

General overview

Corpus	Analytics date	Language
mr_1.jsonl.tsv	3/25/2024	Marathi (mr)

Volumes

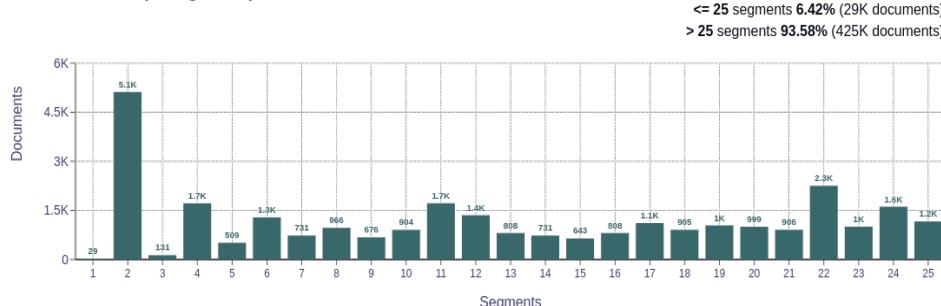
Docs	Segments	Unique segments	Tokens	Size
453,694	56,430,804	53,993 (0.10 %)	647M	7.56 GB

Type-Token Ratio

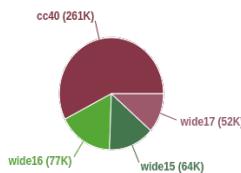
Marathi (mr)

0.01

Documents size (in segments)

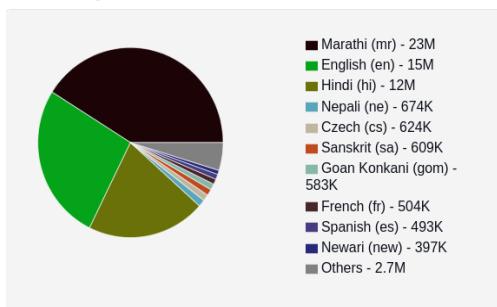


Documents by collection

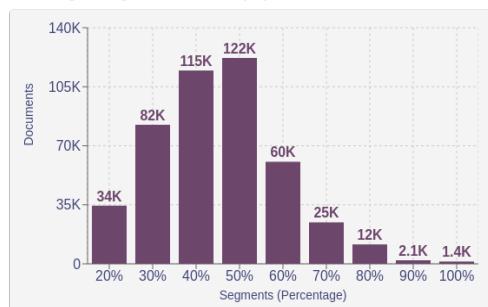


Language Distribution

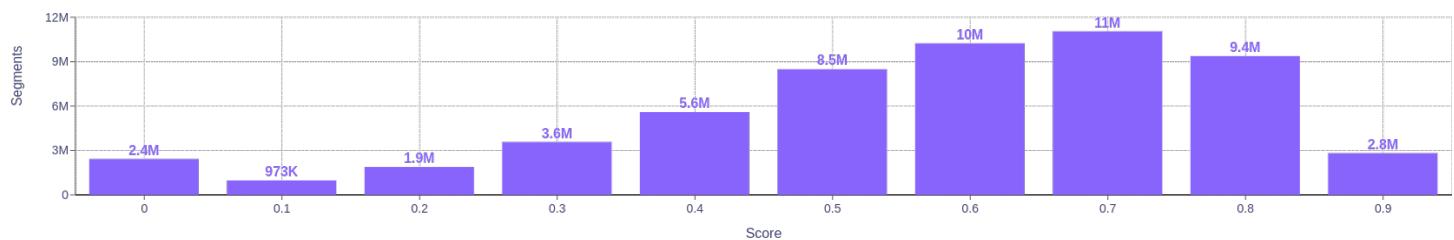
Number of segments



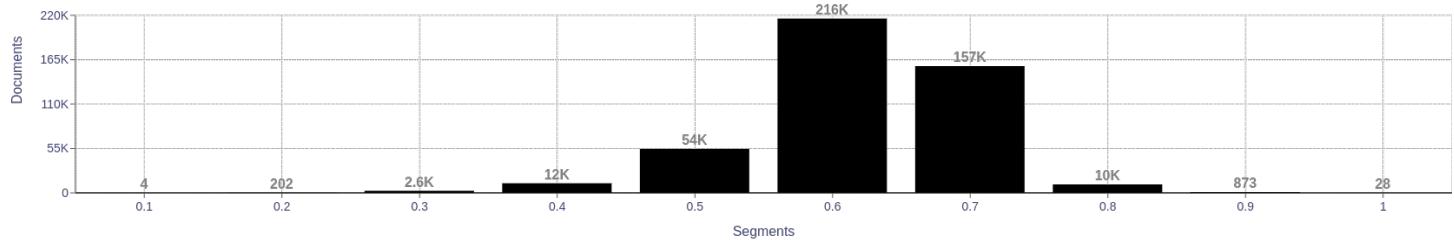
Percentage of segments in Marathi (mr) inside documents



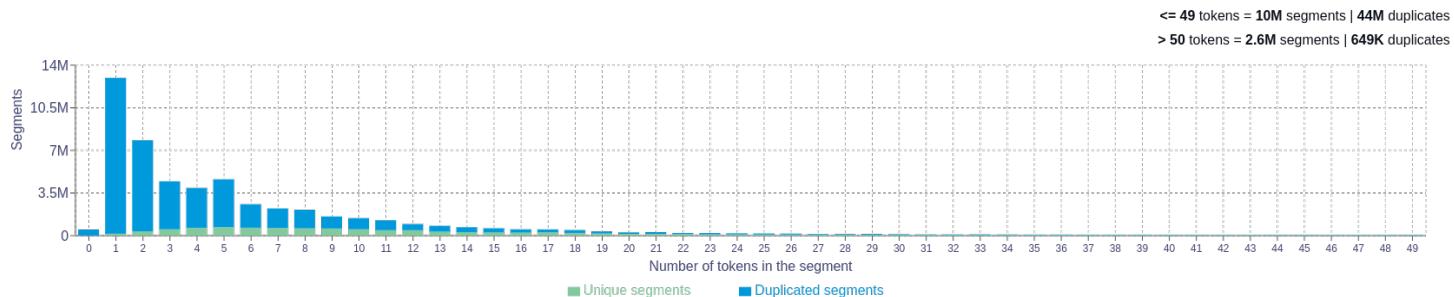
Distribution of segments by fluency score



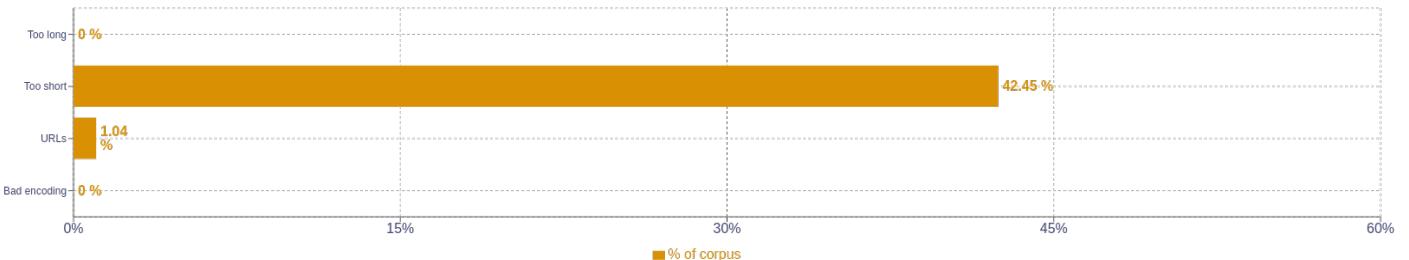
Distribution of documents by average fluency score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	(to 1598348) (in 1515505) (the 1052996) (by 987369) (marathi 980322)
2	(in marathi 312738) (log in 278829) (post comments 258450) (to post 247112) (or register 238207)
3	(to post comments 239262) (or register to 237958) (log in or 237830) (register to post 237727) (in or register 237316)
4	(or register to post 237727) (register to post comments 237725) (log in or register 237305) (in or register to 237303) (opens in new window 118175)
5	(or register to post comments 237725) (log in or register to 237303) (in or register to post 237116) (to twittershare to facebookshare to 89686)
	share to twittershare to facebookshare 89686

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>