

General overview

Corpus	Date	Language
hplt-v3-pap_Latn	9/17/2025	Papiamento

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
181,779	2,414,381	1,819,512 (75.36 %)	97M	462,659,347	446.77 MB

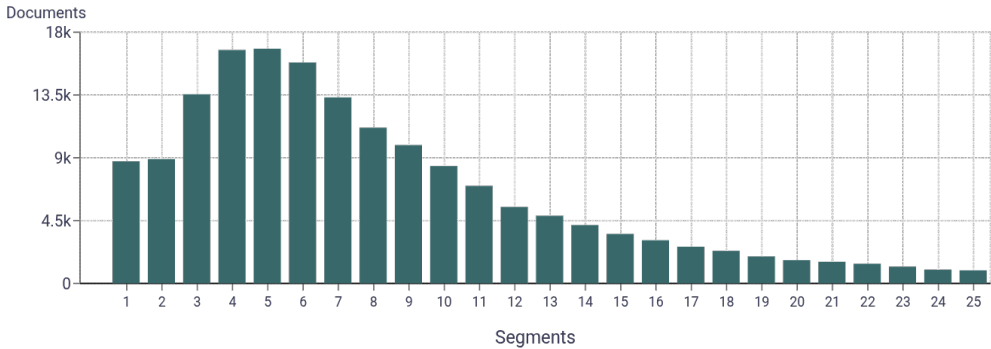
Top 10 domains

Domain	Docs	% of total
diario.aw	16K	8.78%
arubanative.com	16K	8.74%
kikotapasando.com	11K	6.05%
masnoticia.com	9.6K	5.28%
bondia.com	9.5K	5.22%
noticiacila.com	8K	4.40%
awe24.com	6.8K	3.75%
eanews.com	5.8K	3.21%
24ora.com	4.5K	2.45%
awemainta.com	4.3K	2.34%

Top 10 TLDs

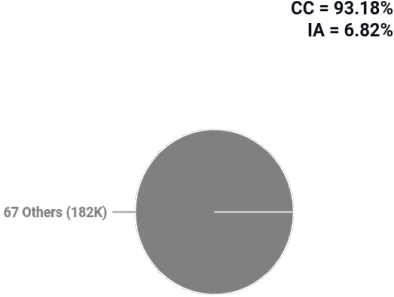
Domain	Docs	% of total
com	125K	68.70%
aw	24K	13.46%
cw	8.8K	4.83%
org	7.7K	4.23%
nl	4.4K	2.40%
news	3.6K	2.00%
nu	3K	1.65%
net	875	0.48%
blog	358	0.20%
today	282	0.16%

Documents size (in segments) ⓘ



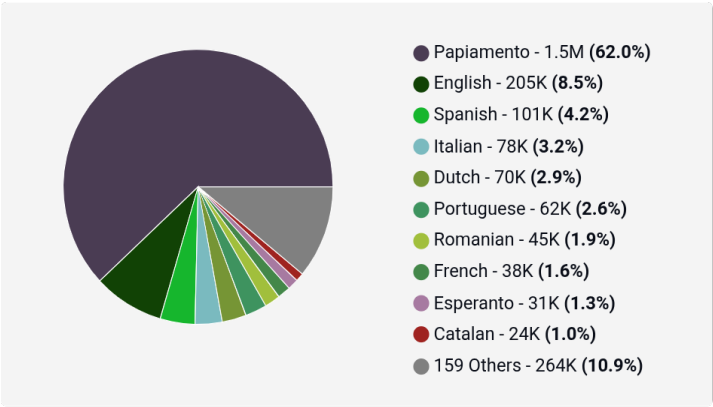
≤ 25 segments **91.42%** (166K documents)
> 25 segments **8.58%** (16K documents)

Document collections

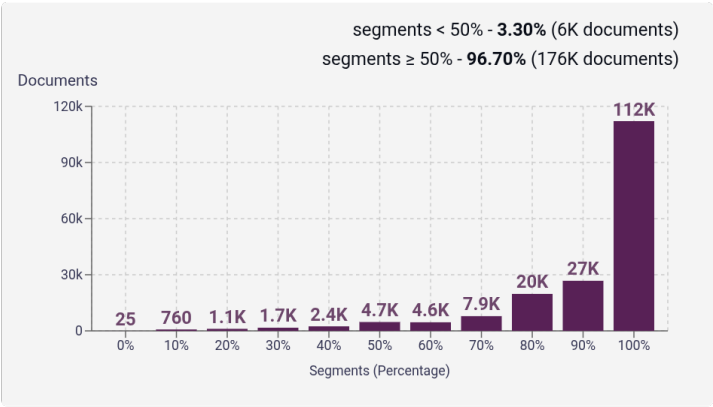


Language Distribution

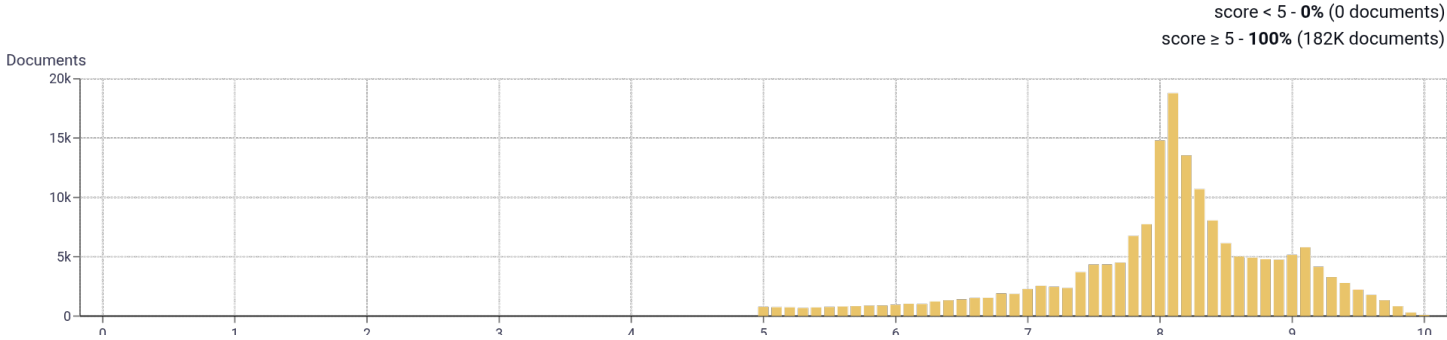
Number of segments in the Papiamento corpus



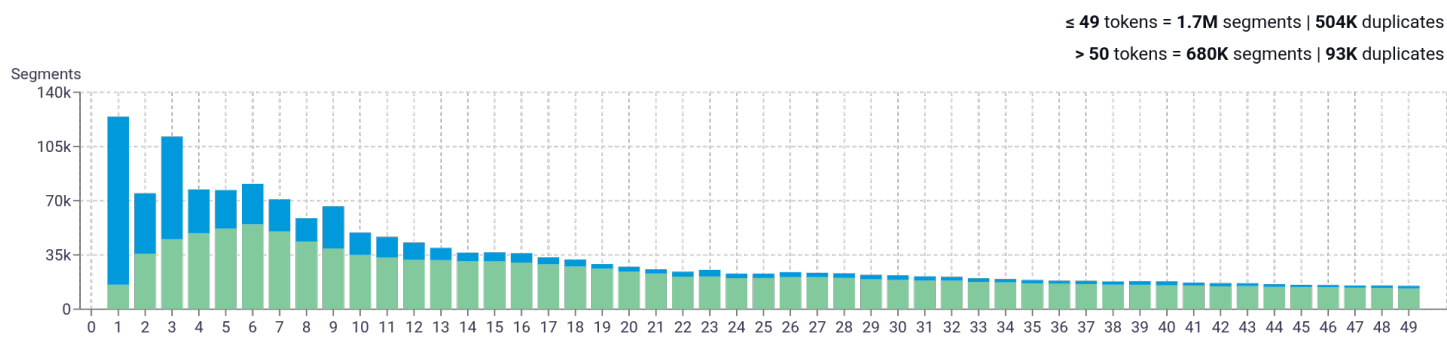
Percentage of segments in Papiamento inside documents



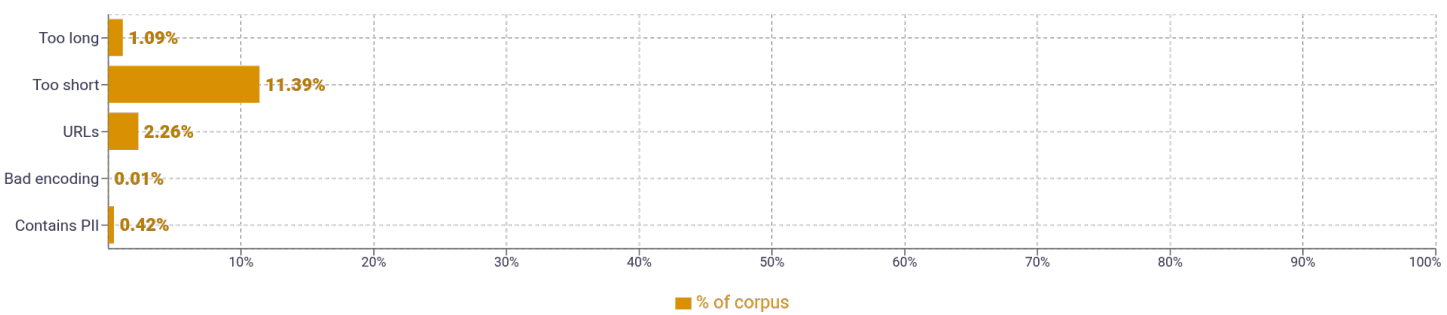
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	pa 2,120,308y 1,479,800na 1,390,732i 740,520nan 715,244	
2	na aruba 73,674pa asina 49,048pa nan 42,229yega na 34,718su mes 34,400	
3	na e sitio 18,031gobierno di aruba 13,407el a bisa 12,706loke ta trata 10,216post a comment 9,787	
4	pa loke ta trata 9,724to post a comment 9,683you must log in 9,415must log in to 9,415log in to post 9,415	
5	in to post a comment 9,682you must log in to 9,415must log in to post 9,415na yegada di e patruya 3,681drenta informe di un accidente 2,001	

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				