

General overview

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

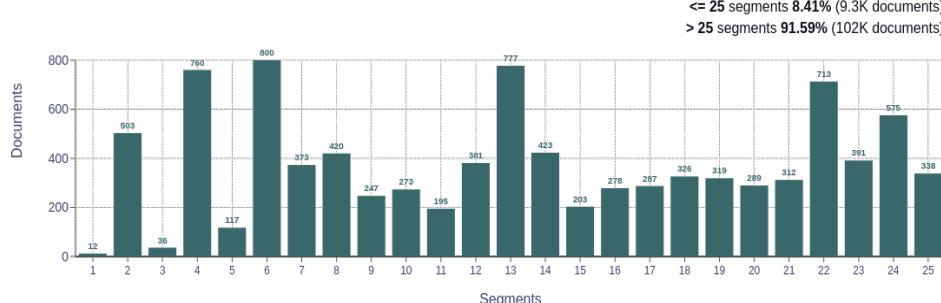
Volumes

Docs	Segments	Unique segments	Tokens	Size
111,254	12,687,508	24,770 (0.20 %)	151M	733.16 MB

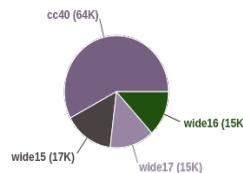
Type-Token Ratio

Welsh (cy)
0.01

Documents size (in segments)

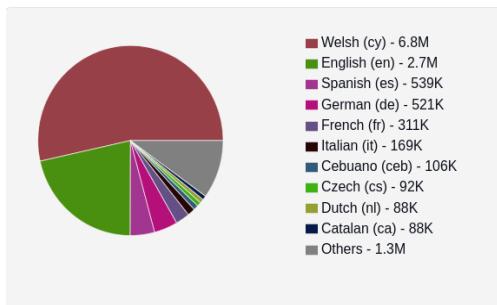


Documents by collection

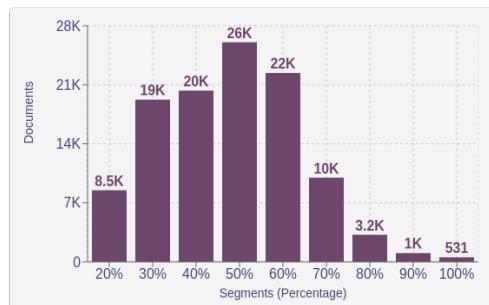


Language Distribution

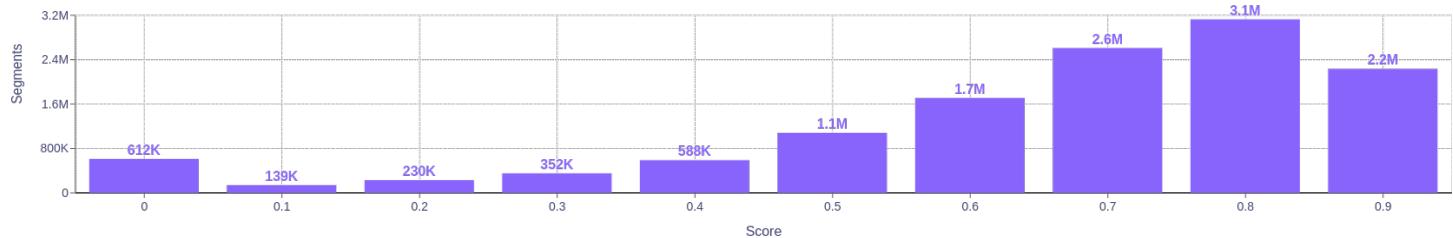
Number of segments



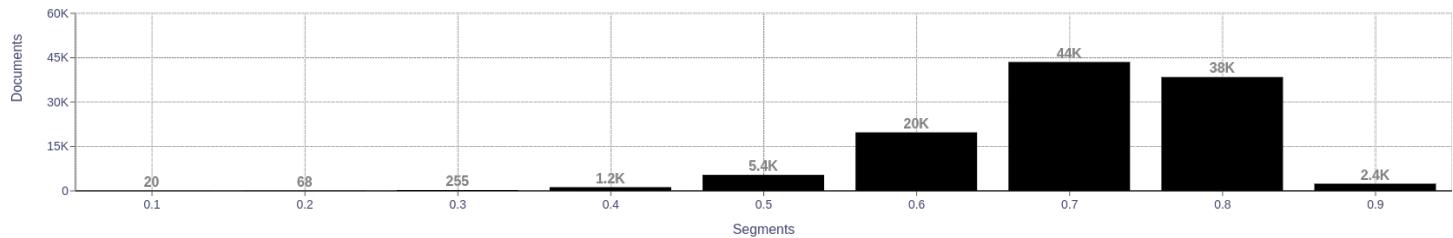
Percentage of segments in Welsh (cy) inside documents



Distribution of segments by fluency score



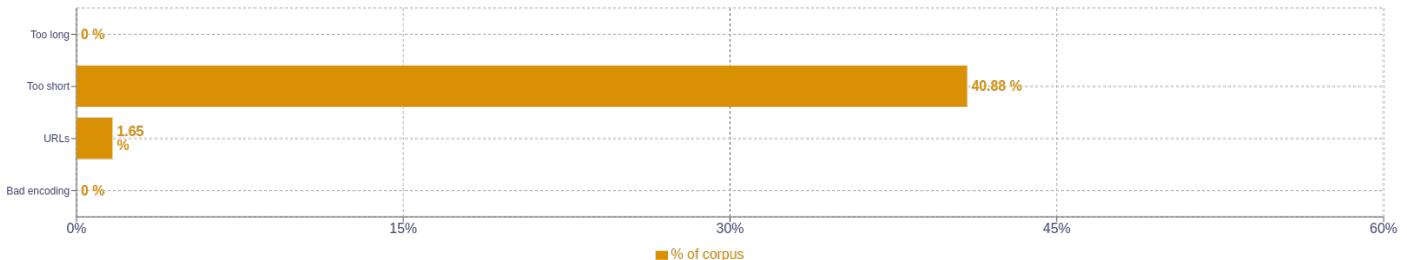
Distribution of documents by average fluency score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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

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>).

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.

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>