# HPLT Analytics report

HPLTAnalytics

## General overview

| Corpus | Analytics date | Language |
|---|---|---|
| gle_Latn.jsonl.tsv | 9/16/2024 | Irish (ga) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Size | Characters |
|---|---|---|---|---|---|
| 490,787 | 10,993,158 | 5,866,989 (53.37 %) | 336M | 1.72 GB | 1,738,847,965 |

## Top 10 domains

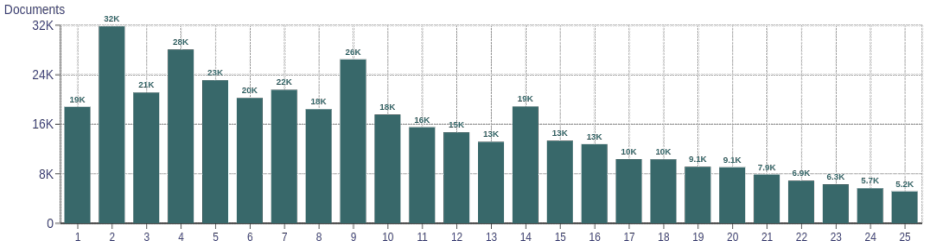
| Domain | Docs | % of total |
|---|---|---|
| wikipedia.org | 63K | 12.94 |
| tuairisc.ie | 25K | 5.11 |
| europa.eu | 11K | 2.28 |
| stealthsettings.com | 8.4K | 1.71 |
| blogspot.com | 8.3K | 1.69 |
| soft-free-download.com | 8.1K | 1.65 |
| duchas.ie | 7.8K | 1.60 |
| itsmygame.org | 6.9K | 1.40 |
| daily-helper.com | 6.6K | 1.35 |
| nos.ie | 6.5K | 1.32 |

## Top 10 TLDs

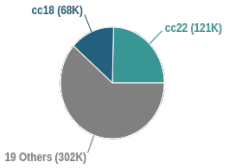
| Domain | Docs | % of total |
|---|---|---|
| com | 179K | 36.49 |
| ie | 149K | 30.43 |
| org | 88K | 18.00 |
| eu | 15K | 2.98 |
| net | 14K | 2.78 |
| mobi | 5.2K | 1.07 |
| pt | 3.6K | 0.74 |
| gov.ie | 2.9K | 0.60 |
| at | 1.8K | 0.36 |
| ru | 1.7K | 0.35 |

## Documents size (in segments)

**<= 25** segments **78.74%** (386K documents)
**> 25** segments **21.26%** (104K documents)

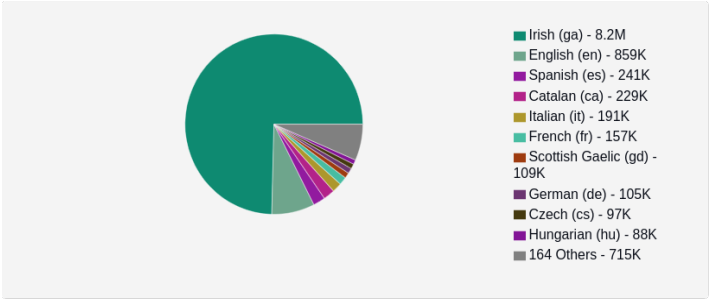



## Documents by collection

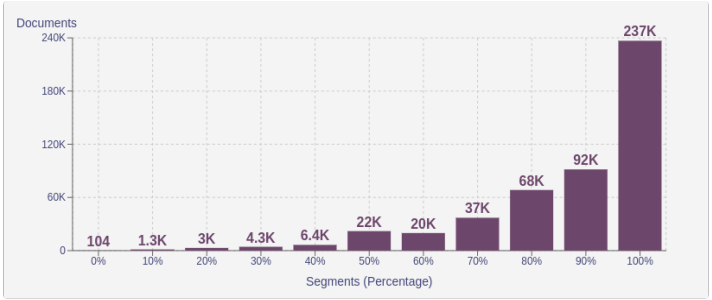



## Language Distribution

### Number of segments

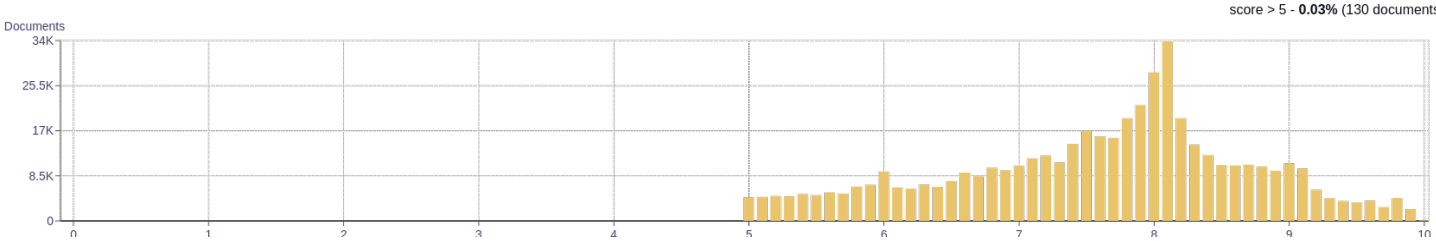



### Percentage of segments in Irish (ga) inside documents



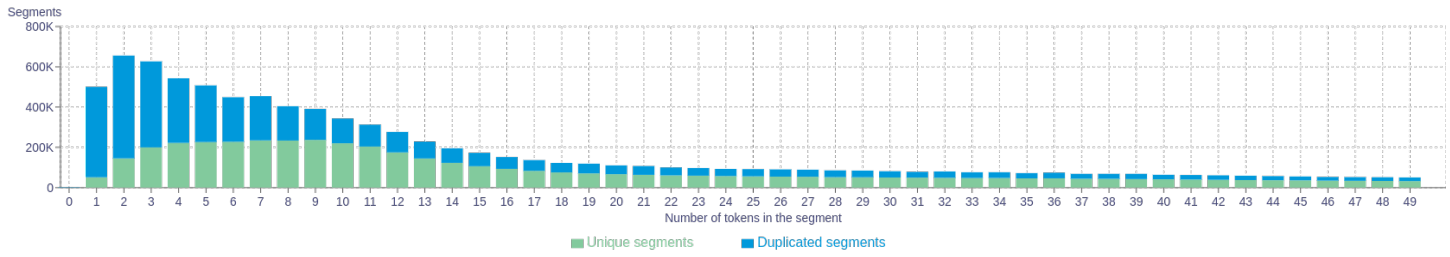

## Distribution of documents by document score

score <= 5 - **99.97%** (491K documents)
score > 5 - **0.03%** (130 documents)

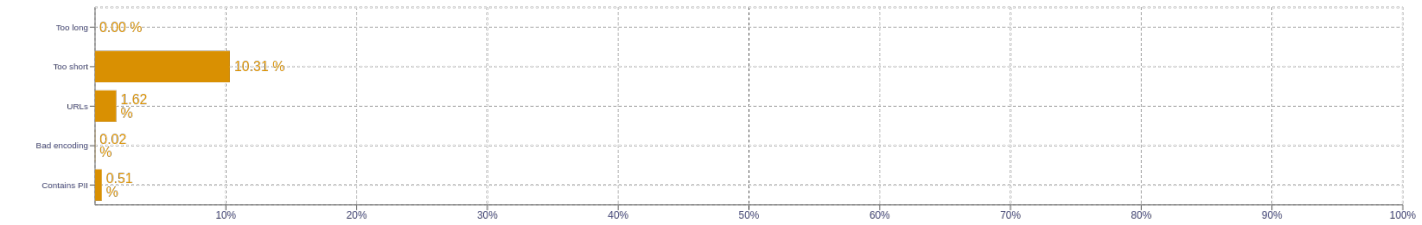



## Segment length distribution by token

**<= 49** tokens = **4.5M** segments | **4.4M** duplicates
**> 50** tokens = **2.1M** segments | **754K** duplicates




## Segment noise distribution



### Frequent n-grams

| Size | n-grams |
|---|---|
| 1 | sin \| 1565921 · bhfuil \| 1395697 · bhí \| 1241338 · atá \| 1195774 · d \| 1123479 |
| 2 | níos mó \| 301320 · féidir leat \| 290291 · átha cliath \| 91311 · níos fearr \| 71883 · sin féin \| 70114 |
| 3 | saor in aisce \| 208120 · chuid is mó \| 89530 · chur ar fáil \| 50016 · fud an domhain \| 47986 · nuair a bhí \| 45820 |
| 4 | lá atá inniu ann \| 23173 · líne saor in aisce \| 19579 · rud é go bhfuil \| 13083 · roinnt i lionraí sóisialta \| 12285 · rud é nach bhfuil \| 11809 |
| 5 | dearmad a mheas an cluiche \| 10997 · play ar líne a flash \| 10954 · ós rud é go bhfuil \| 9689 · más rud é nach bhfuil \| 8855 · más mian leat an cluiche \| 5850 |

## 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