HPLT Analytics report

HPLT Analytics

General overview

Corpus	Date	Language
hplt-v3-gle_Latn	9/17/2025	Irish (ga)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
786,690	18,062,235	12,984,055 (71.89 %)	562M	2,946,523,386	2.92 GB

Top 10 domains

Domain	Docs	% of total
tuairisc.ie	27K	3.41%
wikipedia.org	18K	2.23%
airbnb.ie	17K	2.11%
europa.eu	13K	1.65%
blogspot.com	9.1K	1.16%
eferrit.com	8.3K	1.06%
soft-free-downl...	8.1K	1.03%
stealthsettings...	7.5K	0.95%
eureporter.co	7.5K	0.95%
martech.zone	7.3K	0.93%

Top 10 TLDs

Domain	Docs	% of total
com	372K	47.30%
ie	182K	23.09%
org	50K	6.37%
net	20K	2.60%
eu	19K	2.38%
pt	14K	1.77%
co	8.3K	1.05%
zone	7.3K	0.93%
pw	5K	0.64%
gov.ie	4.9K	0.62%

Register labels

- HI - 3.2%
- ID - 0.4%
- IN - 15.4%
- IP - 14.0%
- LY - 0.4%
- MIX - 2.1%
- NA - 13.3%
- OP - 4.3%
- SP - 0.4%
- UNK - 46.6%

- HI_other - 2.7%
- HI_re - 0.5%
- ID_other - 0.4%
- IN_dtp - 5.0%
- IN_en - 2.7%
- IN_fi - 0.0%
- IN_lt - 1.1%
- IN_other - 6.5%
- IN_ra - 0.0%
- IP_ds - 11.9%
- IP_other - 2.1%
- LY_other - 0.4%

- MIX - 2.1%
- NA_nb - 2.7%
- NA_ne - 7.6%
- NA_other - 2.3%
- NA_sr - 0.8%
- OP_av - 0.3%
- OP_ob - 1.2%
- OP_other - 0.6%
- OP_rs - 0.8%
- OP_rv - 1.3%
- SP_it - 0.2%
- SP_other - 0.2%
- UNK - 46.6%
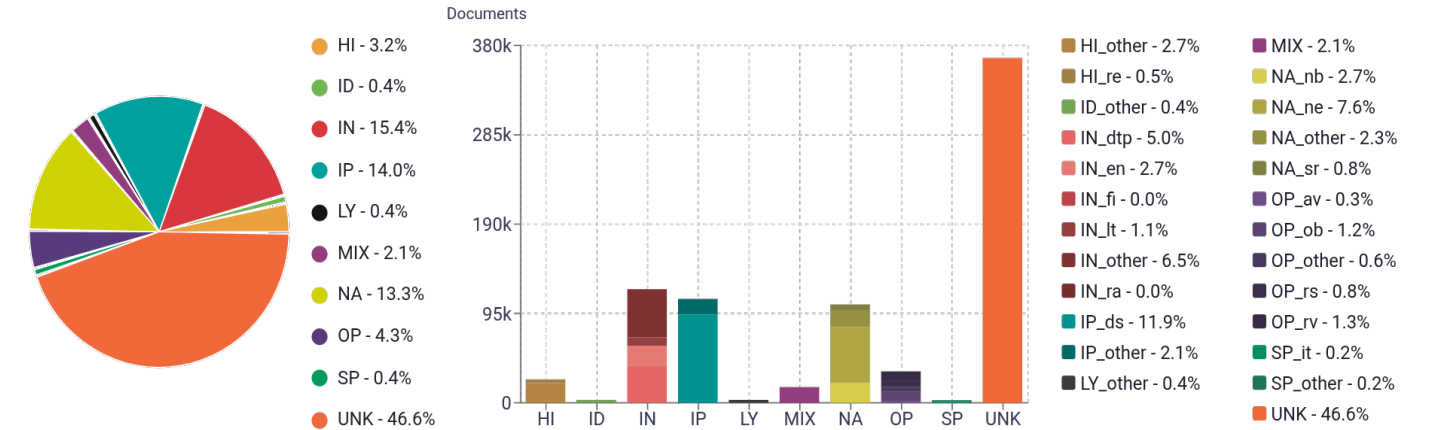
🖲 **MT**:46.4% | 365K Documents

Documents size (in segments) ⓘ

≤ 25 segments **76.32%** (600K documents)
> 25 segments **23.68%** (186K documents)

Document collections

CC = 88.41%
IA = 11.59%

67 Others (787K)
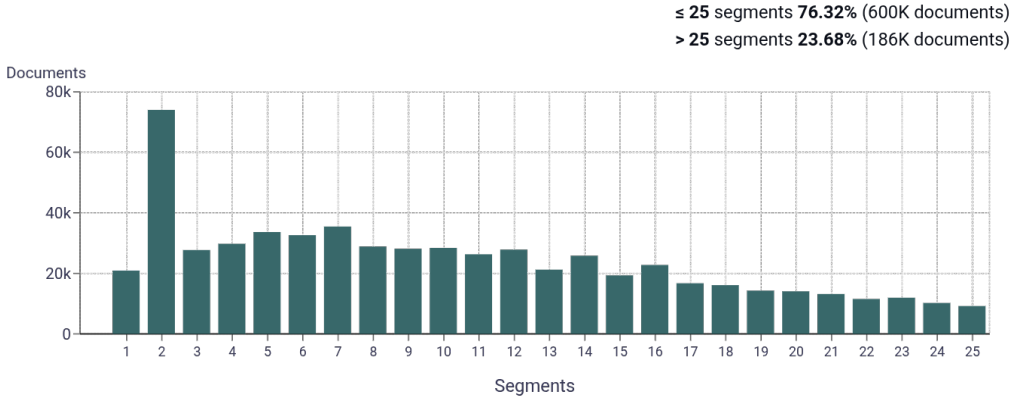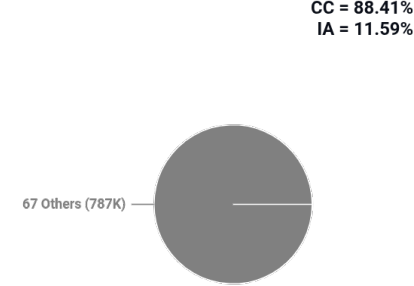
Language Distribution

Number of segments in the Irish (ga) corpus

- Irish (ga) - 14M **(77.9%)**
- English (en) - 1.1M **(6.1%)**
- Spanish (es) - 400K **(2.2%)**
- Catalan (ca) - 304K **(1.7%)**
- Italian (it) - 227K **(1.3%)**
- French (fr) - 214K **(1.2%)**
- Czech (cs) - 182K **(1.0%)**
- Scottish Gaelic (gd) - 163K **(0.9%)**
- Portuguese (pt) - 138K **(0.8%)**
- Hungarian (hu) - 132K **(0.7%)**
- 165 Others - 1.1M **(6.3%)**
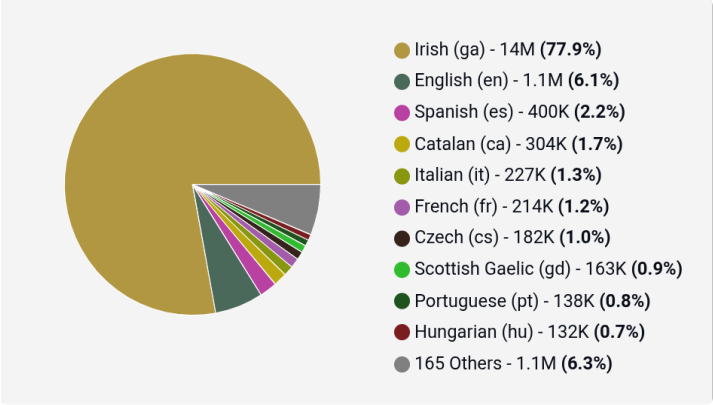
Percentage of segments in Irish (ga) inside documents
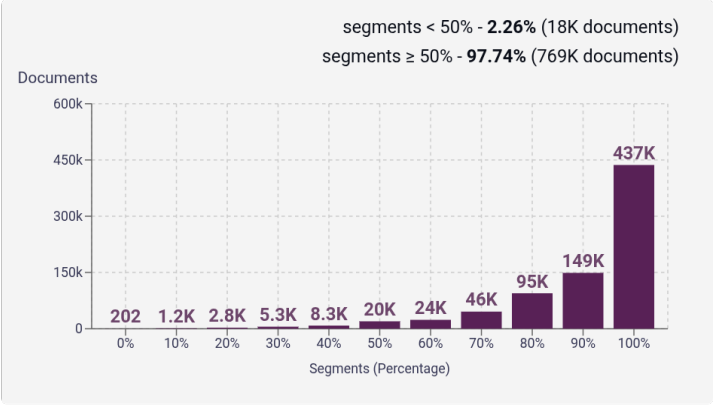
segments < 50% - **2.26%** (18K documents)
segments ≥ 50% - **97.74%** (769K documents)

Documents

202	1.2K	2.8K	5.3K	8.3K	20K	24K	46K	95K	149K	437K
0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

Segments (Percentage)

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (787K documents)

Documents

Segment length distribution by token

≤ **49** tokens = **14M** segments | **4.7M** duplicates
> **50** tokens = **3.6M** segments | **369K** duplicates

Segments

Segment noise distribution

- Too long — **1.05%**
- Too short — **11.77%**
- URLs — **1.43%**
- Bad encoding — **0.02%**
- Contains PII — **0.38%**

■ % of corpus
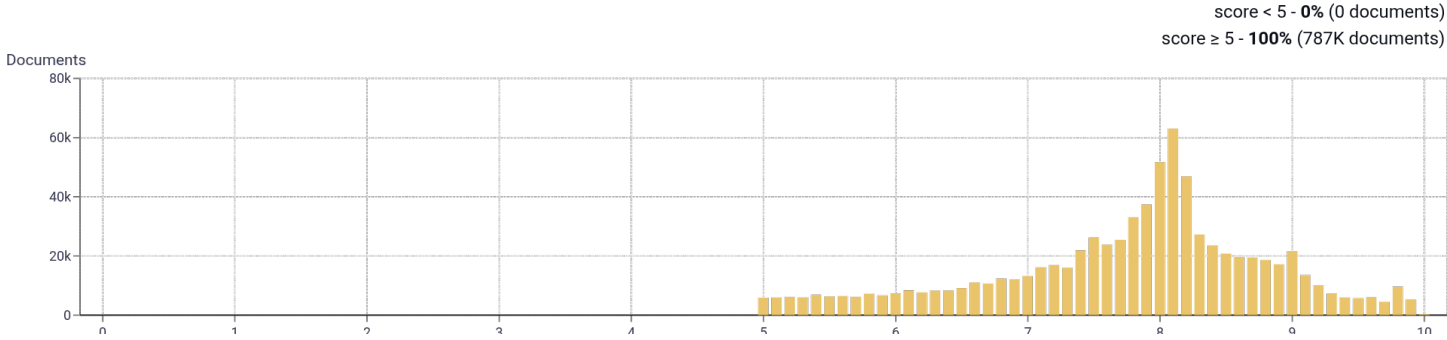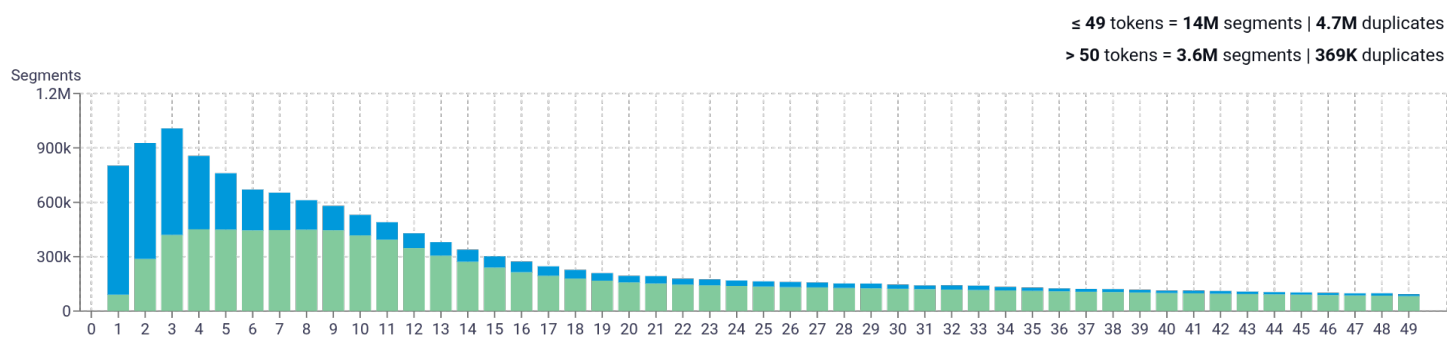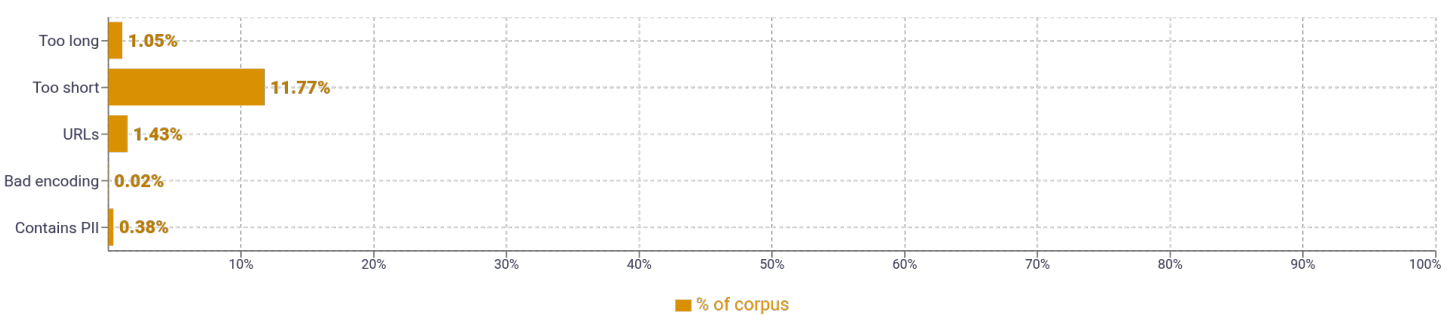
Frequent n-grams

SIZE	N-GRAMS
1	bhfuil | 2,676,008 féidir | 2,068,029 atá | 2,045,718 d | 1,896,962 níos | 1,742,051
2	níos mó | 625,992 féidir leat | 617,534 níos fearr | 159,702 bhaint amach | 129,762 mian leat | 124,659
3	saor in aisce | 357,229 chuid is mó | 164,780 nuair a bhíonn | 104,773 fud an domhain | 86,396 uair an chloig | 79,137
4	lá atá inniu ann | 39,861 rud é go bhfuil | 26,503 féidir leat a bheith | 22,192 arís agus arís eile | 21,617 lóistíní saoire ar cíos | 21,279
5	rud a fhágann go bhfuil | 24,783 ós rud é go bhfuil | 20,549 más rud é nach bhfuil | 15,275 bí ar an chéad trácht | 15,099 iiiiiiiiiii iiiiiiiiiii iiiiiiiiiii iiiiiiiiiii iiiiiiiiiii | 13,169

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				