HPLT Analytics report

HPLT Analytics

General overview

Corpus	Date	Language
hplt-v3-spa_Latn	9/24/2025	Spanish (es)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
725,580,145	16,311,481,939	9,110,045,395 (55.85 %)	512B	2,737,390,990,249	2.54 TB

Top 10 domains

Domain	Docs	% of total
blogspot.com	29M	4.05%
wordpress.com	8.5M	1.17%
blogspot.com.es	4.5M	0.62%
buenastareas.com	3.5M	0.48%
blogspot.com.ar	2.4M	0.33%
blogspot.mx	2.2M	0.31%
elpais.com	1.9M	0.26%
web.app	1.4M	0.20%
as.com	1.2M	0.16%
yahoo.com	1.1M	0.16%

Top 10 TLDs

Domain	Docs	% of total
com	371M	51.18%
es	101M	13.90%
com.ar	34M	4.67%
org	27M	3.73%
net	23M	3.18%
com.mx	20M	2.81%
cl	20M	2.70%
mx	16M	2.26%
pe	6.8M	0.93%
info	5.9M	0.82%

Register labels

- HI - 2.5%
- ID - 2.0%
- IN - 14.8%
- IP - 18.6%
- LY - 0.2%
- MIX - 4.4%
- NA - 42.3%
- OP - 6.8%
- SP - 0.5%
- UNK - 8.0%

- HI_other - 1.7%
- HI_re - 0.8%
- ID_other - 2.0%
- IN_dtp - 6.6%
- IN_en - 0.4%
- IN_fi - 0.0%
- IN_lt - 0.6%
- IN_other - 6.9%
- IN_ra - 0.3%
- IP_ds - 15.3%
- IP_other - 3.3%
- LY_other - 0.2%
- MIX - 4.4%
- NA_nb - 3.9%
- NA_ne - 30.4%
- NA_other - 3.1%
- NA_sr - 4.9%
- OP_av - 0.7%
- OP_ob - 1.9%
- OP_other - 1.6%
- OP_rs - 0.8%
- OP_rv - 1.9%
- SP_it - 0.4%
- SP_other - 0.1%
- UNK - 8.0%
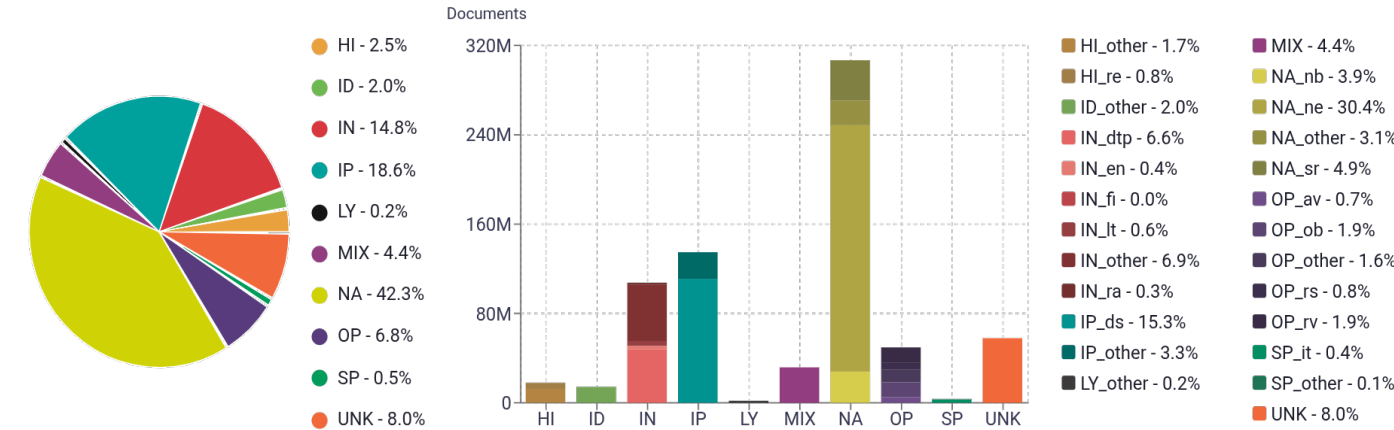
🤖 **MT**:3.8% | 28M Documents

Documents size (in segments) ⓘ

≤ 25 segments **79.15%** (574M documents)
**> 25** segments **20.85%** (151M documents)


Document collections

CC = 88.49%
IA = 11.51%

67 Others (726M)
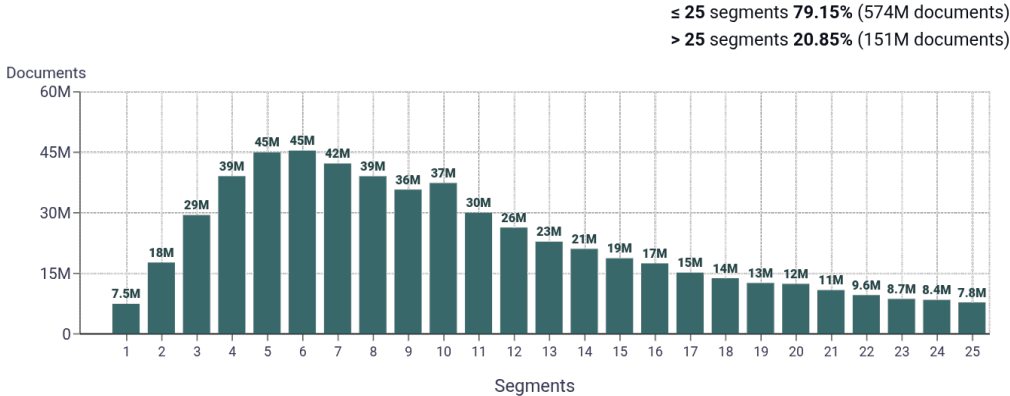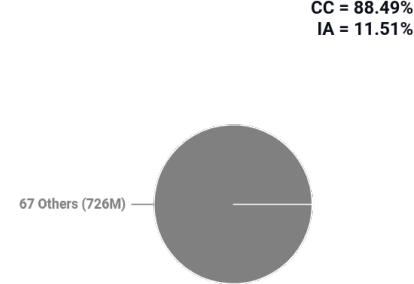
Language Distribution

Number of segments in the Spanish (es) corpus

- Spanish (es) - 14B **(87.1%)**
- English (en) - 698M **(4.3%)**
- Italian (it) - 353M **(2.2%)**
- Portuguese (pt) - 215M **(1.3%)**
- Galician (gl) - 193M **(1.2%)**
- French (fr) - 161M **(1.0%)**
- German (de) - 65M **(0.4%)**
- Catalan (ca) - 60M **(0.4%)**
- Asturian (ast) - 29M **(0.2%)**
- Dutch (nl) - 28M **(0.2%)**
- 165 Others - 295M **(1.8%)**
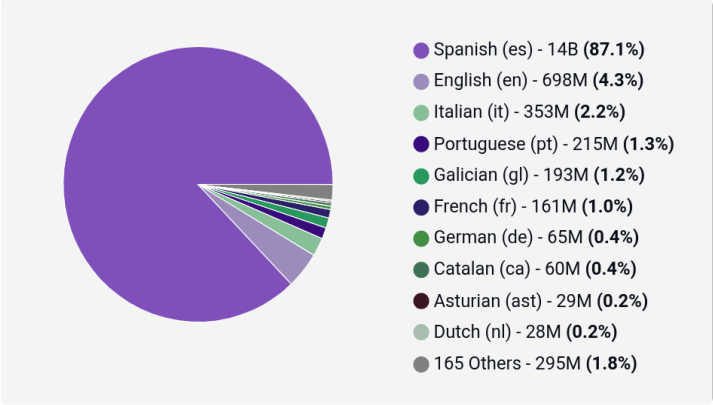
Percentage of segments in Spanish (es) inside documents
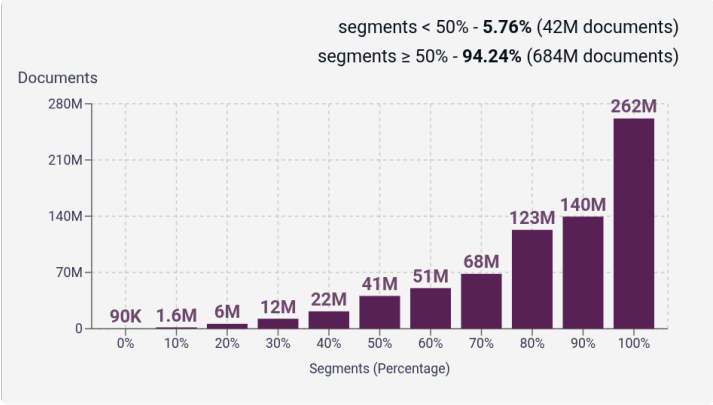
segments < 50% - **5.76%** (42M documents)
segments ≥ 50% - **94.24%** (684M documents)

Documents

90K	1.6M	6M	12M	22M	41M	51M	68M	123M	140M	262M

Segments (Percentage): 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (726M documents)

Documents

Segment length distribution by token

≤ **49** tokens = **13B** segments | **6.4B** duplicates
> **50** tokens = **3.3B** segments | **794M** duplicates

Segments

Segment noise distribution

- Too long — **0.90%**
- Too short — **13.02%**
- URLs — **1.70%**
- Bad encoding — **0.18%**
- Contains PII — **0.52%**

■ % of corpus
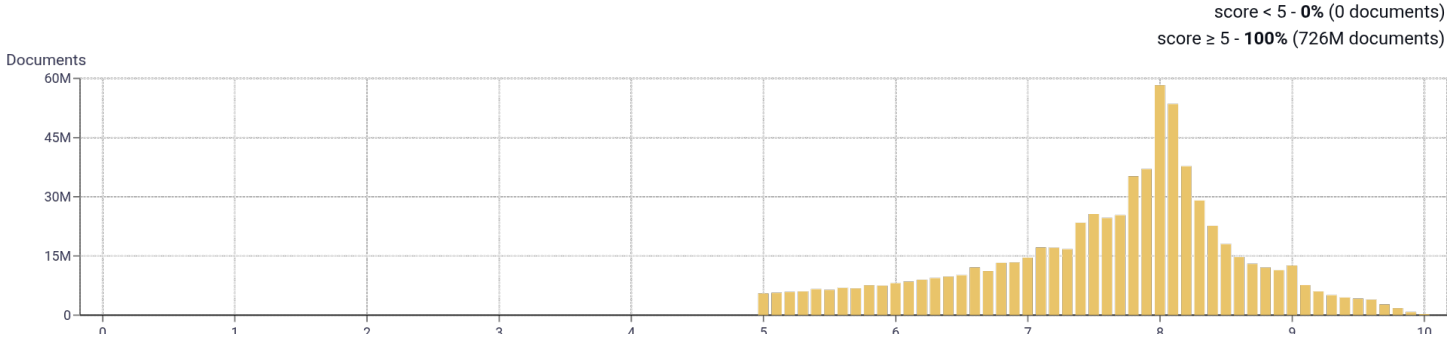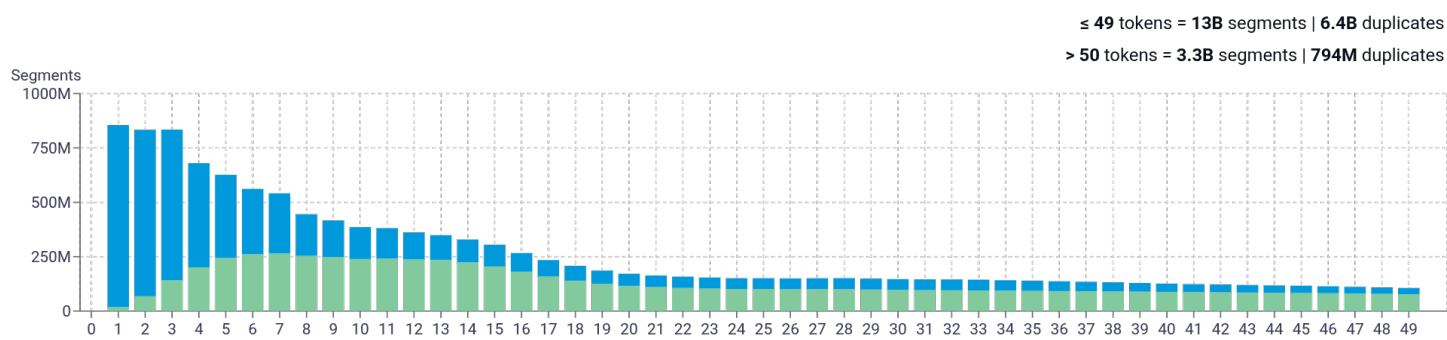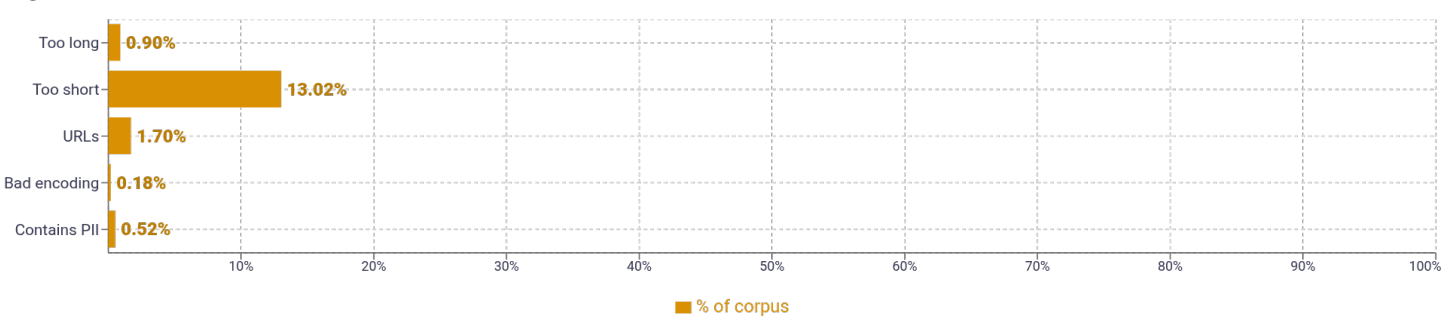
Frequent n-grams

SIZE	N-GRAMS	
1	ser | 643,143,061 prostitutas | 548,934,410 años | 501,140,618 dos | 455,327,367 así | 428,749,066	⧉
2	redes sociales | 51,660,916 alta calidad | 45,661,142 sitio web | 43,361,566 prostitutas prostitutas | 42,781,955 primera vez | 38,493,484	⧉
3	molino de bolas | 32,120,939 tener en cuenta | 28,348,056 fin de semana | 27,532,359 trituradora de piedra | 27,041,319 publicar un comentario | 24,649,442	⧉
4	opinión de un comprador | 10,623,690 comentario en la entrada | 5,530,382 importante tener en cuenta | 5,093,737 presidente de la república | 4,844,115 provincia de buenos aires | 4,752,698	⧉
5	organización mundial de la salud | 3,046,120 mejor forma de comprar online | 2,470,833 servicio de atención al cliente | 2,461,309 américa latina y el caribe | 2,358,188 mejorar la calidad de vida | 2,131,368	⧉

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				