

General overview

Corpus	Date	Language
hplt-v3-quy_Latn	9/24/2025	Ayacucho Quechua (quy)

Volumes

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

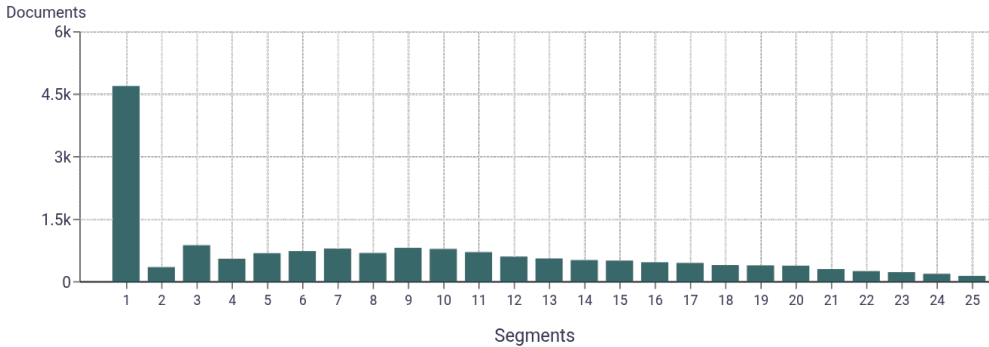
Top 10 domains

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%

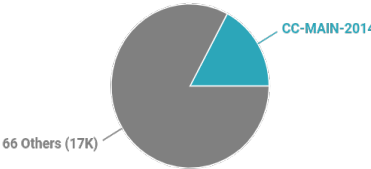
Documents size (in segments) ⓘ



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

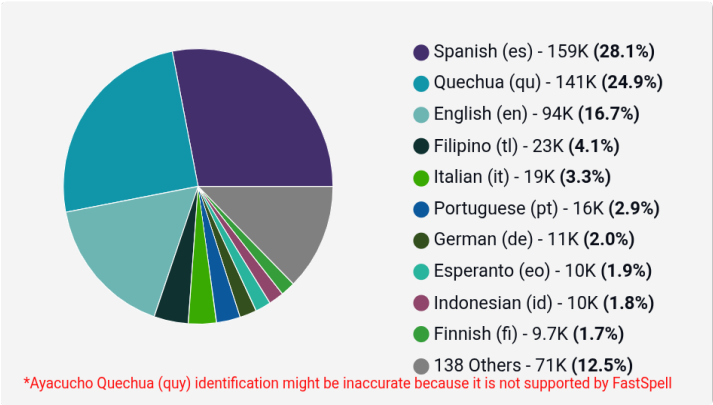
Document collections

CC = 89.83%
IA = 10.17%

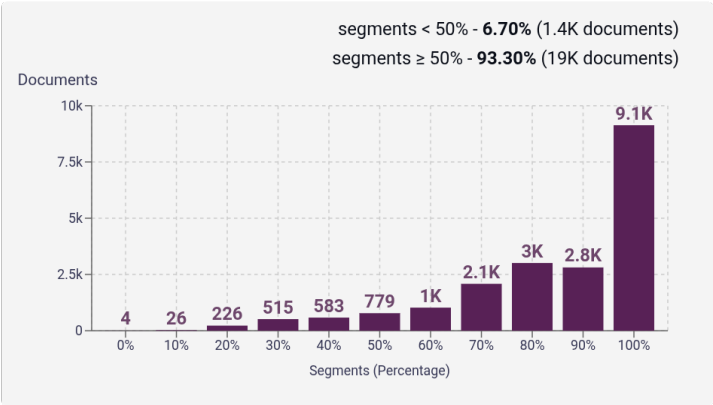


Language Distribution

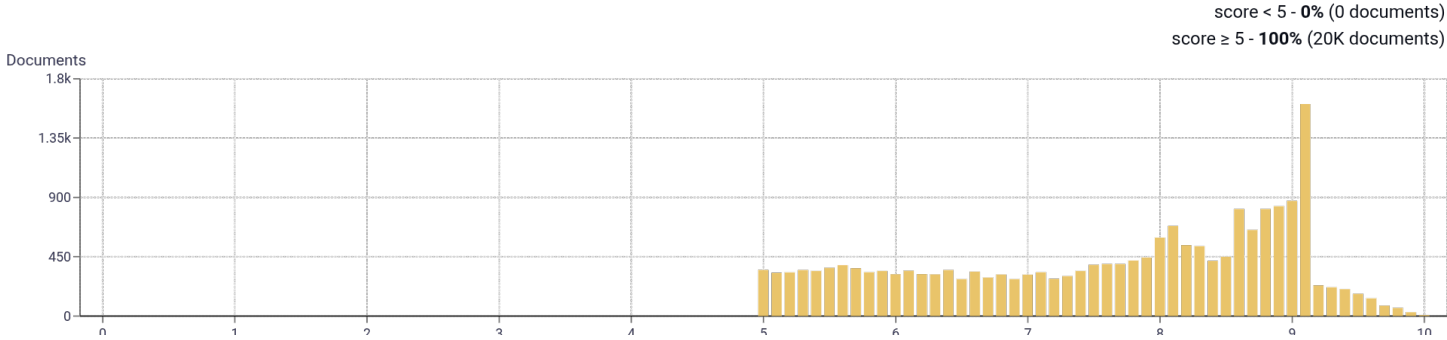
Number of segments in the Ayacucho Quechua (quy) corpus



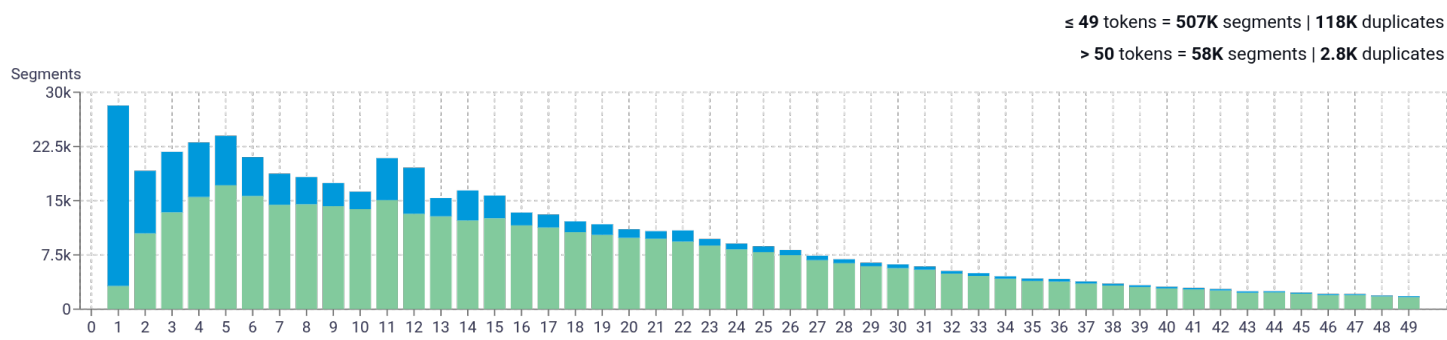
Percentage of segments in Ayacucho Quechua (quy) inside documents



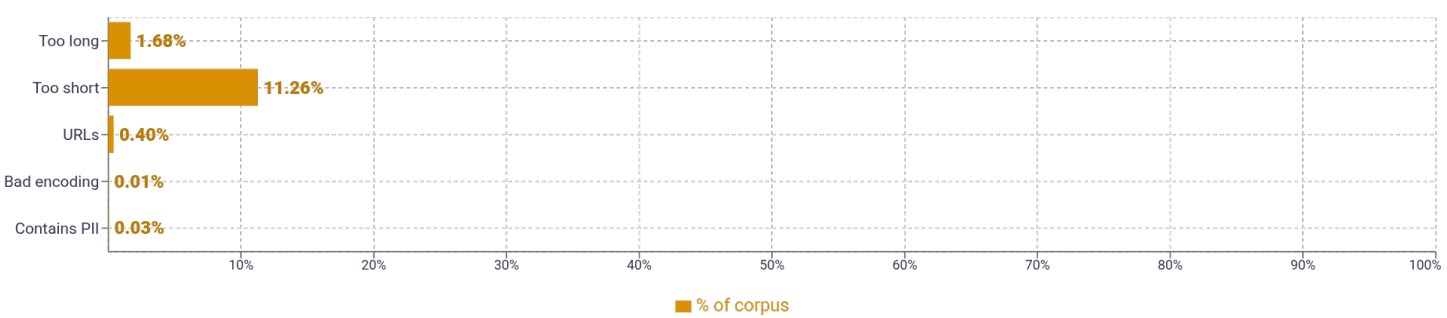
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>diospa 53,691</div> <div>tayta 53,487</div> <div>tukuy 50,238</div> <div>ima 48,943</div> <div>y 48,176</div>	
2	<div>pukyuta llamk 11,115</div> <div>tayta dios 8,165</div> <div>tata dios 6,324</div> <div>yaya diospa 6,116</div> <div>de la 5,396</div>	
3	<div>nisqaqa multimidya kapuyninkunayuqmi 1,904</div> <div>kapuyninkunayuqmi kay hawa 1,904</div> <div>commons nisqaqa multimidya 1,904</div> <div>tayta diosninchipa shiminta 691</div> <div>tata diospa rimananta 672</div>	
4	<div>multimidya kapuyninkunayuqmi kay hawa 1,904</div> <div>commons nisqaqa multimidya kapuyninkunayuqmi 1,904</div> <div>suyukunata uyarinakunatapas tarinki kaymantam 484</div> <div>nisqapi suyukunata uyarinakunatapas tarinki 484</div> <div>commons nisqapi suyukunata uyarinakunatapas 484</div>	
5	<div>nisqaqa multimidya kapuyninkunayuqmi kay hawa 1,904</div> <div>nisqapi suyukunata uyarinakunatapas tarinki kaymantam 484</div> <div>commons nisqapi suyukunata uyarinakunatapas tarinki 484</div> <div>simi icha huk indihina simi 391</div> <div>ashaninka simi icha huk indihina 391</div>	

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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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	ntp
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				