HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-quy_Latn	9/18/2025	Quechua (qu)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
20,199	565,117	444,585 (78.67 %)	17M	113,453,667	110.43 MB

Top 10 domains

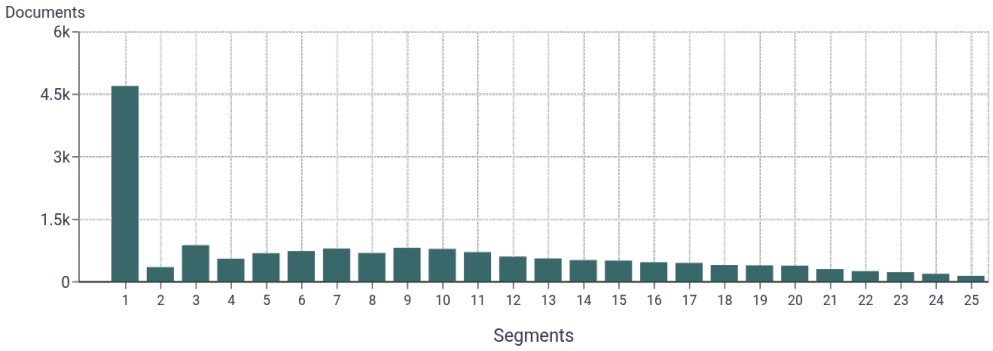
Domain	Docs	% of total
bible.is	4.6K	22.60%
wikipedia.org	4.3K	21.12%
jw.org	4.1K	20.36%
ebible.org	1.3K	6.36%
policiaecuador....	674	3.34%
breakeveryyoke.com	555	2.75%
ladecana.pe	552	2.73%
policia.gob.ec	324	1.60%
asambleanaciona...	262	1.30%
bibles.org	241	1.19%

Top 10 TLDs

Domain	Docs	% of total
org	11K	52.77%
is	4.6K	22.60%
com	1.5K	7.62%
gob.ec	1.5K	7.59%
pe	846	4.19%
gob.pe	288	1.43%
net	273	1.35%
ru	51	0.25%
com.ec	36	0.18%
de	34	0.17%

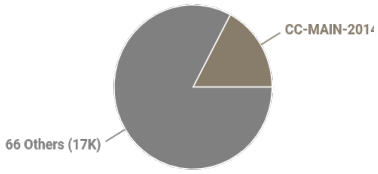
Documents size (in segments)

≤ 25 segments **85.03%** (17K documents)
> 25 segments **14.97%** (3K documents)



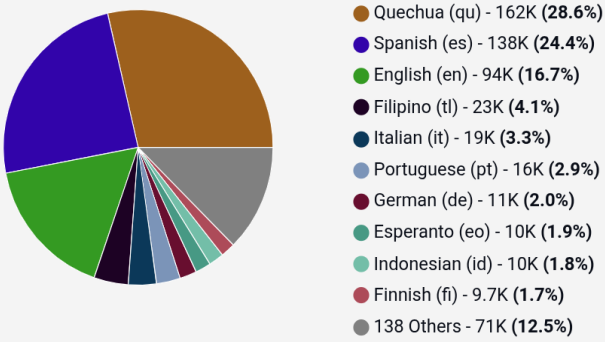
Document collections

CC = 89.83%
IA = 10.17%



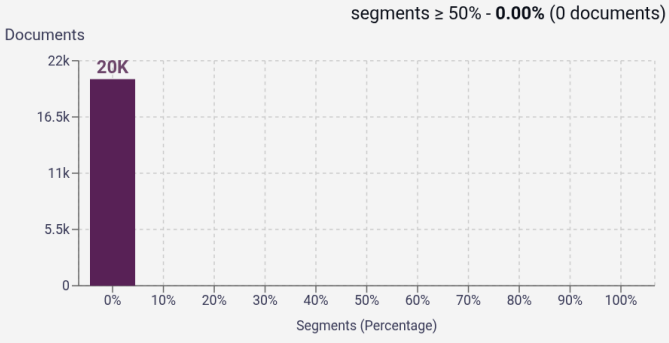
Language Distribution

Number of segments in the Quechua (qu) corpus



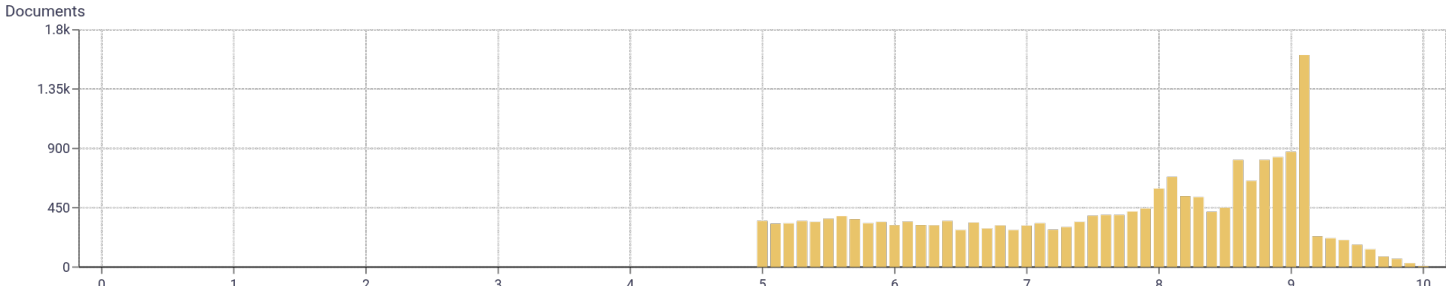
Percentage of segments in Quechua (qu) inside documents

segments < 50% - **100.00%** (20K documents)
segments ≥ 50% - **0.00%** (0 documents)



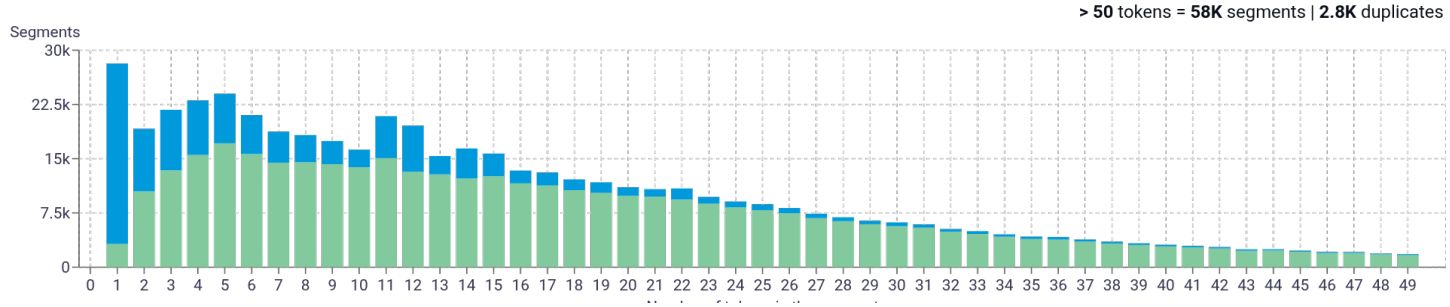
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (20K documents)

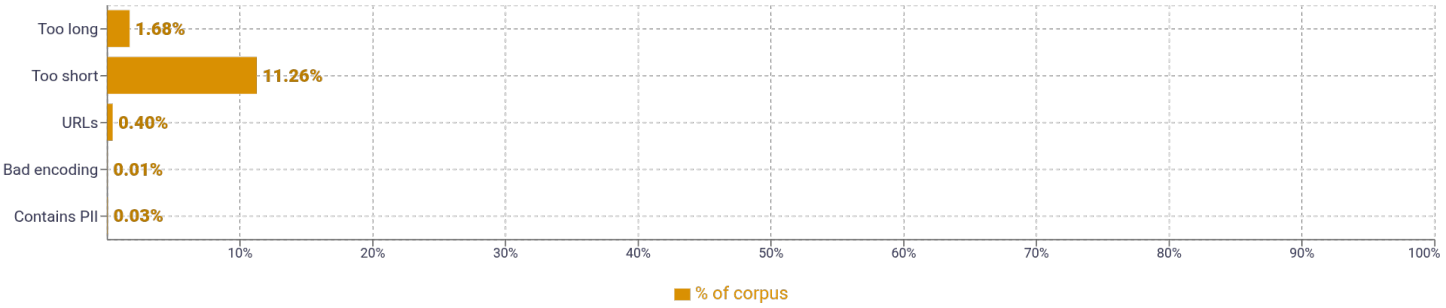


Segment length distribution by token

≤ 49 tokens = **507K** segments | **118K** duplicates
> 50 tokens = **58K** segments | **2.8K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	mana | 199,216 · chay | 148,413 · kay | 78,504 · diospa | 53,691 · tayta | 53,487
2	pukyuta llamk | 11,115 · tayta dios | 8,165 · tata dios | 6,324 · yaya diospa | 6,116 · mana alli | 6,081
3	nisqaqa multimidya kapuyninkunayuqmi | 1,904 · multimidya kapuyninkunayuqmi kay | 1,904 · kapuyninkunayuqmi kay hawa | 1,904 · commons nisqaqa multimidya | 1,904 · tayta diosninchipa shiminta | 691
4	nisqaqa multimidya kapuyninkunayuqmi kay | 1,904 · multimidya kapuyninkunayuqmi kay hawa | 1,904 · commons nisqaqa multimidya kapuyninkunayuqmi | 1,904 · suyukunata uyarinakunatapas tarinki kaymantam | 484 · nisqapi suyukunata uyarinakunatapas tarinki | 484
5	nisqaqa multimidya kapuyninkunayuqmi kay hawa | 1,904 · commons nisqaqa multimidya kapuyninkunayuqmi kay | 1,904 · nisqapi suyukunata uyarinakunatapas tarinki kaymantam | 484 · commons nisqapi suyukunata uyarinakunatapas tarinki | 484 · simi icha huk indihina simi | 391

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				