

General overview

Corpus	Date	Language
hplt-v3-gla_Latn	9/17/2025	Scottish Gaelic (gd)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
204,013	3,769,630	2,932,827 (77.80 %)	130M	626,415,896	616.35 MB

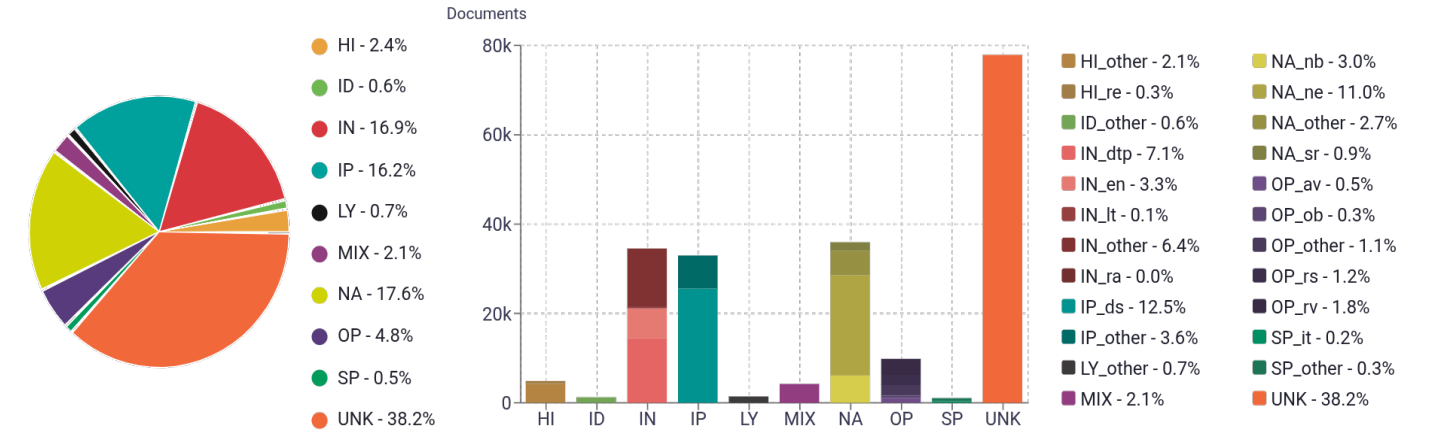
Top 10 domains

Domain	Docs	% of total
ambaile.org.uk	9K	4.41%
bbc.co.uk	7.3K	3.59%
wikipedia.org	5.8K	2.84%
martech.zone	5.4K	2.67%
bbc.com	4.5K	2.21%
versionsmart.news	3.3K	1.61%
eturbonews.com	2.8K	1.35%
uhi.ac.uk	1.7K	0.83%
wordpress.com	1.6K	0.77%
cm-santiago-do-...	1.5K	0.73%

Top 10 TLDs

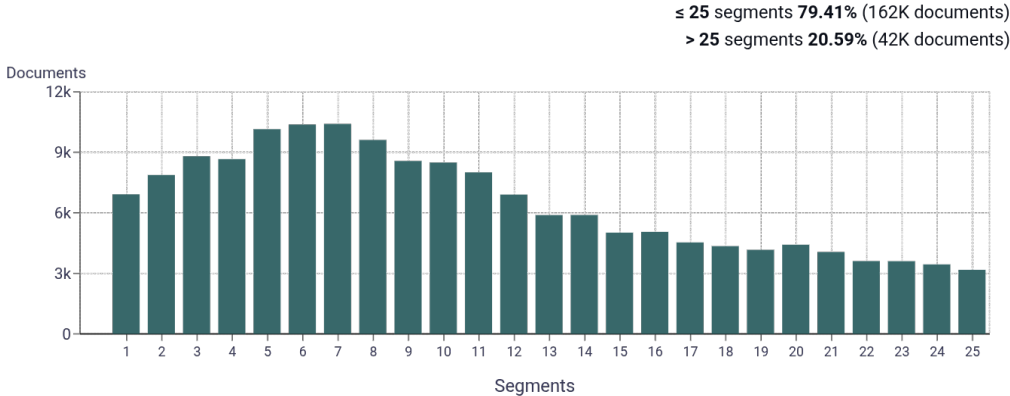
Domain	Docs	% of total
com	117K	57.12%
org	16K	7.89%
co.uk	11K	5.53%
org.uk	9.9K	4.85%
pt	7K	3.43%
zone	5.4K	2.67%
news	3.7K	1.81%
net	3.6K	1.76%
scot	3.2K	1.58%
ac.uk	2.2K	1.07%

Register labels

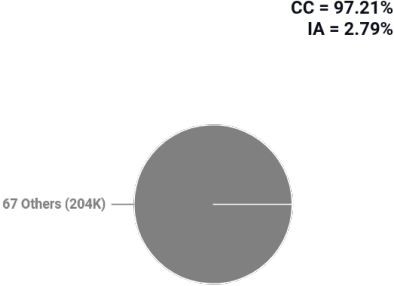


MT:34.6% | 71K Documents

Documents size (in segments) ⓘ

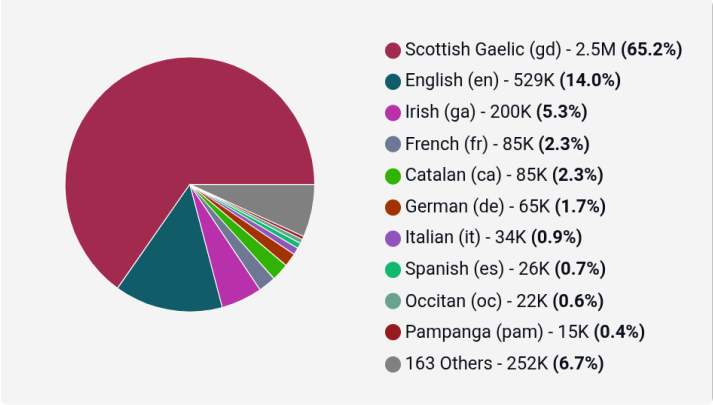


Document collections

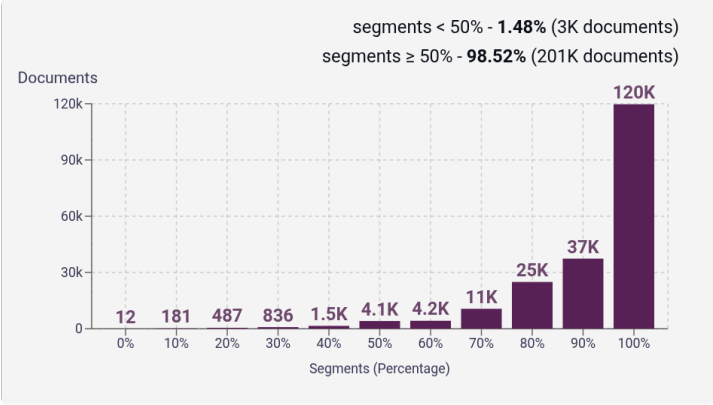


Language Distribution

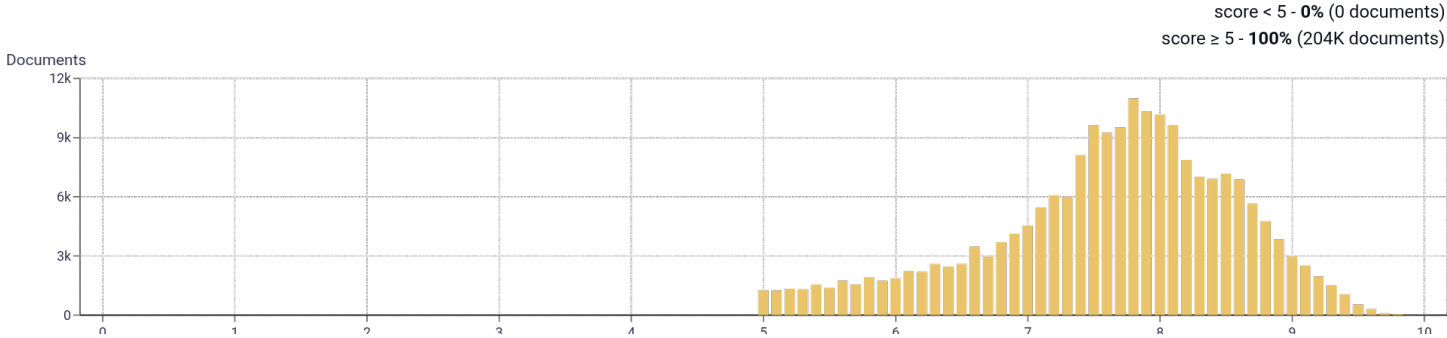
Number of segments in the Scottish Gaelic (gd) corpus



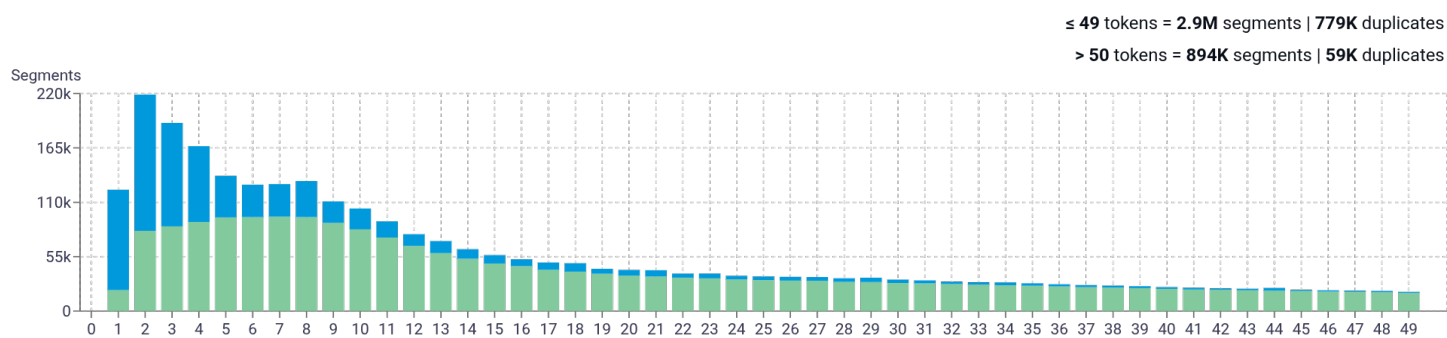
Percentage of segments in Scottish Gaelic (gd) inside documents



Distribution of documents by document score

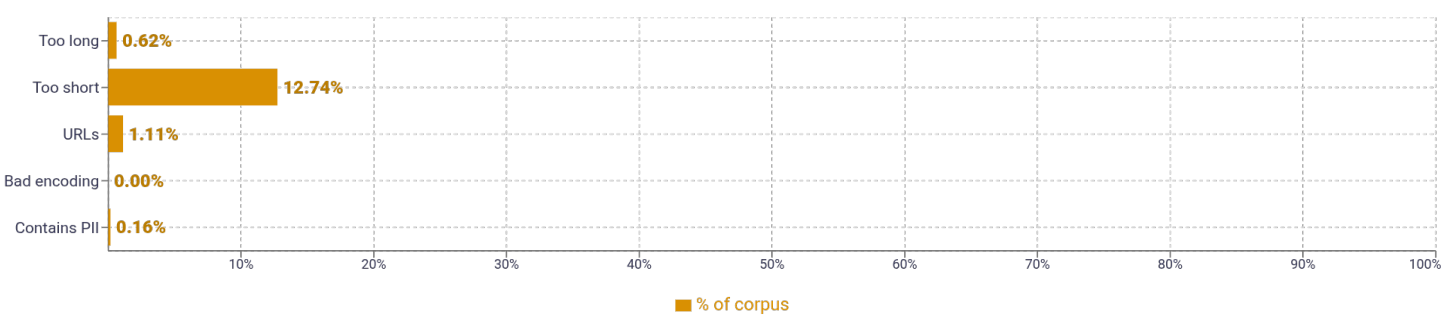


Segment length distribution by token



≤ 49 tokens = 2.9M segments | 779K duplicates
> 50 tokens = 894K segments | 59K duplicates

Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	agus 2,384,134 gu 1,290,205 ann 1,194,038 h 514,489 aig 487,054	
2	gu bheil 272,599 sam bith 112,670 gu math 97,431 gu h 77,763 às deidh 76,886	
3	far a bheil 25,993 cinnteach gu bheil 23,118 aig a bheil 22,331 feadh an t 19,397 àm ri teachd 18,264	
4	aig an aon àm 17,066 dèanamh cinnteach gu bheil 14,104 san àm ri teachd 11,199 ag ràdh gu bheil 8,973 nas fhaide air adhart 7,642	
5	chiad fhear a thog beachd 7,473 gus dèanamh cinnteach gu bheil 6,275 agus mar sin air adhart 4,565 fiosrachaidh mu cheannach is prìsean 3,392 ma tha thu ag iarraidh 3,067	

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , 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				