

## General overview

Corpus	Analytics date	Language
ml_1.jsonl.tsv	3/23/2024	Malayalam (ml)

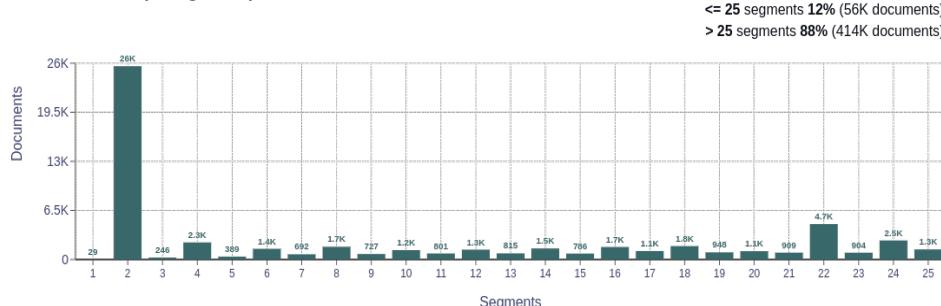
## Volumes

Docs	Segments	Unique segments	Tokens	Size
469,980	57,033,693	53,375 (0.09 %)	633M	9.96 GB

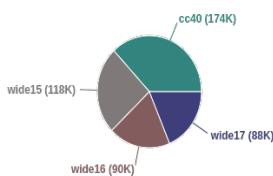
## Type-Token Ratio

Malayalam (ml)
0.03

## Documents size (in segments)

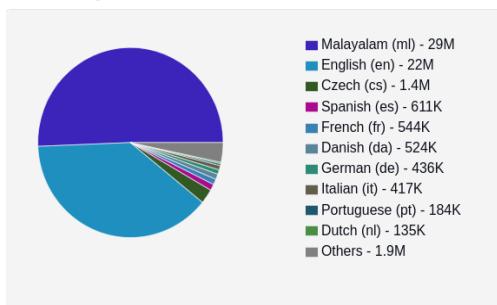


## Documents by collection

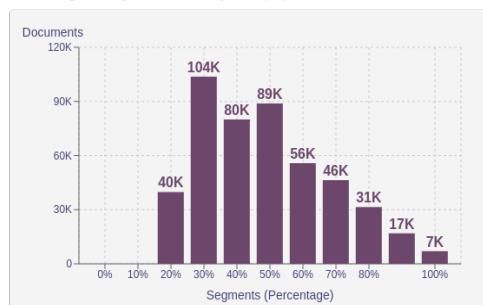


## Language Distribution

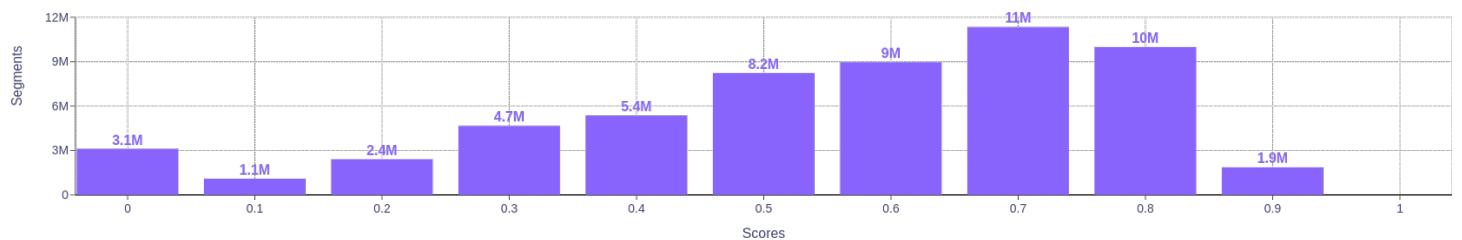
## Number of segments



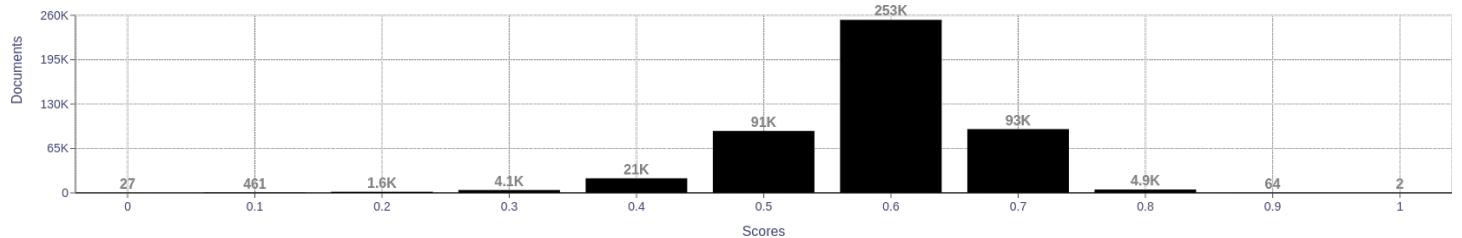
## Percentage of segments in Malayalam (ml) inside documents



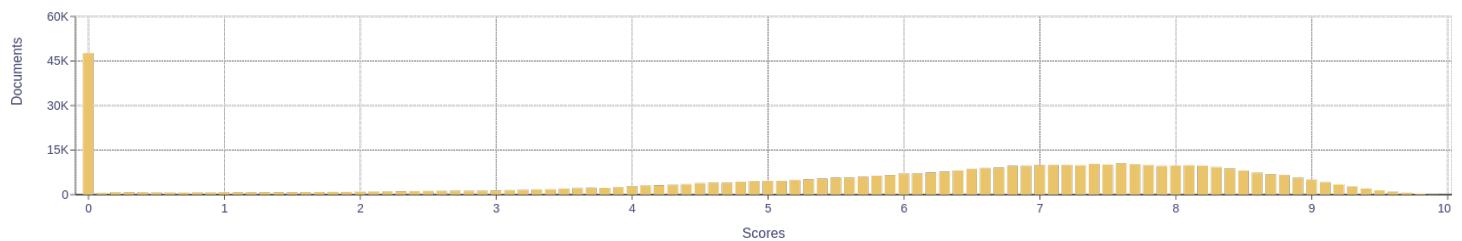
## Distribution of segments by fluency score



## Distribution of documents by average fluency score

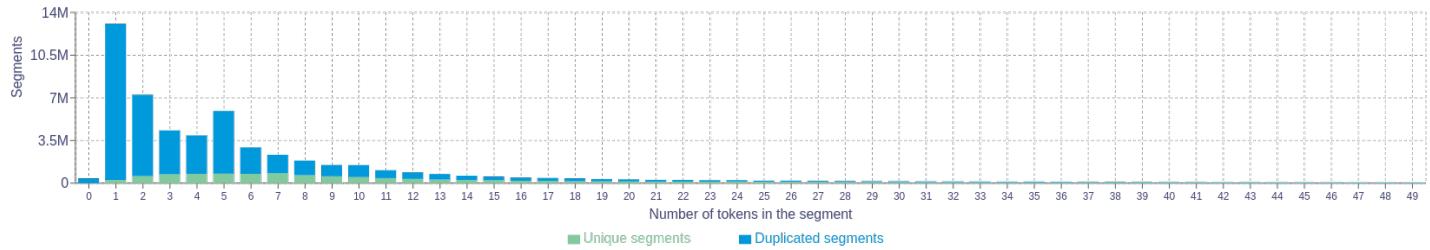


## Distribution of documents by document score

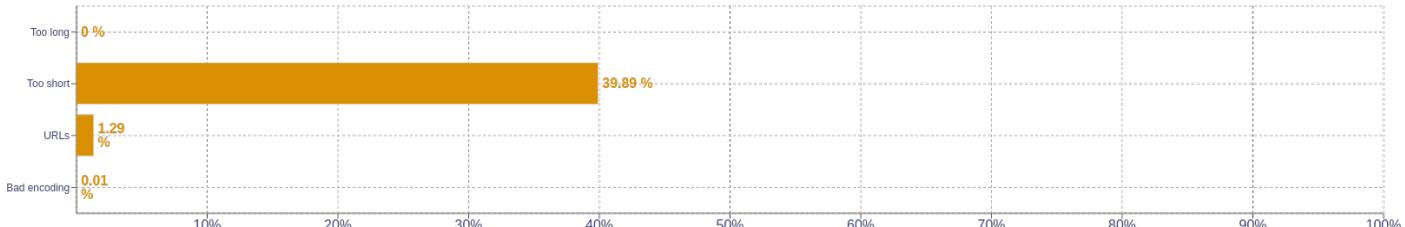


## Segment length distribution by token

<= 49 tokens = 11M segments | 44M duplicates  
 > 50 tokens = 2M segments | 622K duplicates



## Segment noise distribution



## Frequent n-grams

Size	n-grams
1	(em   3754966) (en   2090109) (to   1762323) (the   1650615) (news   1601203)
2	(span style=   275173) (read more   265772) (posted by   258577) (of the   203874) (about us   194597)
3	(to twittershare to   177459) (share to twittershare   177459) (twittershare to facebookshare   172892) (to facebookshare to   172892) (facebookshare to pinterest   172892)
4	(share to twittershare to   177459) (twittershare to facebookshare to   172892) (to twittershare to facebookshare   172892) (to facebookshare to pinterest   172892) (links to this post   78281)
5	(twittershare to facebookshare to pinterest   172892) (to twittershare to facebookshare to   172892) (share to twittershare to facebookshare   172892) (en facebook en facebook en facebook   36035) (twitter en facebook en facebook   36035)

## 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 or 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>