

General overview

Corpus	Analytics date	Language
cy_1.jsonl.tsv	3/16/2024	Welsh (cy)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
111,254	12,687,508	3,487,441 (27.49 %)	151M	733.16 MB	

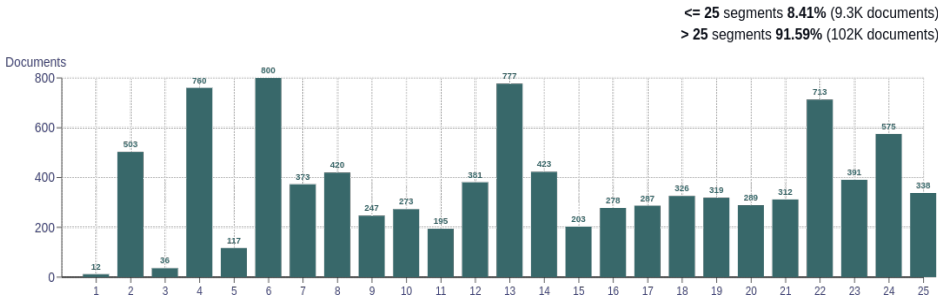
Top 10 domains

Domain	Docs	% of total
wikipedia.org	7.9K	7.12
sgames.org	3.2K	2.88
golwg360.cymru	2.2K	1.97
llyfrgell.cymru	2.1K	1.89
southwales.ac.uk	1.8K	1.59
testunau.org	1.6K	1.44
blogspot.co.uk	1.5K	1.39
whatdotheyknow.com	1.3K	1.21
ofunnygames.com	1.3K	1.15
llyw.cymru	1.2K	1.10

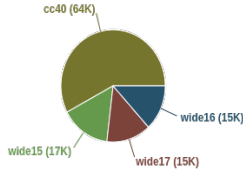
Top 10 TLDs

Domain	Docs	% of total
com	27K	24.37
cymru	25K	22.35
org	21K	19.27
gov.uk	8.9K	8.01
ac.uk	8.3K	7.50
co.uk	7.8K	7.01
org.uk	3.7K	3.32
wales	2.8K	2.56
net	1.4K	1.23
news	676	0.61

Documents size (in segments)

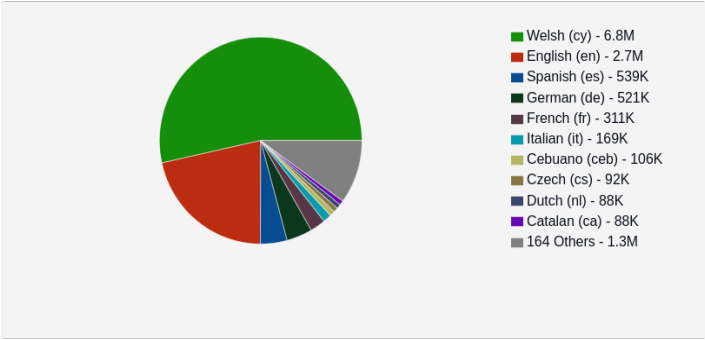


Documents by collection

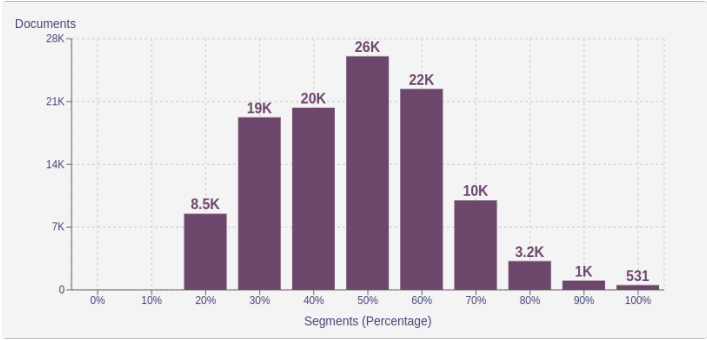


Language Distribution

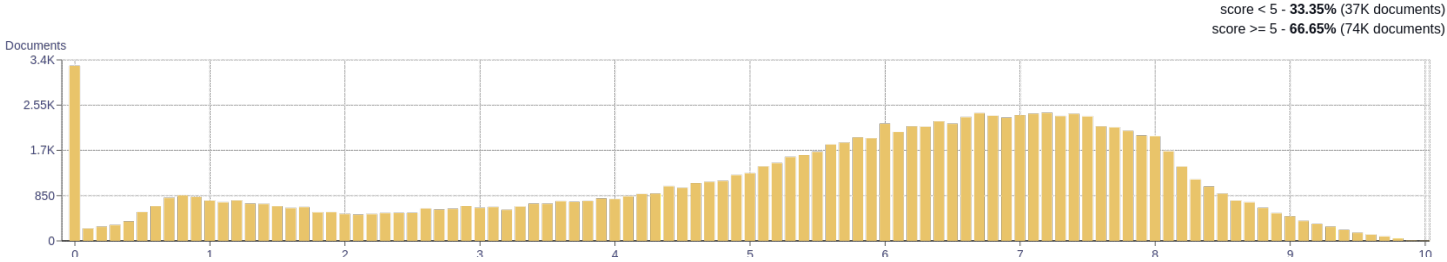
Number of segments



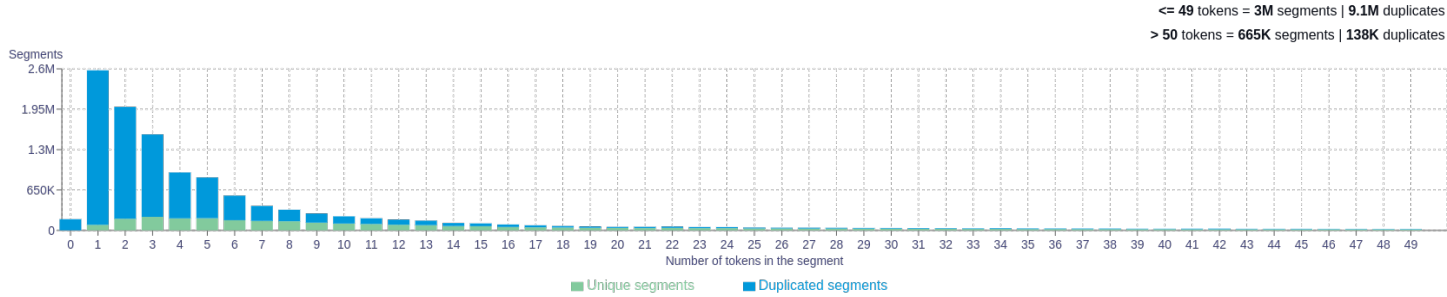
Percentage of segments in Welsh (cy) inside documents



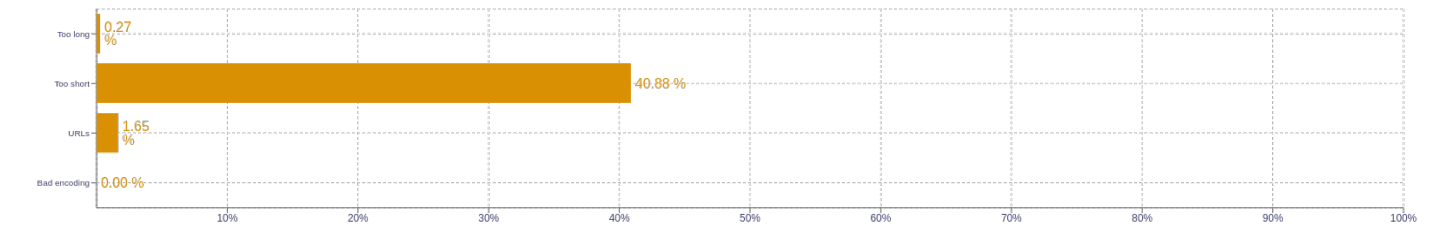
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>r   1130142n   691744cymru   447316newydd   194136gemau   142327</div>
2	<div>llywodraeth cymru   30872polisi preifatrwydd   27898support script   23697senedd chevron   20876gemau ar-lein   20452</div>
3	<div>cod y dudalen   18584telerau ac amodau   18499share to facebook   14947share to twitter   14944copy to clipboard   14331</div>
4	<div>browser does not support   23791golygu cod y dudalen   18512share to twitter share   14943share to facebook share   14334copy to clipboard share   14330</div>
5	<div>browser does not support script   23697clipboard share to facebook share   14328facebook share to twitter share   14327twitter share to linkedin video   9548newidiwyd y dudalen hon ddiwethaf   6644</div>

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 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>, <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>