HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-glg_Latn	9/18/2025	Galician (gl)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
4,033,272	66,522,728	44,114,752 (66.32 %)	2.2B	11,636,455,615	11.12 GB

Top 10 domains

Domain	Docs	% of total
blogspot.com	300K	7.44%
xunta.gal	125K	3.11%
wikipedia.org	116K	2.88%
wordpress.com	98K	2.42%
galiciaconfiden...	65K	1.61%
consumer.es	60K	1.48%
blogspot.com.es	57K	1.42%
pontevedraviva.com	54K	1.33%
nosdiario.gal	47K	1.17%
crtvg.es	46K	1.15%

Top 10 TLDs

Domain	Docs	% of total
com	1.6M	39.43%
gal	962K	23.86%
es	671K	16.63%
org	467K	11.57%
net	58K	1.45%
com.es	57K	1.43%
info	48K	1.20%
eu	45K	1.11%
gob.es	13K	0.32%
co	10K	0.25%

Register labels

- HI - 1.1%
- ID - 0.3%
- IN - 15.8%
- IP - 11.2%
- LY - 0.1%
- MIX - 6.1%
- NA - 53.7%
- OP - 5.3%
- SP - 0.8%
- UNK - 5.5%

- HI_other - 0.8%
- HI_re - 0.2%
- ID_other - 0.3%
- IN_dtp - 5.6%
- IN_en - 3.4%
- IN_fi - 0.0%
- IN_lt - 1.3%
- IN_other - 5.4%
- IN_ra - 0.1%
- IP_ds - 6.3%
- IP_other - 4.9%
- LY_other - 0.1%
- MIX - 6.1%
- NA_nb - 9.4%
- NA_ne - 37.2%
- NA_other - 4.0%
- NA_sr - 3.0%
- OP_av - 0.2%
- OP_ob - 2.8%
- OP_other - 1.2%
- OP_rs - 0.3%
- OP_rv - 0.9%
- SP_it - 0.7%
- SP_other - 0.1%
- UNK - 5.5%
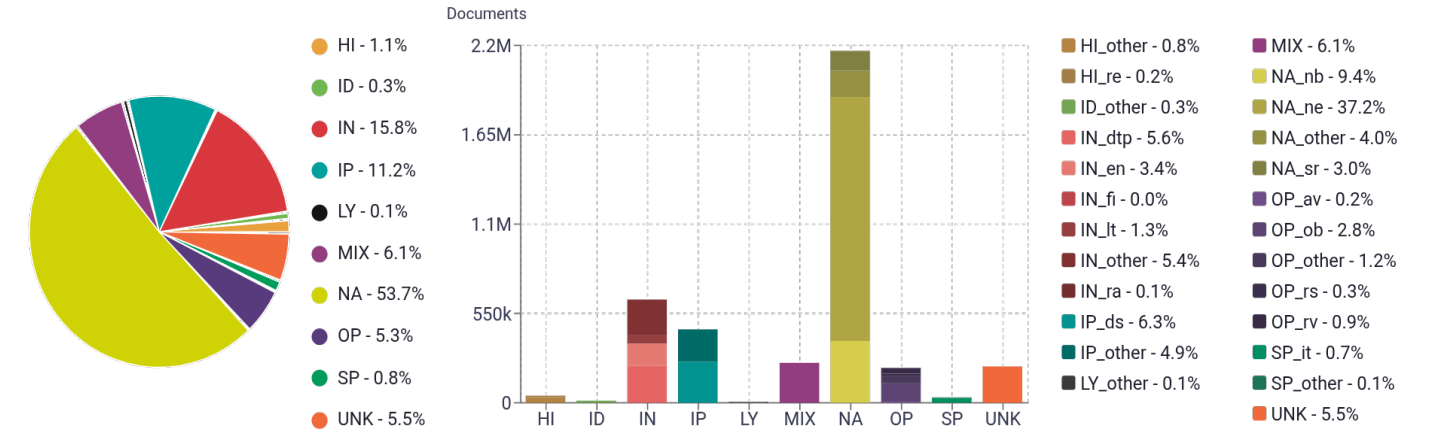
🖳 **MT**:2.6% | 107K Documents

Documents size (in segments) ⓘ

≤ 25 segments **86.71%** (3.5M documents)
> 25 segments **13.29%** (536K documents)


Document collections

CC = 92.97%
IA = 7.03%

67 Others (4M)
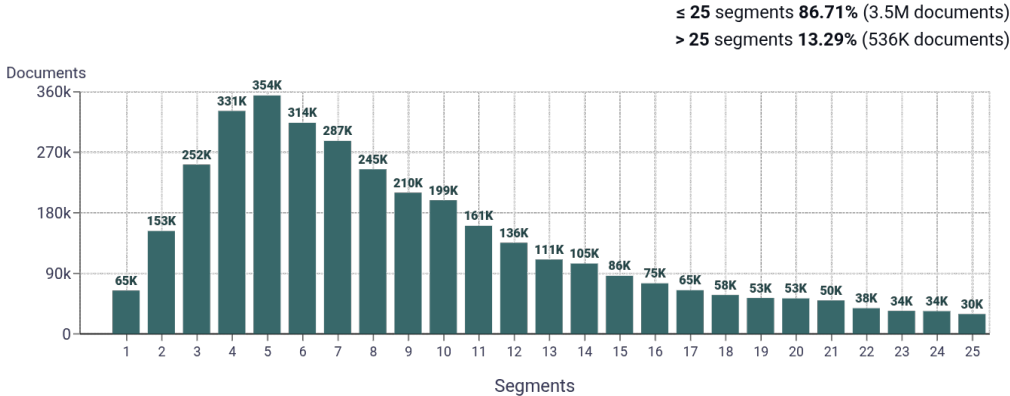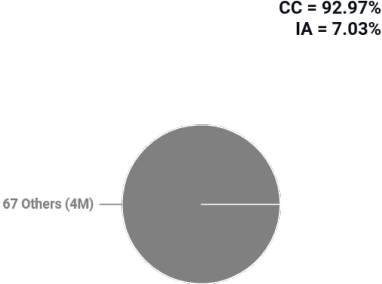
Language Distribution

Number of segments in the Galician (gl) corpus

- Galician (gl) - 56M **(84.9%)**
- Spanish (es) - 3.1M **(4.7%)**
- English (en) - 2.2M **(3.3%)**
- Italian (it) - 1.3M **(1.9%)**
- Portuguese (pt) - 1M **(1.5%)**
- French (fr) - 593K **(0.9%)**
- Catalan (ca) - 287K **(0.4%)**
- German (de) - 240K **(0.4%)**
- Lithuanian (lt) - 95K **(0.1%)**
- Dutch (nl) - 84K **(0.1%)**
- 165 Others - 1.2M **(1.8%)**
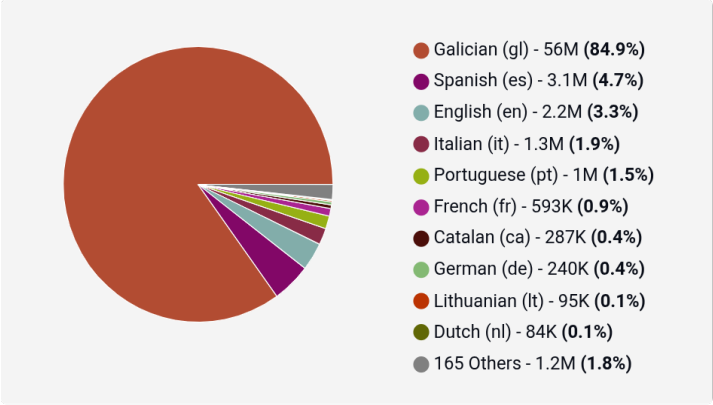
Percentage of segments in Galician (gl) inside documents
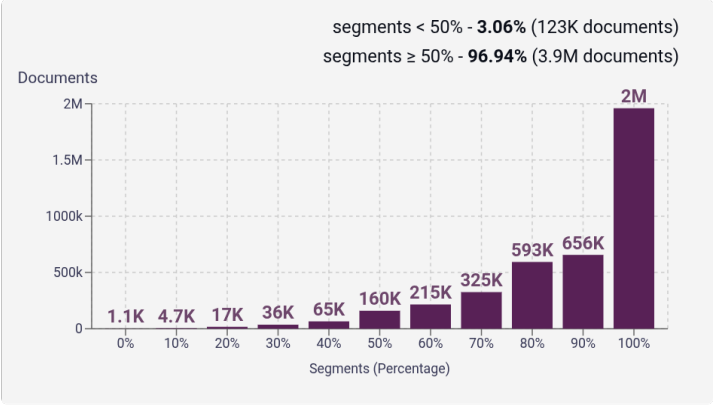
segments < 50% - **3.06%** (123K documents)
segments ≥ 50% - **96.94%** (3.9M documents)

Documents

- 0%: 1.1K
- 10%: 4.7K
- 20%: 17K
- 30%: 36K
- 40%: 65K
- 50%: 160K
- 60%: 215K
- 70%: 325K
- 80%: 593K
- 90%: 656K
- 100%: 2M

Segments (Percentage)

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (4M documents)

Documents

Segment length distribution by token

≤ 49 tokens = **52M** segments | **20M** duplicates
> 50 tokens = **15M** segments | **2.7M** duplicates

Segments

Segment noise distribution

- Too long: **0.77%**
- Too short: **11.03%**
- URLs: **1.55%**
- Bad encoding: **0.15%**
- Contains PII: **0.32%**

■ % of corpus
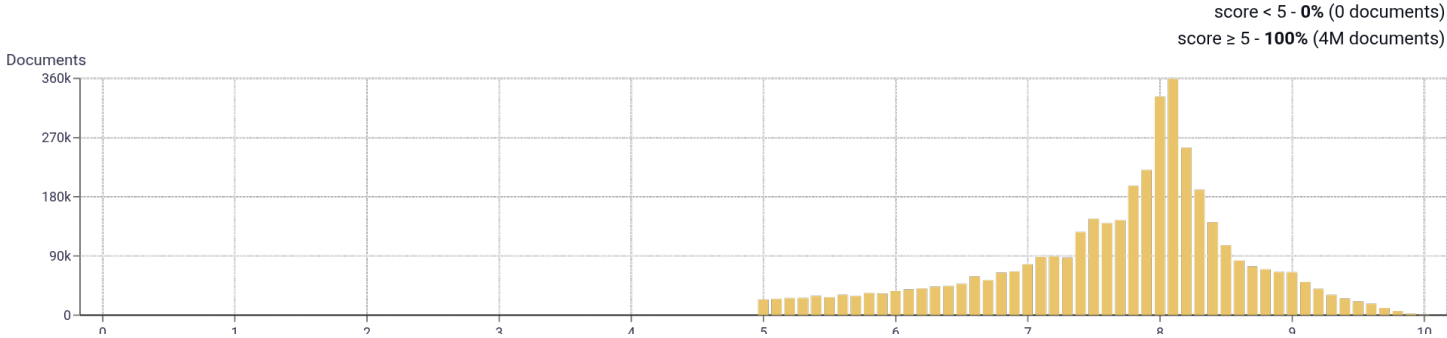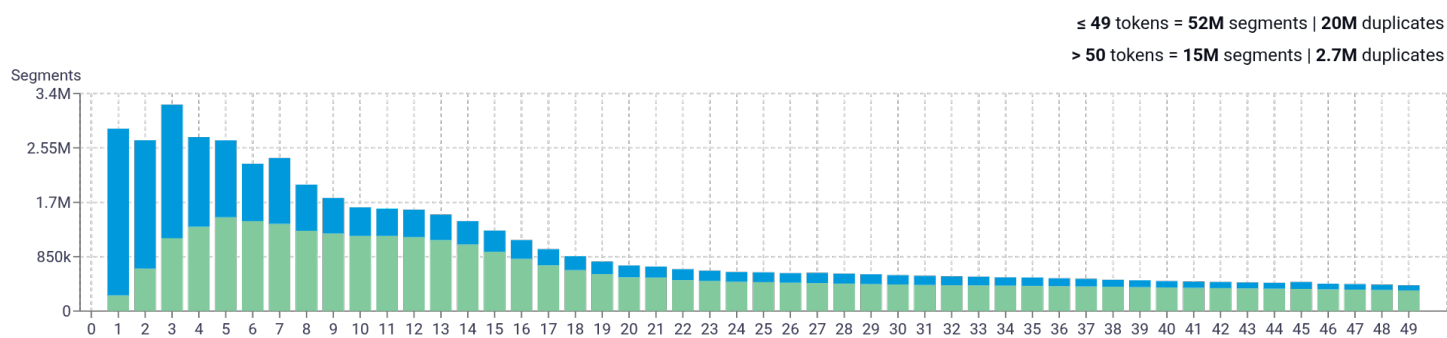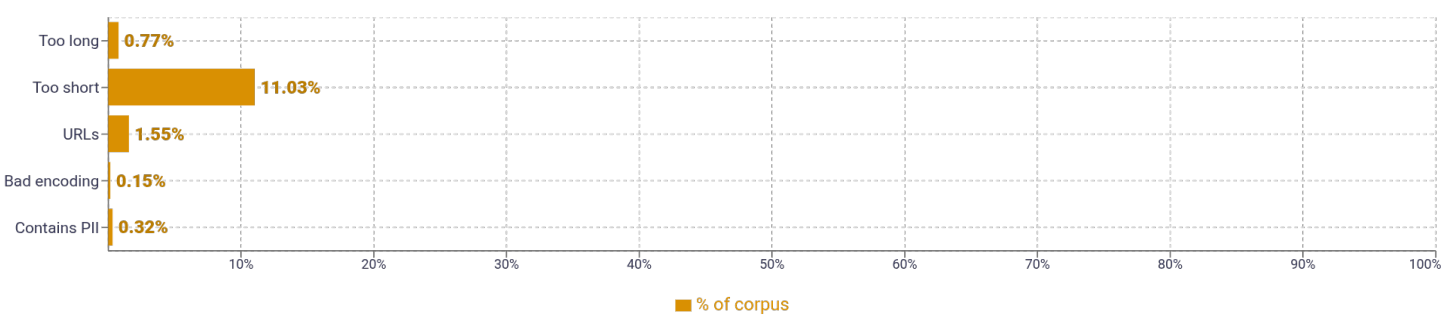
Frequent n-grams

SIZE	N-GRAMS	
1	galicia | 3,119,188 anos | 2,557,254 día | 2,340,782 ano | 2,287,743 persoas | 2,179,117	⧉
2	terá lugar | 233,273 medio ambiente | 225,464 primeira vez | 161,229 sitio web | 158,543 correo electrónico | 157,857	⧉
3	santiago de compostela | 414,089 xunta de galicia | 408,106 editar a fonte | 407,835 millóns de euros | 270,183 fin de semana | 204,076	⧉
4	diario oficial de galicia | 69,943 comunidade autónoma de galicia | 68,467 día das letras galegas | 51,905 sen ánimo de lucro | 36,903 consello da cultura galega | 35,582	⧉
5	universidade de santiago de compostela | 64,973 prazo de presentación de solicitudes | 34,362 contra a violencia de xénero | 31,320 electrónica da xunta de galicia | 27,699 concello de val do dubra | 25,829	⧉

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				