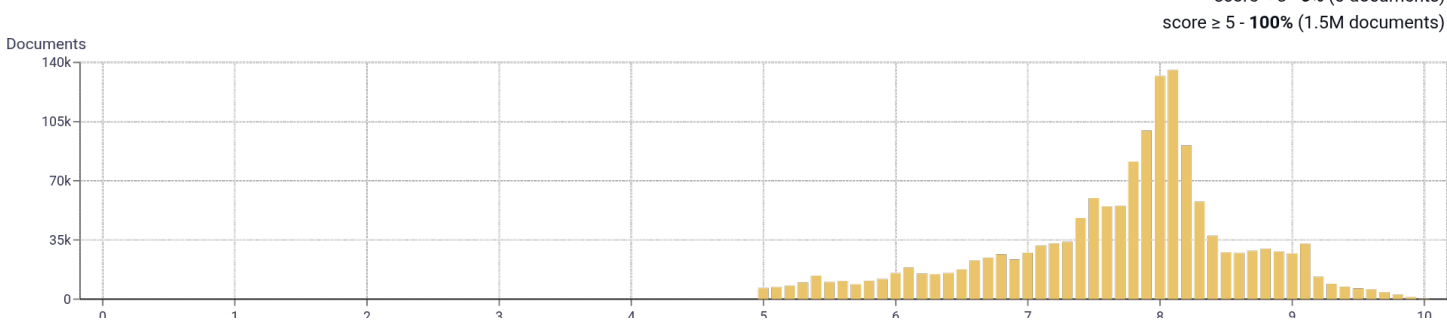
Number of segments in the Panjabi corpus



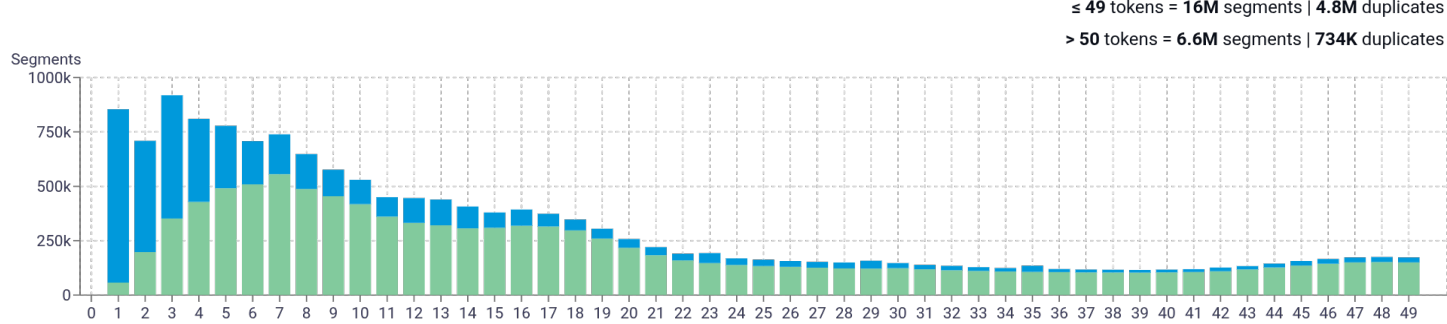
Percentage of segments in Panjabi inside documents



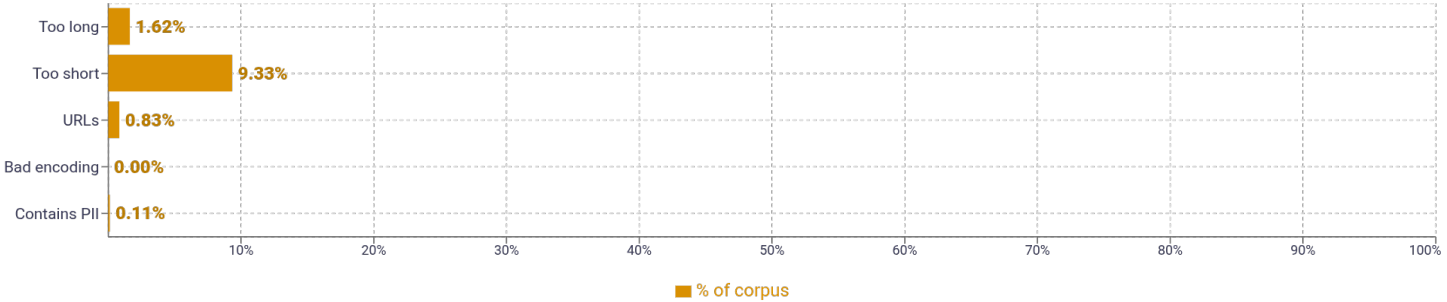
Distribution of documents by document score








Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	ਤੁਸੀਂ   1,472,341   ਰਾ   1,386,752   'ਤੇ   1,353,468   'ਚ   1,291,013   ਸੋ   1,173,454	
2	ਅਕਾਲੀ ਦਲ   249,886   read more   193,768   ਸੋਸ਼ਲ ਮੀਡੀਆ   162,462   ਕਰੋੜ ਰੁਪਏ   137,364   ਕੈਪਟਨ ਅਮਰਿੰਦਰ   113,911	
3	ਸ਼੍ਰੋਮਣੀ ਅਕਾਲੀ ਦਲ   109,750   sadhu singh hamdard   78,686   singh hamdard trust   78,682   ਗੁਰਦੁਆਰਾ ਪ੍ਰਬੰਧਕ ਕਮੇਟੀ   61,666   ਵਿਧਾਨ ਸਭਾ ਚੋਣਾਂ   52,982	
4	sadhu singh hamdard trust   78,682   ਸ਼੍ਰੋਮਣੀ ਗੁਰਦੁਆਰਾ ਪ੍ਰਬੰਧਕ ਕਮੇਟੀ   37,127   whole or in part   26,236   in whole or in   26,236   without the prior written   26,228	
5	in whole or in part   26,236   the prior written consent of   26,228   the ajit logo is copyright   26,228   without the prior written consent   26,227   written consent of the trust   26,226	

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 of 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				