HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-war_Latn	9/18/2025	Waray (war)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
9,350	119,195	103,742 (87.04 %)	5.1M	25,305,511	24.28 MB

Top 10 domains

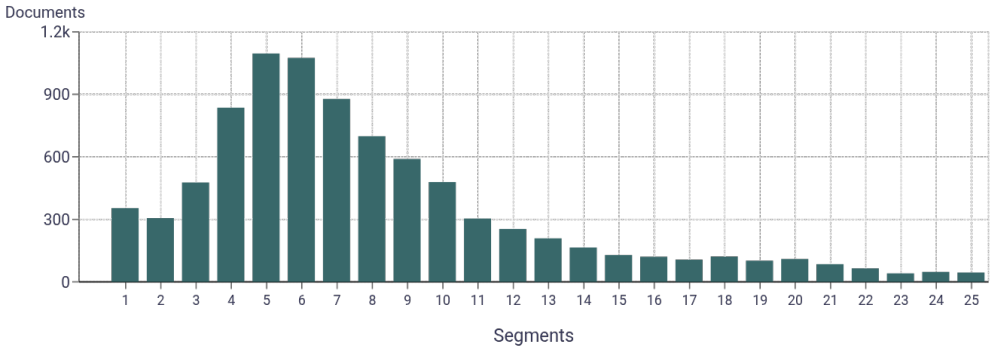
Domain	Docs	% of total
wikipedia.org	2.3K	24.28%
isumat.com	1.6K	16.70%
bomboradyo.com	1.5K	16.10%
pia.gov.ph	926	9.90%
jw.org	913	9.76%
rmn.ph	312	3.34%
bible.is	308	3.29%
wordpress.com	183	1.96%
tacloban.gov.ph	133	1.42%
blogspot.com	119	1.27%

Top 10 TLDs

Domain	Docs	% of total
com	3.6K	38.91%
org	3.3K	35.66%
gov.ph	1.1K	12.27%
ph	490	5.24%
is	308	3.29%
de	133	1.42%
net	46	0.49%
pl	45	0.48%
nu	42	0.45%
click	32	0.34%

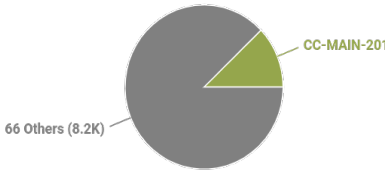
Documents size (in segments)

≤ 25 segments **93.02%** (8.7K documents)
> 25 segments **6.98%** (653 documents)



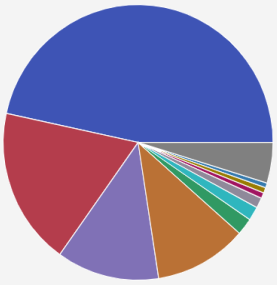
Document collections

CC = 92.37%
IA = 7.63%



Language Distribution

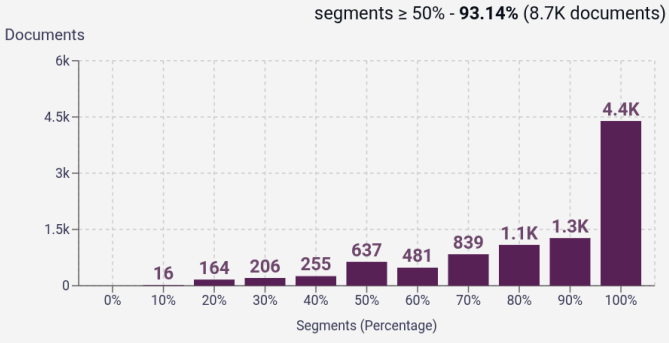
Number of segments in the Waray (war) corpus



- Waray (war) - 56K **(46.6%)**
- Filipino (tl) - 22K **(18.5%)**
- English (en) - 15K **(12.3%)**
- Cebuano (ceb) - 13K **(11.1%)**
- Spanish (es) - 2.4K **(2.0%)**
- Italian (it) - 1.9K **(1.6%)**
- Indonesian (id) - 1.4K **(1.2%)**
- Iloko (ilo) - 794 **(0.7%)**
- Croatian (hr) - 767 **(0.6%)**
- Portuguese (pt) - 670 **(0.6%)**
- 118 Others - 5.7K **(4.8%)**

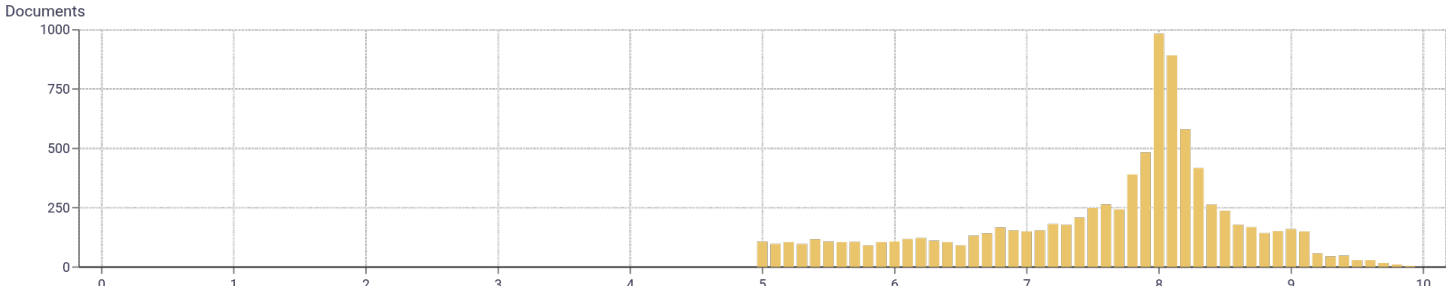
Percentage of segments in Waray (war) inside documents

segments < 50% - **6.86%** (641 documents)
segments ≥ 50% - **93.14%** (8.7K documents)



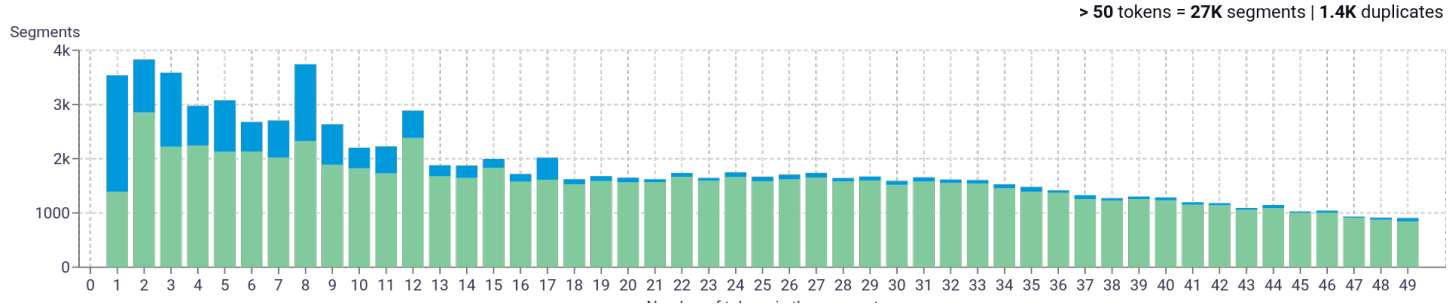
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (9.4K documents)

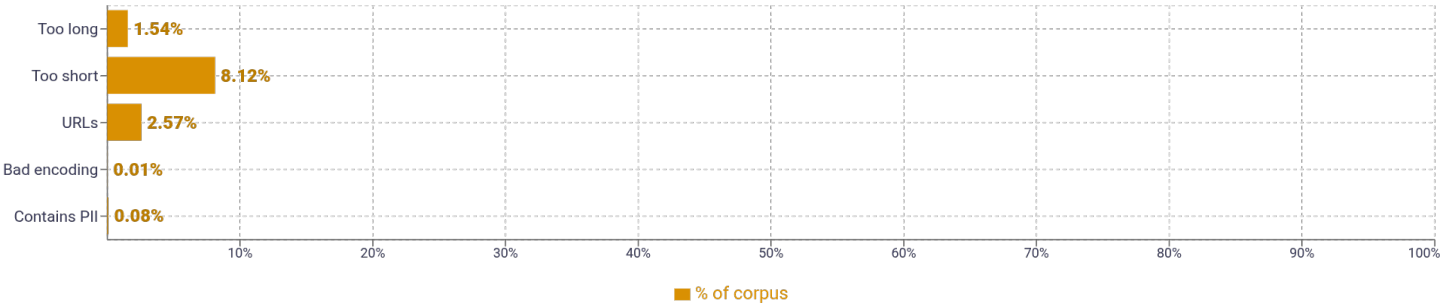


Segment length distribution by token

≤ 49 tokens = **92K** segments | **14K** duplicates
> 50 tokens = **27K** segments | **1.4K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	an | 203,333 · han | 198,672 · ha | 154,863 · mga | 122,993 · hin | 59,977
2	han mga | 21,417 · an mga | 20,677 · ha mga | 9,335 · ha iya | 8,266 · han iya | 7,799
3	an mga tawo | 1,399 · an iya mga | 1,322 · han mga tawo | 1,266 · igliwat an wikitext | 1,264 · ini nga mga | 1,091
4	hinigugma ko na mga | 655 · mga saksi ni jehova | 521 · salamat han iyo pag | 509 · bersyon nga angay ighubad | 467 · artikulo nga aada han | 466
5	hinigugma ko na mga anak | 654 · bersyon nga angay ighubad ha | 467 · mayda impormasyon hini nga artikulo | 461 · hini nga artikulo nga aada | 461 · iningles nga bersyon nga angay | 399

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				