HPLT Analytics report

General overview

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

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%
noticiacla.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

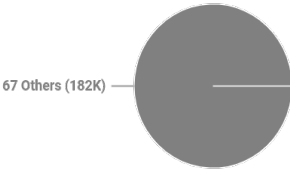| Domain	Docs	% of total
com	125K	68.70%
aw	24K	13.46%
cw	8.8K	4.83%
org	7.7K	4.23%
nl	4.4K	2.40%
news	3.6K	2.00%
nu	3K	1.65%
net	875	0.48%
blog	358	0.20%
today	282	0.16%

Documents size (in segments) ⓘ
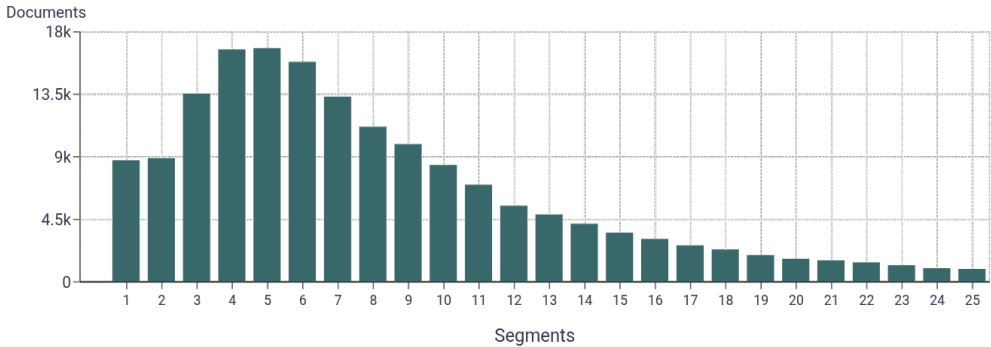
≤ 25 segments **91.42%** (166K documents)
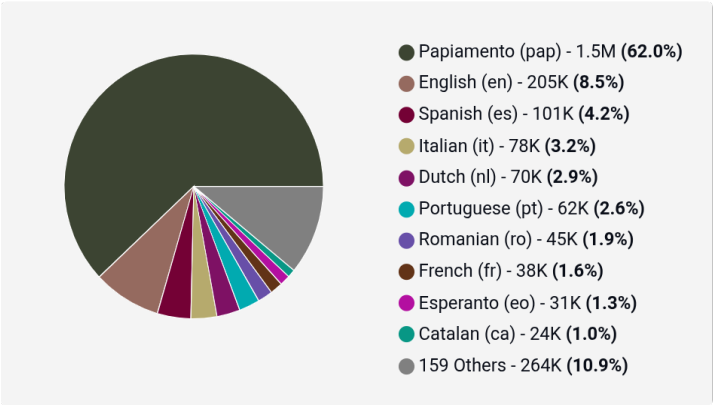> 25 segments **8.58%** (16K documents)
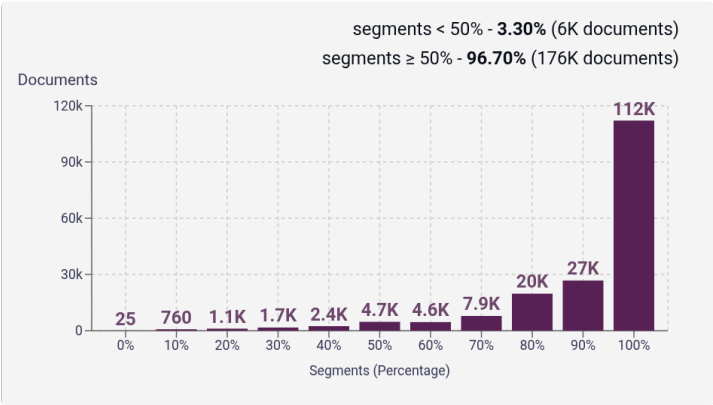

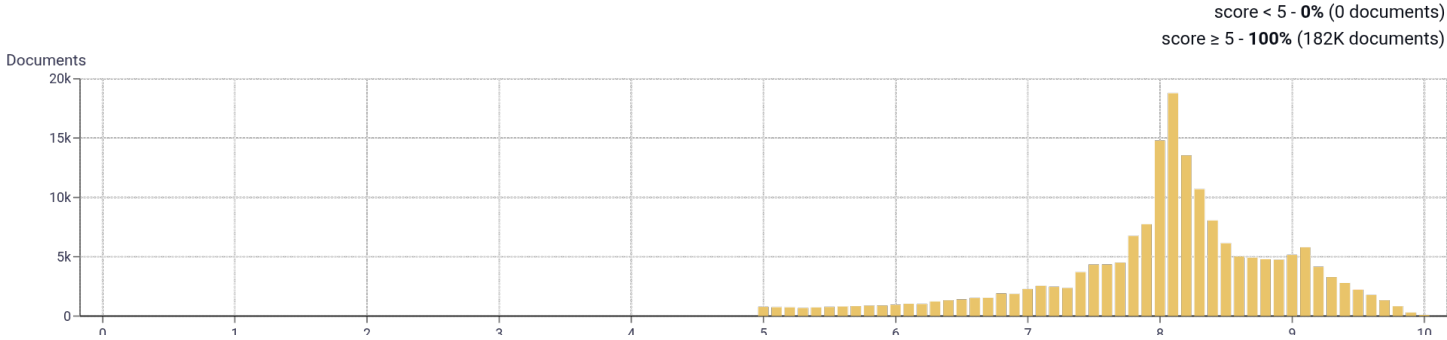## Document collections

CC = **93.18%**
IA = **6.82%**

67 Others (182K)

Language Distribution

Number of segments in the Papiamento (pap) corpus

- Papiamento (pap) - 1.5M **(62.0%)**
- English (en) - 205K **(8.5%)**
- Spanish (es) - 101K **(4.2%)**
- Italian (it) - 78K **(3.2%)**
- Dutch (nl) - 70K **(2.9%)**
- Portuguese (pt) - 62K **(2.6%)**
- Romanian (ro) - 45K **(1.9%)**
- French (fr) - 38K **(1.6%)**
- Esperanto (eo) - 31K **(1.3%)**
- Catalan (ca) - 24K **(1.0%)**
- 159 Others - 264K **(10.9%)**

Percentage of segments in Papiamento (pap) inside documents

segments < 50% - **3.30%** (6K documents)
segments ≥ 50% - **96.70%** (176K documents)

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (182K documents)

Documents

20k

15k

10k

5k

0

0 1 2 3 4 5 6 7 8 9 10

Segment length distribution by token

≤ **49** tokens = **1.7M** segments | **504K** duplicates
> **50** tokens = **680K** segments | **93K** duplicates

Segments

140k

105k

70k

35k

0

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49

Segment noise distribution

Too long	**1.09%**
Too short	**11.39%**
URLs	**2.26%**
Bad encoding	**0.01%**
Contains PII	**0.42%**

10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

■ % of corpus
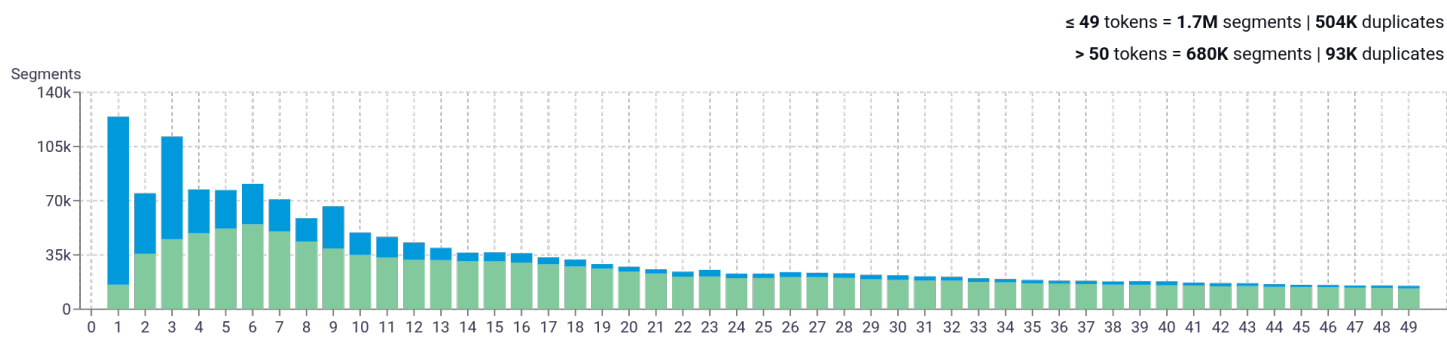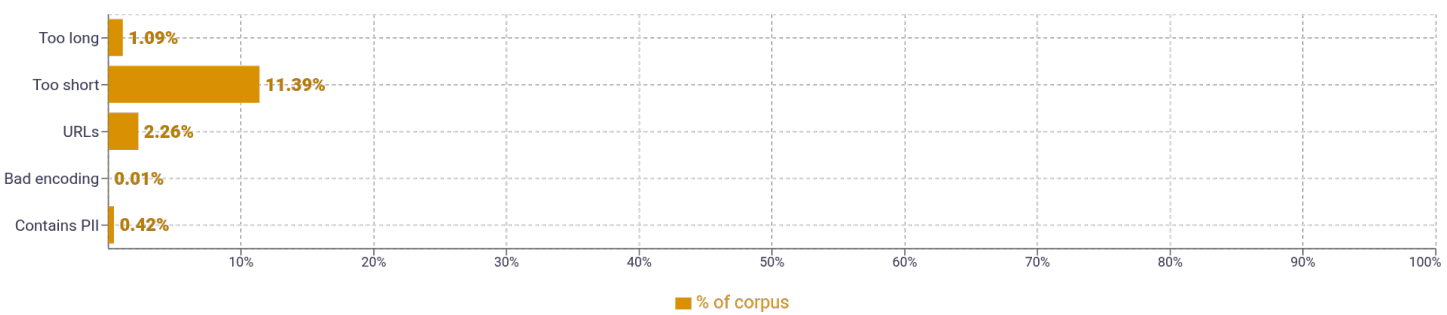
Frequent n-grams

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

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				