HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
gla_Latn.jsonl.tsv	12/5/2024	Scottish Gaelic (gd)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
137,411	3,306,787	1,928,255 (58.31 %)	102M	480,450,664	471.56 MB

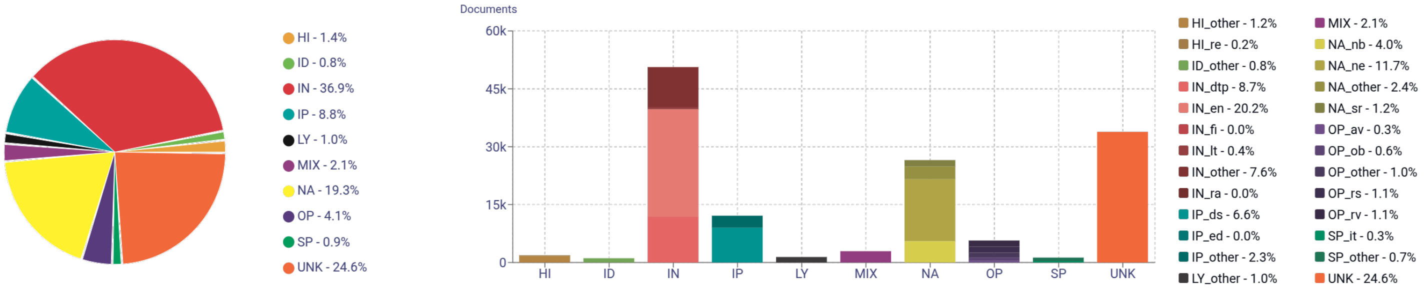
Top 10 domains

Domain	Docs	% of total
wikipedia.org	29K	21.02%
ambaile.org.uk	11K	7.76%
bbc.co.uk	8.7K	6.35%
bbc.com	3.2K	2.36%
learngaelic.net	2.7K	1.99%
versionsmart.news	2.2K	1.59%
uhi.ac.uk	2.1K	1.52%
vsaduidoma.com	1.8K	1.34%
ucc.ie	1.7K	1.24%
learngaelic.scot	1.7K	1.23%

Top 10 TLDs

Domain	Docs	% of total
com	49K	35.56%
org	35K	25.72%
co.uk	12K	9.06%
org.uk	12K	8.50%
net	5.6K	4.05%
scot	3K	2.21%
pt	2.5K	1.82%
ac.uk	2.5K	1.82%
news	2.3K	1.67%
ie	2K	1.46%

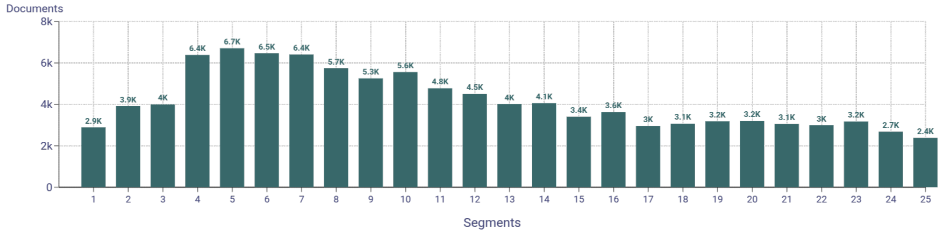
Register labels



🤖 **MT**:21.1% | 29K Documents

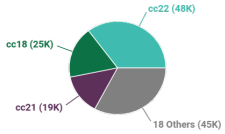
Documents size (in segments)

<= 25 segments **75.96%** (104K documents)
> 25 segments **24.04%** (33K documents)



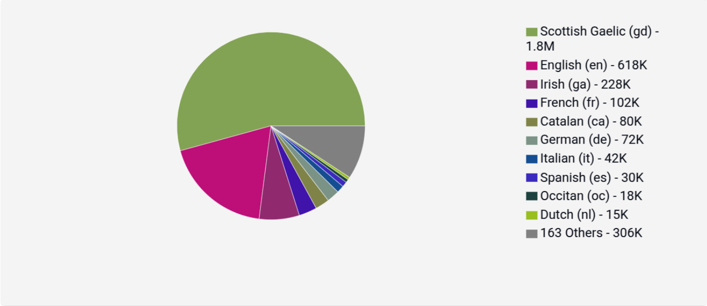
Documents by collection

CC = 82.14%
IA = 17.86%

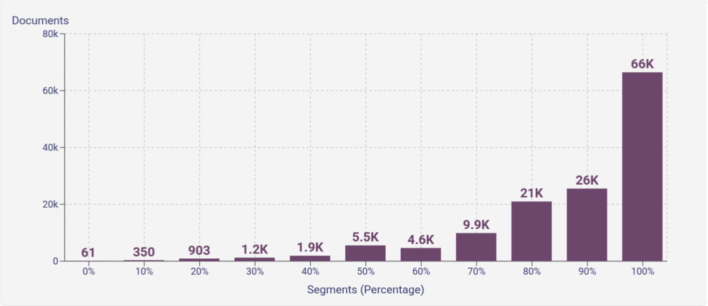


Language Distribution

Number of segments in the Scottish Gaelic (gd) corpus

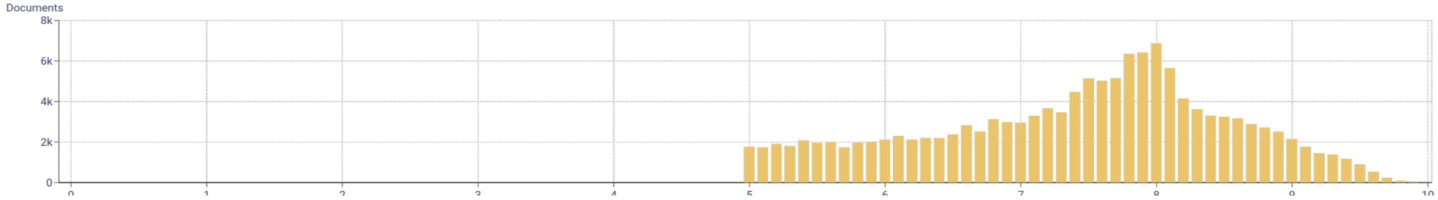


Percentage of segments in Scottish Gaelic (gd) inside documents



Distribution of documents by document score

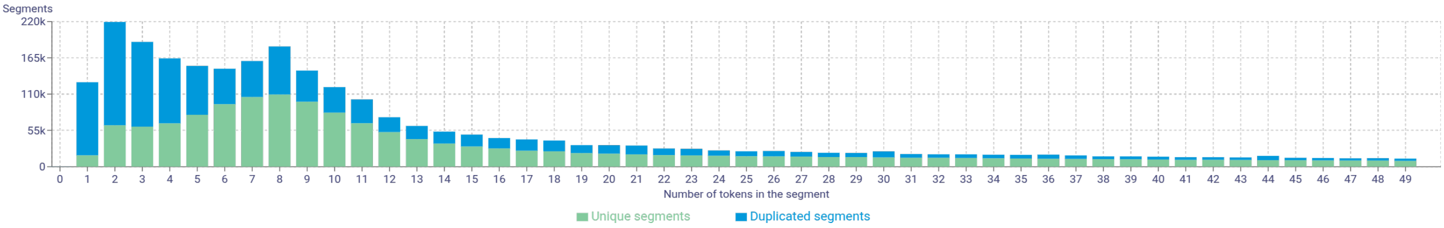
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (137K documents)



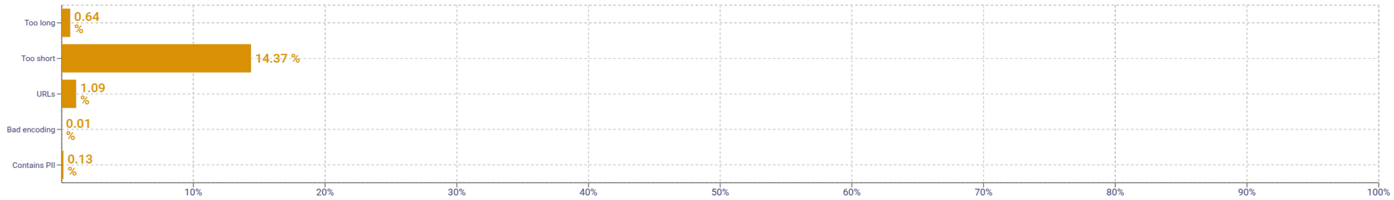
Segment length distribution by token

≤ 49 tokens = **1.5M** segments | **1.2M** duplicates

> 50 tokens = **623K** segments | **184K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	agus | 1776239 ann | 989906 gu | 884596 h | 436550 s | 403330
2	gu bheil | 162109 sam bith | 82620 gu math | 59991 gu h | 52490 bu chòir | 45040
3	deasaich an tùs | 54314 far a bheil | 19825 aig a bheil | 13763 ann an alba | 12798 aig an àm | 11840
4	aig an aon àm | 9843 dèanamh cinnteach gu bheil | 6331 ag ràdh gu bheil | 6192 san àm ri teachd | 5965 tuilleadh fiosrachaidh mu cheannach | 4486
5	fiosrachaidh mu cheannach is prìsean | 3577 os cionn ìre na mara | 3223 dòcha gum biodh e feumail | 2984 you may need to install | 2949 you have problems viewing them | 2949

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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av