HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-ceb_Latn	9/17/2025	Cebuano

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
354,235	6,777,731	5,305,947 (78.29 %)	249M	1,251,036,410	1.18 GB

Top 10 domains

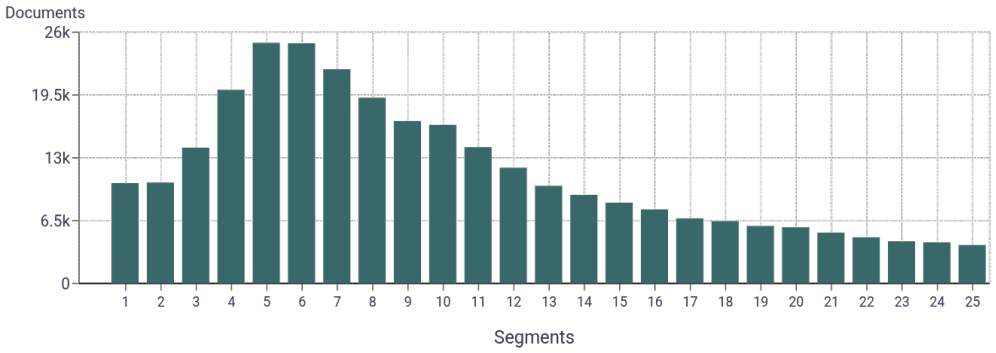
Domain	Docs	% of total
sunstar.com.ph	27K	7.50%
rmn.ph	18K	5.02%
philstar.com	14K	4.03%
jw.org	9.1K	2.56%
biblica.com	7.8K	2.20%
martech.zone	7.6K	2.14%
bomboradyo.com	6.4K	1.81%
rpnradio.com	5.8K	1.63%
aksyonradyoiloi...	5.3K	1.50%
blogspot.com	5.1K	1.44%

Top 10 TLDs

Domain	Docs	% of total
com	212K	59.87%
com.ph	33K	9.34%
ph	23K	6.56%
org	23K	6.51%
gov.ph	12K	3.50%
zone	7.6K	2.14%
net	7.5K	2.12%
news	4.1K	1.15%
is	2.3K	0.64%
ru	2.2K	0.61%

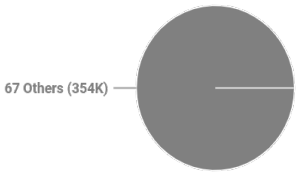
Documents size (in segments)

≤ 25 segments **81.35%** (288K documents)
> 25 segments **18.65%** (66K documents)



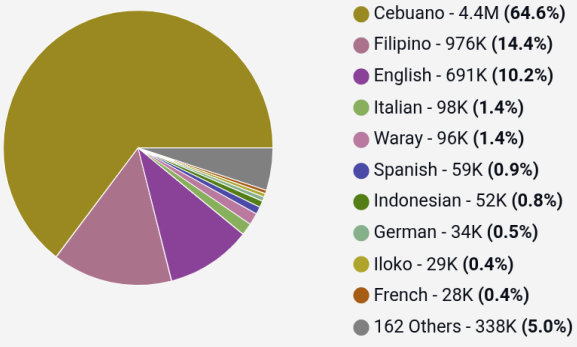
Document collections

CC = 95.66%
IA = 4.34%



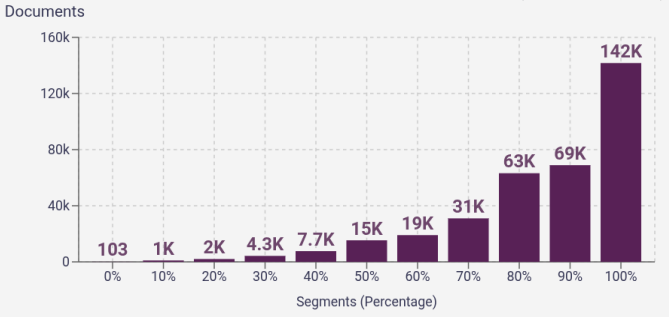
Language Distribution

Number of segments in the Cebuano corpus



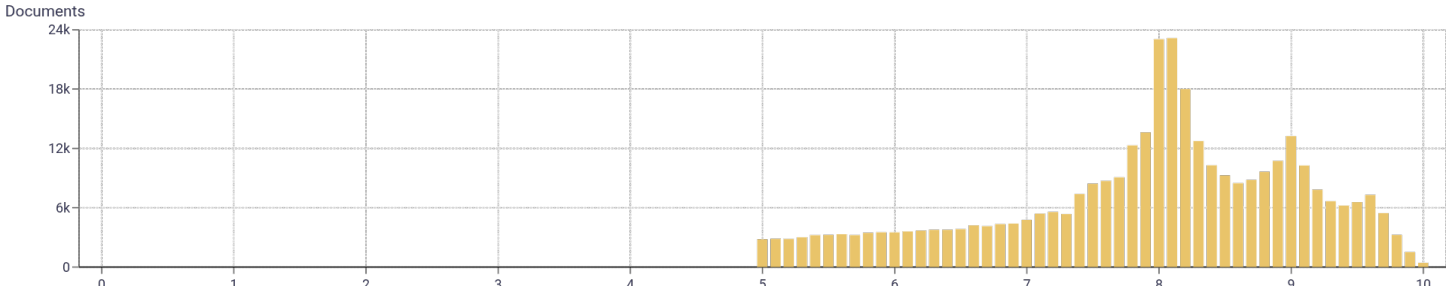
Percentage of segments in Cebuano inside documents

segments < 50% - **4.26%** (15K documents)
segments ≥ 50% - **95.74%** (339K documents)



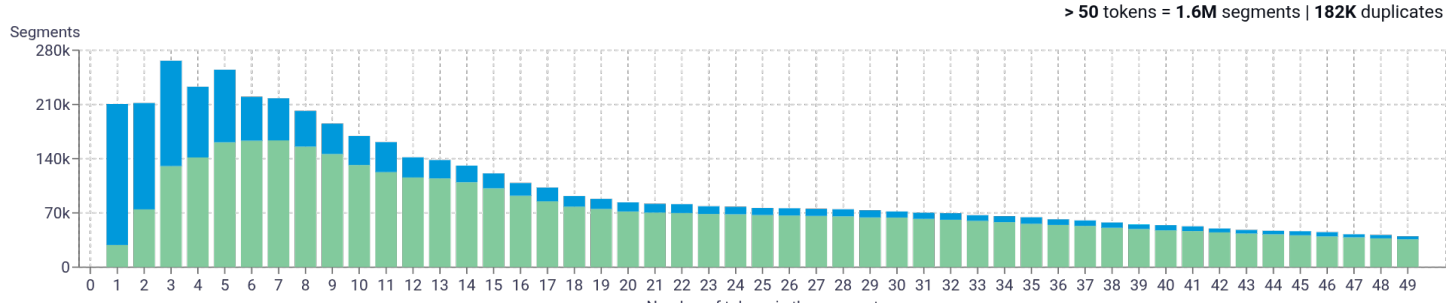
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (354K documents)

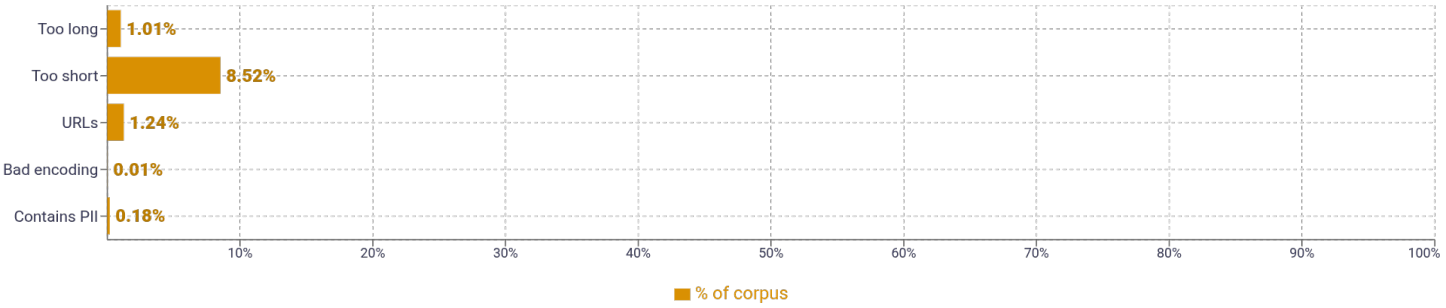


Segment length distribution by token

≤ 49 tokens = **5.2M** segments | **1.3M** duplicates
> 50 tokens = **1.6M** segments | **182K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	ka | 3,634,806 · pag | 1,001,396 · alang | 978,374 · adunay | 846,295 · pa | 789,356
2	duha ka | 118,951 · labi ka | 103,749 · ka tuig | 97,039 · bisan pa | 91,518 · wala pa | 84,607
3	adunay usa ka | 100,121 · ingon usa ka | 41,476 · og usa ka | 34,346 · isip usa ka | 18,631 · mao ang labing | 17,016
4	alang sa usa ka | 57,620 · mao ang usa ka | 39,902 · ngadto sa usa ka | 15,669 · paghimo sa usa ka | 14,687 · ingon sa usa ka | 11,139
5	himoa ang una nga makomentaryo | 7,883 · us aka us aka us | 5,516 · aka us aka us aka | 5,512 · palihug sa pagsangyaw ug ayaw | 4,809 · naghimo niini nga usa ka | 4,070

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				