

General overview

Corpus	Date	Language
hplt-v3-fi_Latn	10/3/2025	Filipino (fil)

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
3,441,238	83,672,999	58,869,555 (70.36 %)	29.64%	2.6B	14,038,567,730	13.14 GB

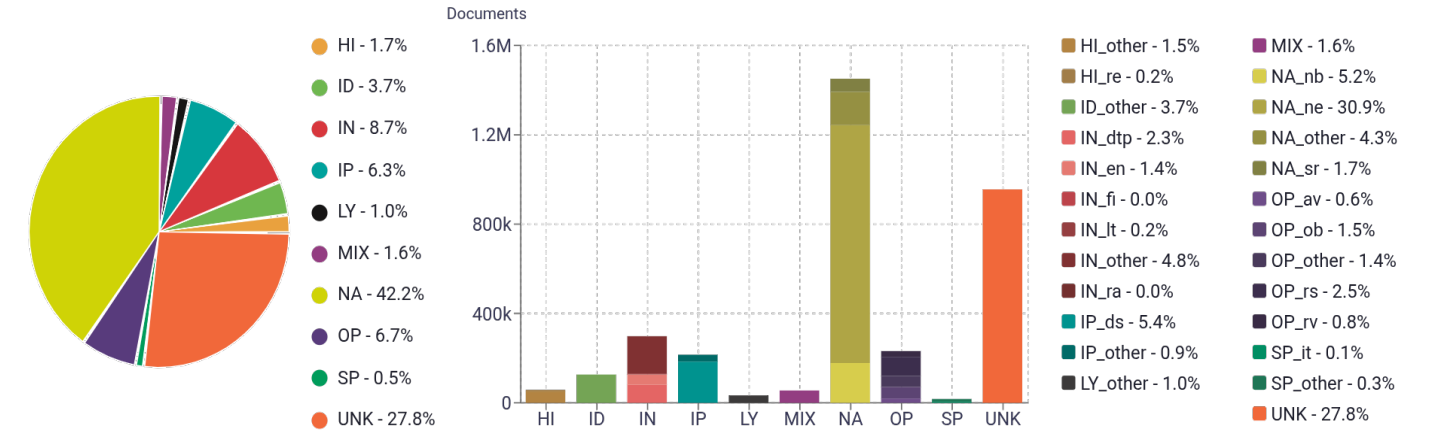
Top 10 domains

Domain	Docs	% of total
blogspot.com	108K	3.13%
abante.com.ph	103K	2.99%
remate.ph	83K	2.42%
dwiz882am.com	72K	2.11%
rmn.ph	68K	1.97%
wordpress.com	66K	1.92%
hatawtabloid.com	50K	1.46%
abs-cbn.com	46K	1.34%
pinoyparazzi.com	45K	1.32%
pep.ph	45K	1.30%

Top 10 TLDs

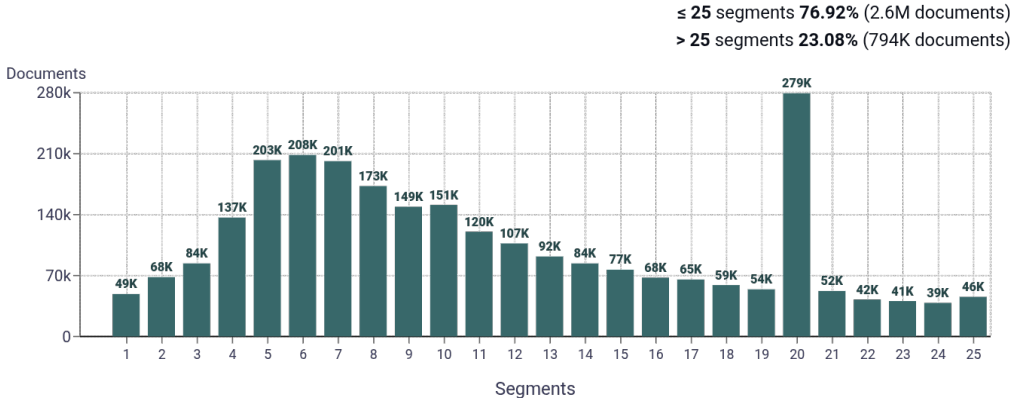
Domain	Docs	% of total
com	1.8M	51.41%
ph	314K	9.13%
com.ph	244K	7.09%
ru	219K	6.35%
org	185K	5.37%
net	180K	5.24%
tk	52K	1.51%
gov.ph	38K	1.11%
pl	35K	1.02%
info	34K	0.98%

Register labels



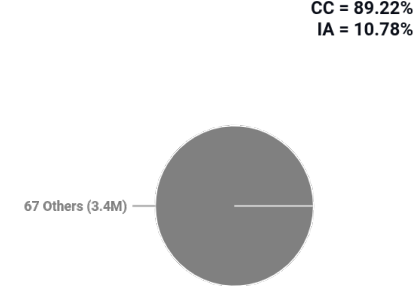
MT:24.4% | 840K Documents

Documents size (in segments)



≤ 25 segments 76.92% (2.6M documents)
> 25 segments 23.08% (794K documents)

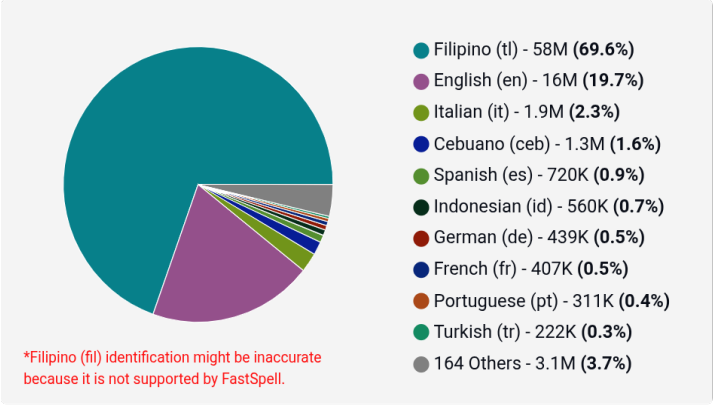
Document collections



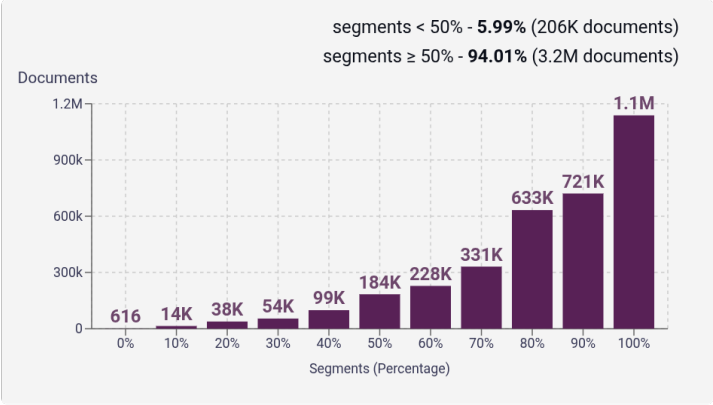
CC = 89.22%
IA = 10.78%

Language Distribution

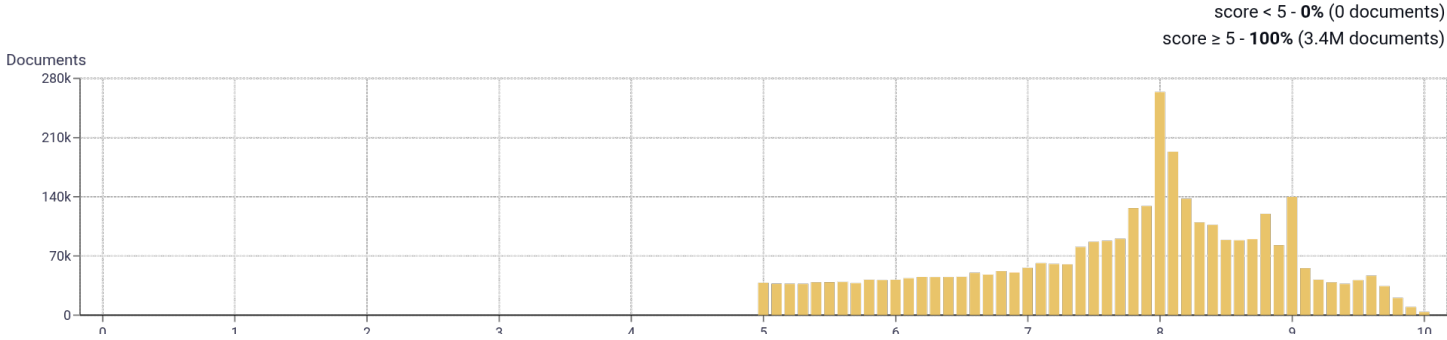
Number of segments in the Filipino (fil) corpus



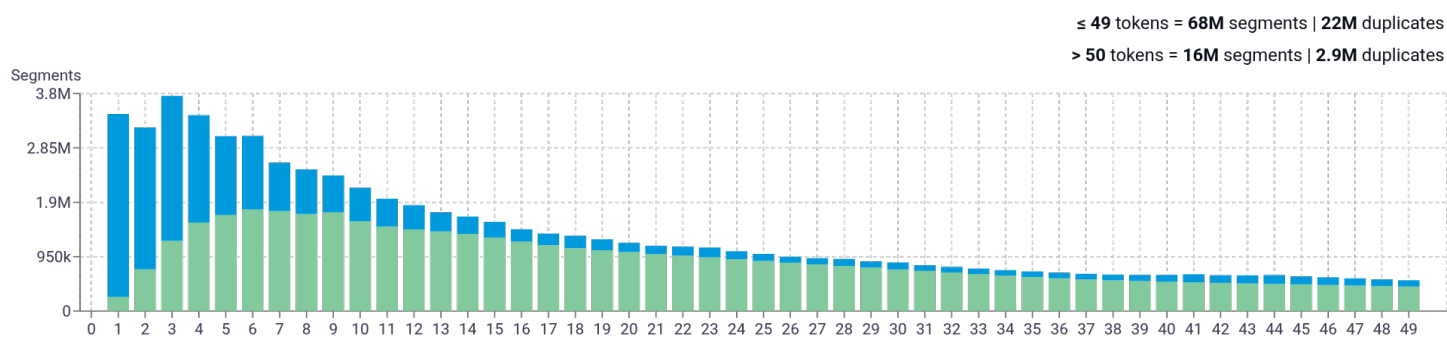
Percentage of segments in Filipino (fil) inside documents



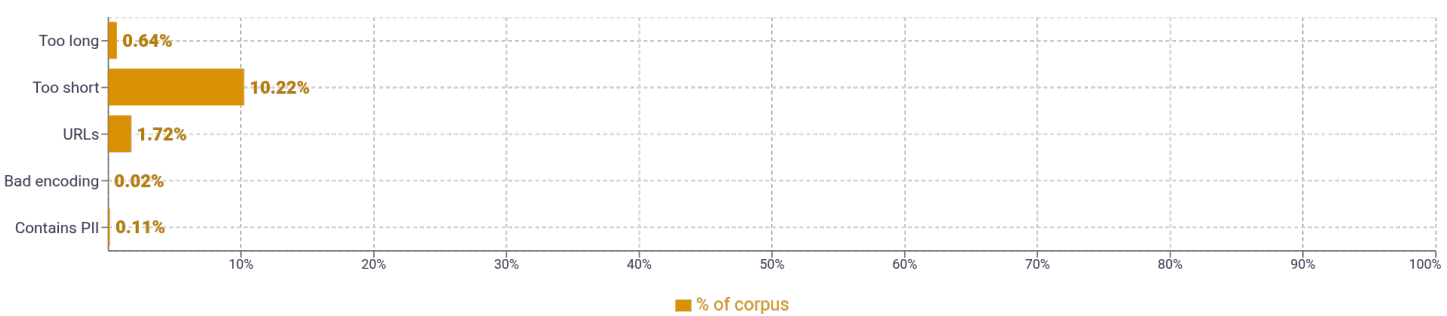
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	upang 8,496,889the 8,261,316si 7,760,934mo 7,686,659lang 5,980,947	
2	kuko halamang-singaw 1,097,612of the 1,057,891't ibang 669,474in the 626,465t ibang 588,785	
3	pagbaba ng timbang 696,113mawalan ng timbang 467,121pati na rin 386,124maliit na suso 315,108paggamot ng kuko 234,477	
4	mask para sa buhok 243,036paggamot ng kuko halamang-singaw 187,164langis para sa buhok 167,999upang mawala ang timbang 148,691paligid ng mga mata 117,047	
5	bags sa ilalim ng mata 116,288halamang-singaw sa kanyang mga paa 112,779circles sa ilalim ng mata 95,578serye ng hst na solong 70,857sistema na hinihimok ng haydroliko 70,716	

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 *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>, , , 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				