HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
ga_1.jsonl.tsv	3/16/2024	Irish (ga)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
115,529	13,949,561	3,466,337 (24.85 %)	152M	810.08 MB	

Top 10 domains

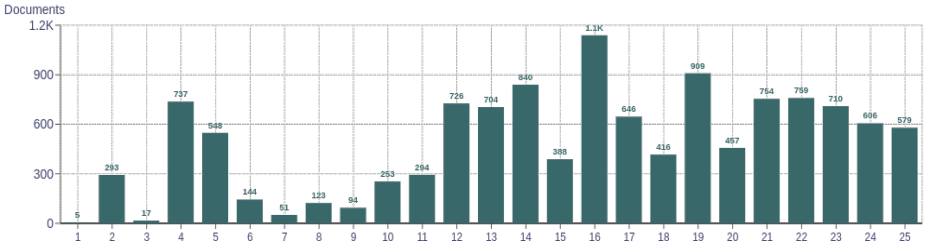
Domain	Docs	% of total
duchas.ie	8.2K	7.13
teanglann.ie	6.5K	5.59
librarything.com	5.2K	4.46
europa.eu	5.1K	4.44
tuairisc.ie	4.8K	4.13
wikipedia.org	4.7K	4.05
pornk-org.com	3.9K	3.38
sgames.org	3.5K	3.02
ainm.ie	3.5K	2.99
potafocal.com	2.7K	2.35

Top 10 TLDs

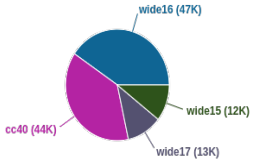
Domain	Docs	% of total
ie	49K	42.70
com	38K	32.55
org	11K	9.44
eu	5.8K	5.00
net	1.5K	1.34
pt	1.5K	1.28
cn	1.1K	0.92
news	946	0.82
at	745	0.64
gov.ie	664	0.57

Documents size (in segments)

<= 25 segments **10.56%** (12K documents)
> 25 segments **89.44%** (103K documents)

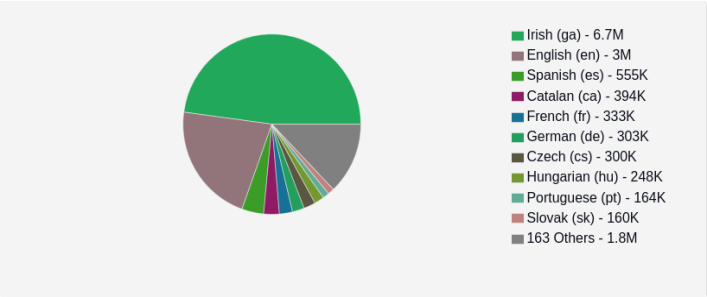


Documents by collection

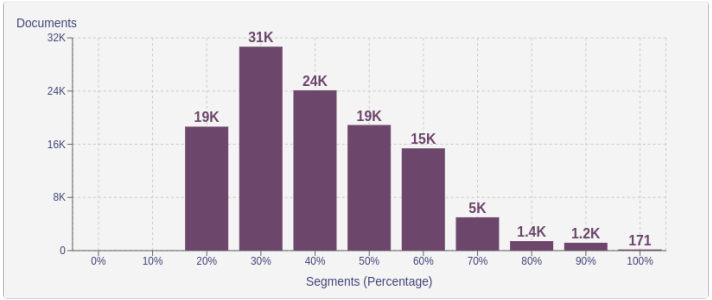


Language Distribution

Number of segments

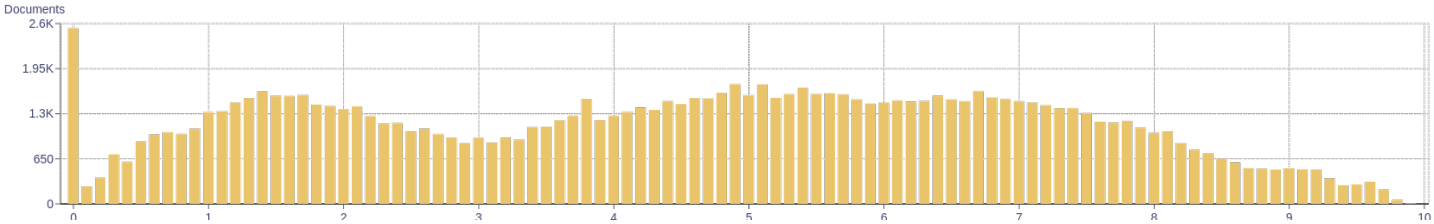


Percentage of segments in Irish (ga) inside documents



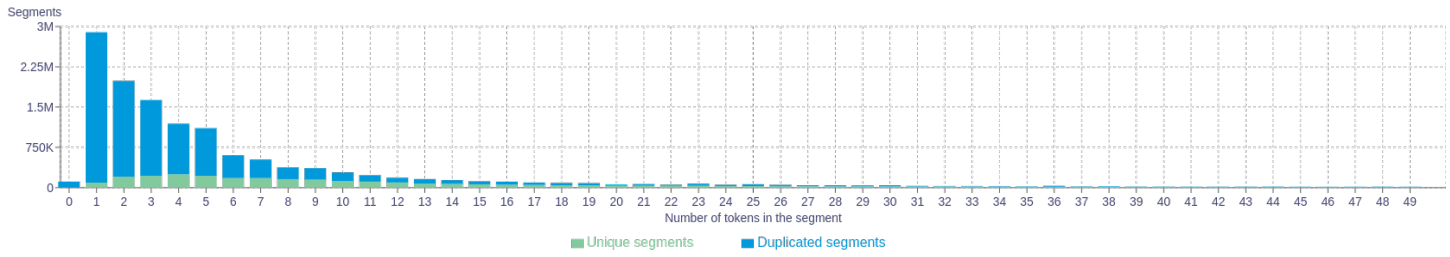
Distribution of documents by document score

score <= 5 - **53.1%** (61K documents)
score > 5 - **46.9%** (54K documents)

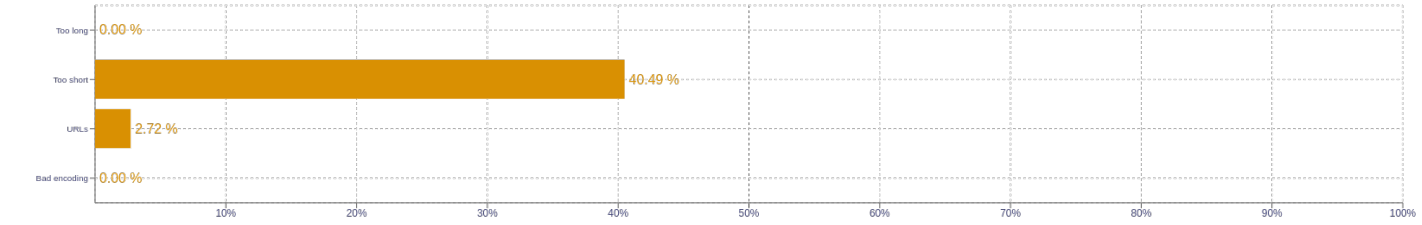


Segment length distribution by token

<= 49 tokens = **3M** segments | **10M** duplicates
> 50 tokens = **639K** segments | **144K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	the | 1106634 · of | 666287 · and | 525367 · to | 499552 · sin | 421573
2	of the | 167794 · meán fómhair | 82694 · níos mó | 77263 · féidir leat | 70355 · deireadh fómhair | 64487
3	cumann na sagart | 268768 · léachtaí an aifrinn | 186331 · machnamh ar léachtaí | 83362 · saor in aisce | 65913 · bailiúchán na scol | 38115
4	maidir leis na forbairtí | 11167 · sheoladh lenár liosta ríomhphoist | 11163 · gcoimeádfaí ar an eolas | 11161 · maith leat go gcoimeádfaí | 11160 · thionscadail eile de chuid | 11157
5	machnamh ar léachtaí an aifrinn | 81148 · macnamh ar léachtaí an aifrinn | 14051 · léim go príomhábhar an leathanaigh | 11406 · cuir do sheoladh lenár liosta | 11163 · más maith leat go gcoimeádfaí | 11160

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