

General overview

Corpus	Analytics date	Language
cym_Latn.jsonl.tsv	9/16/2024	Welsh (cy)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
758,127	15,568,883	7,875,808 (50.59 %)	491M	2.25 GB	2,387,082,330

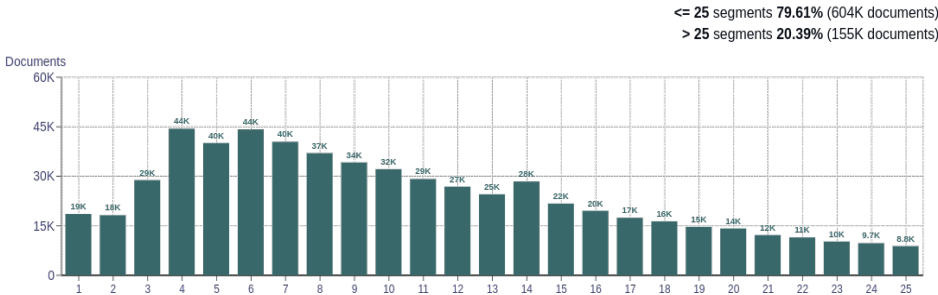
Top 10 domains

Domain	Docs	% of total
wikipedia.org	126K	16.56
bbc.co.uk	25K	3.33
bbc.com	19K	2.49
testunau.org	14K	1.86
cardiff.ac.uk	9.9K	1.30
llyw.cymru	9.9K	1.30
blogspot.com	9.5K	1.25
aber.ac.uk	9K	1.19
soft-free-download.com	7.9K	1.04
playgame24.com	7.3K	0.97

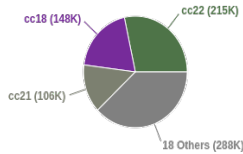
Top 10 TLDs

Domain	Docs	% of total
com	191K	25.17
org	190K	25.03
cymru	101K	13.31
co.uk	65K	8.51
ac.uk	60K	7.89
gov.uk	49K	6.50
org.uk	41K	5.46
wales	23K	2.97
net	11K	1.44
police.uk	3.8K	0.50

Documents size (in segments)

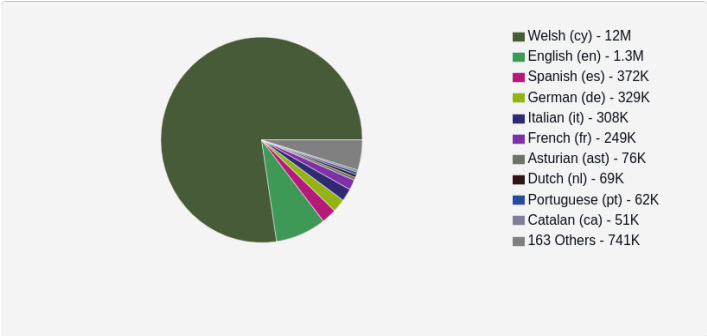


Documents by collection

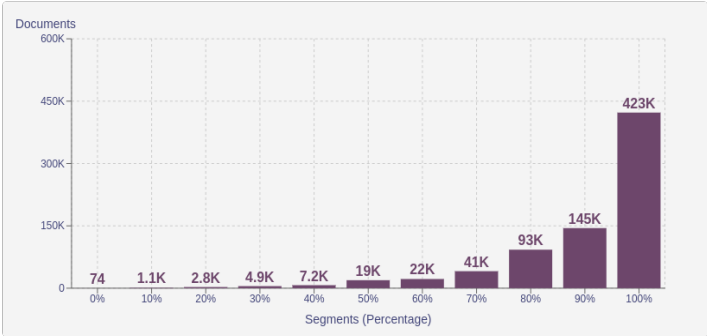


Language Distribution

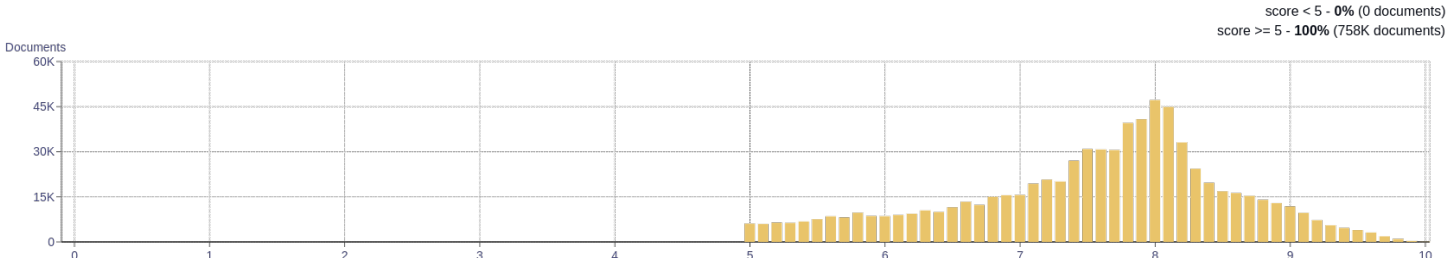
Number of segments



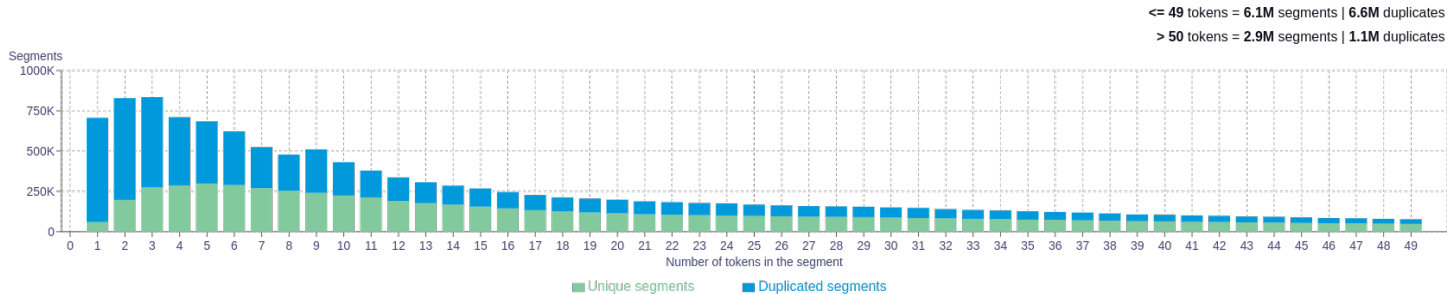
Percentage of segments in Welsh (cy) inside documents



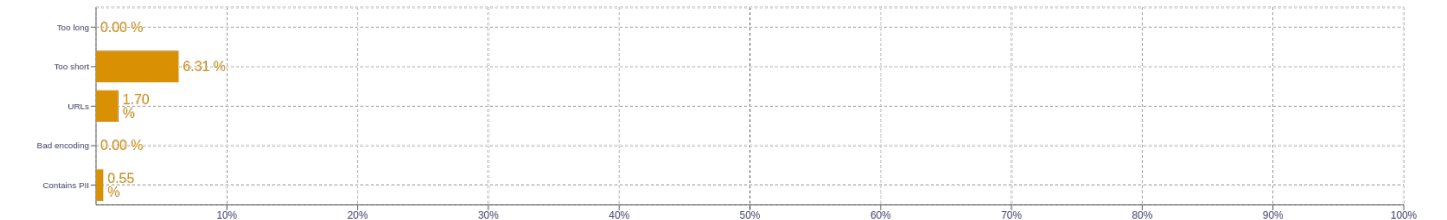
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>r 4049653n 2517313cymru 1146027newydd 675258golygu 512099</div>
2	<div>golygu cod 147766llywodraeth cymru 109475plaid cymru 44567unol daleithiau 42895iechyd meddwl 38558</div>
3	<div>cod y dudalen 147345rhagor o wybodaeth 32955ragor o wybodaeth 19505llywodraeth y du 17851peidiwch ag anghofio 17208</div>
4	<div>golygu cod y dudalen 147331gêm hon gyda 'ch 13000plant a phobl ifanc 12447barn am y gêm 10073peidiwch ag anghofio barn 10062</div>
5	<div>rhannu gêm hon gyda 'ch 12984gêm hon gyda 'ch ffrindiau 12947anghofio barn am y gêm 10062cymru y drindod dewi sant 7145lwytho i lawr y gêm 5286</div>

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>