

General overview

Corpus	Analytics date	Language
si_1.jsonl.tsv	3/23/2024	Sinhala (si)

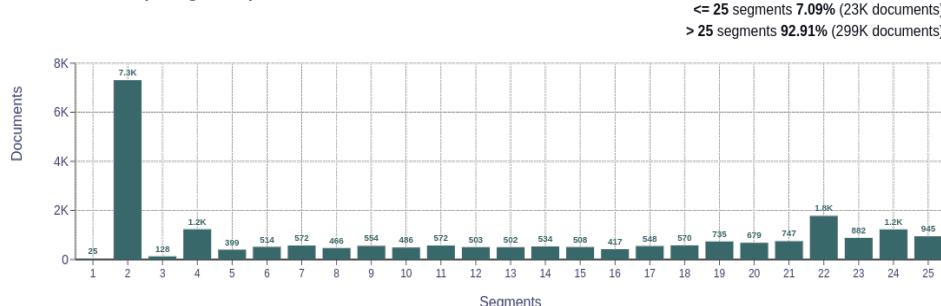
Volumes

Docs	Segments	Unique segments	Tokens	Size
322,515	57,918,718	90,047 (0.16 %)	764M	7.45 GB

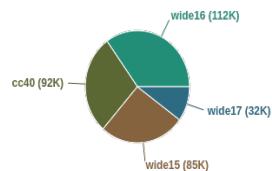
Type-Token Ratio

Sinhala (si)
0.01

Documents size (in segments)

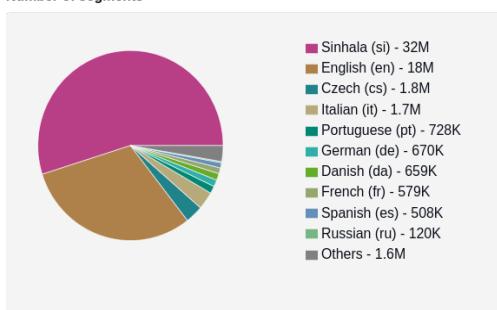


Documents by collection

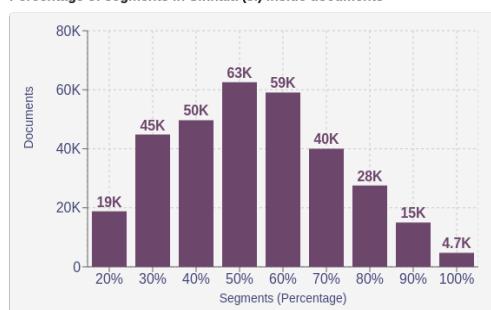


Language Distribution

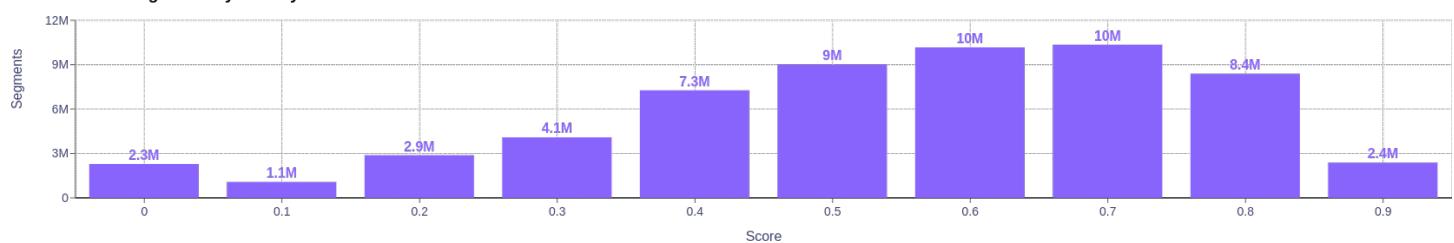
Number of segments



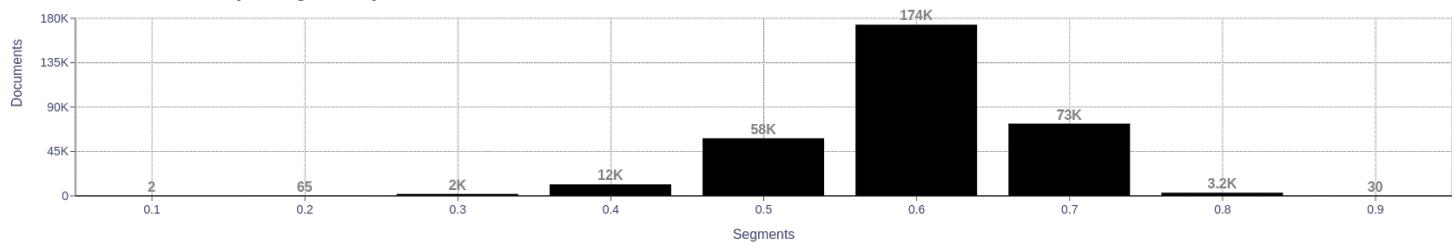
Percentage of segments in Sinhala (si) inside documents



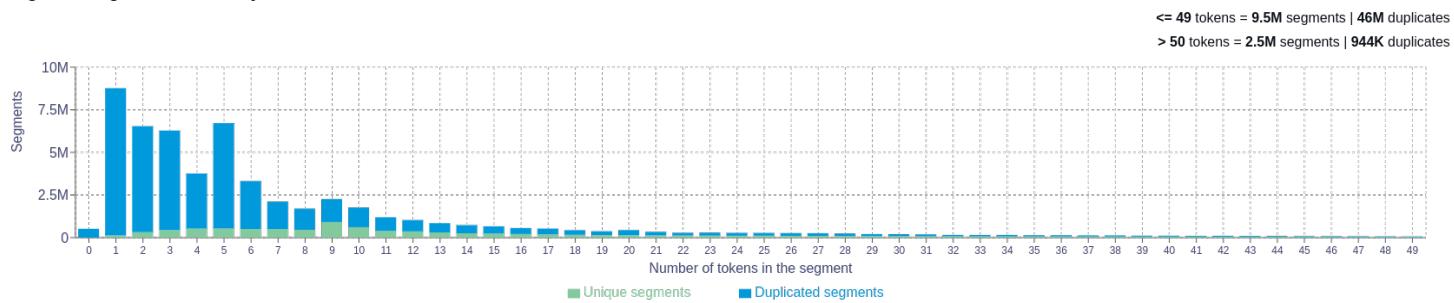
Distribution of segments by fluency score



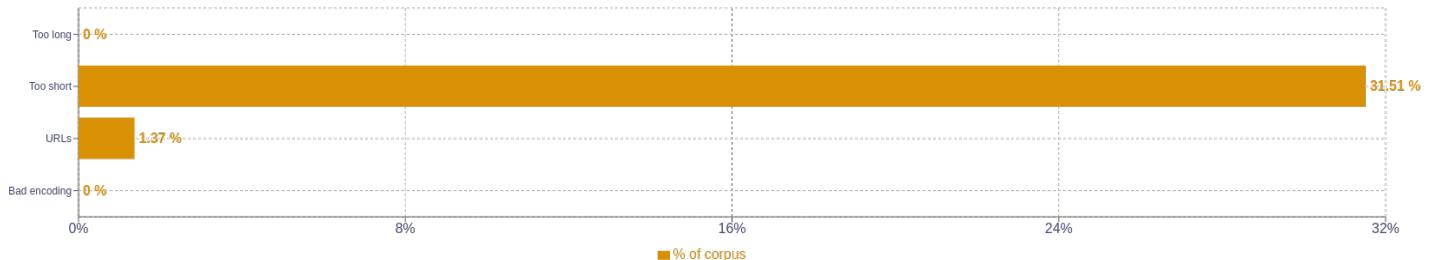
Distribution of documents by average fluency score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	(ago 2483679) (font 1962532) (quot 1806603) (අක 1736470) (style= 1732372)
2	(years ago 639009) (sinhala subtitles 555687) (months ago 546022) (iskoola pota 424779) (>span style= 399335)
3	(with sinhala subtitles 362734) (to twittershare to 216822) (share to twittershare 216817) (to facebookshare to 207116) (twittershare to facebookshare 207099)
4	(share to twittershare to 216817) (to twittershare to facebookshare 207099) (twittershare to facebookshare to 207096) (to facebookshare to pinterest 205587) විෂයට අදාළ මියන ලදී 125349
5	(to twittershare to facebookshare to 207096) (share to twittershare to facebookshare 207095) (twittershare to facebookshare to pinterest 205587) කාන්තෙනය විෂයට අදාළ මියන ලදී 104333) (සන්නිවේදන කාන්තෙනය විෂයට අදාළ මියන 101488)

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/mabanon/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>