HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-ilo_Latn	9/18/2025	Iloko

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
43,850	849,837	642,950 (75.66 %)	26M	134,062,555	128.74 MB

Top 10 domains

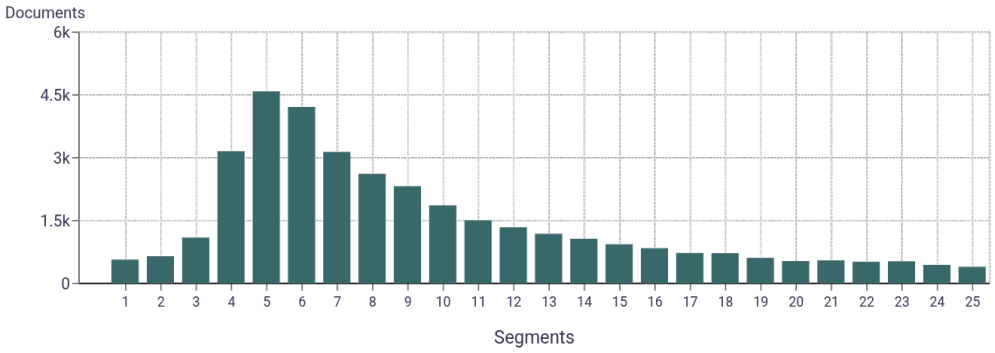
Domain	Docs	% of total
tawidnewsmag.com	12K	27.56%
wikipedia.org	9.5K	21.76%
bomboradyo.com	5.5K	12.49%
jw.org	4.2K	9.68%
blogspot.com	2.4K	5.40%
mb.com.ph	2K	4.62%
breakeveryyoke.com	777	1.77%
wordpress.com	703	1.60%
rpnradio.com	548	1.25%
pressreader.com	372	0.85%

Top 10 TLDs

Domain	Docs	% of total
com	24K	55.08%
org	15K	33.27%
com.ph	2.1K	4.78%
gov.ph	692	1.58%
net	534	1.22%
nu	339	0.77%
ph	314	0.72%
is	282	0.64%
online	260	0.59%
gov	102	0.23%

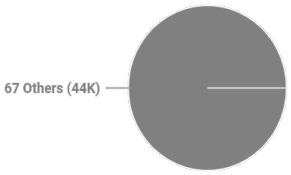
Documents size (in segments)

≤ 25 segments **82.32%** (36K documents)
> 25 segments **17.68%** (7.8K documents)



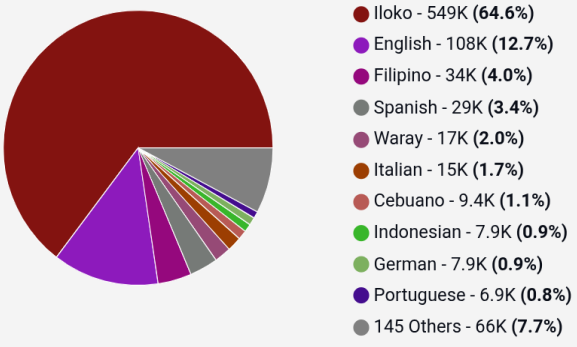
Document collections

CC = 96.17%
IA = 3.83%



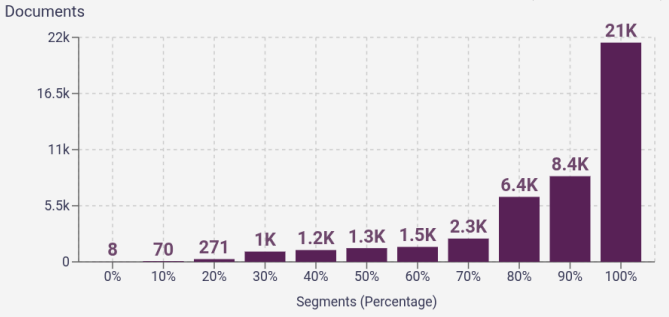
Language Distribution

Number of segments in the Iloko corpus



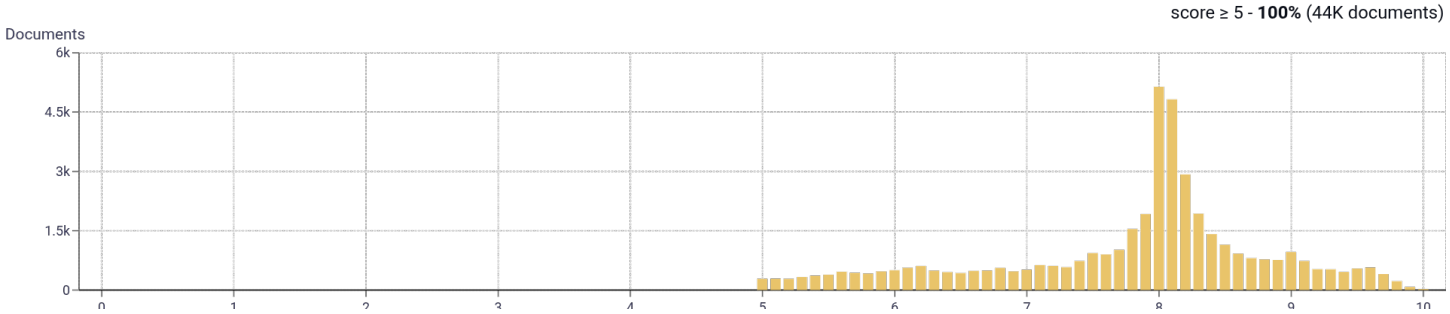
Percentage of segments in Iloko inside documents

segments < 50% - **5.74%** (2.5K documents)
segments ≥ 50% - **94.26%** (41K documents)



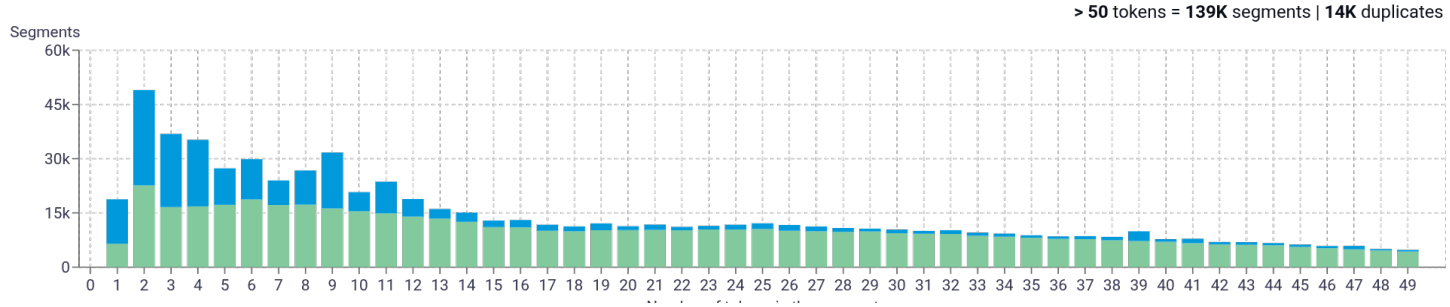
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (44K documents)

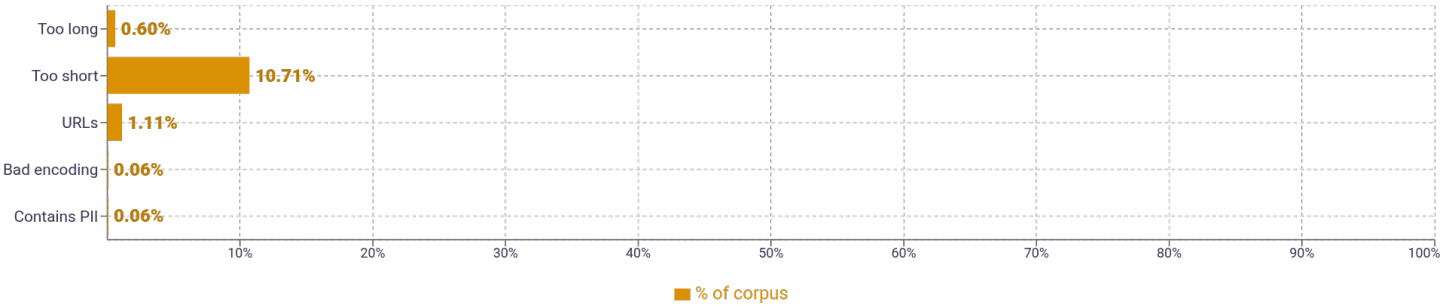


Segment length distribution by token

≤ 49 tokens = **711K** segments | **193K** duplicates
> 50 tokens = **139K** segments | **14K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	ket | 291,111 · kadagiti | 193,237 · pay | 108,318 · met | 101,779 · ta | 92,498
2	ilocos sur | 18,896 · gapu ta | 11,734 · dadduma pay | 11,336 · of the | 9,792 · para kadagiti | 8,915
3	urnosen ti taudan | 23,171 · comments are closed | 5,776 · ditoy nga ili | 3,724 · residente iti barangay | 3,226 · datos ti populasion | 2,715
4	populasion babaen ti probinsia | 2,112 · dagup ti populasion babaen | 2,110 · ket maysa a maika | 1,312 · ket nabingbingay a politikal | 1,235 · kadagiti ili ken siudad | 989
5	klase nga ili iti probinsia | 1,703 · pagalagadan a kodigo ti heograpia | 1,338 · datos ti populasion ti lwua | 1,337 · baro a lubong a patarus | 1,145 · opisial a resulta ti panagbutos | 1,071

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				