

## General overview

Corpus	Analytics date	Language
my_1.jsonl.tsv	3/26/2024	Burmese (my)

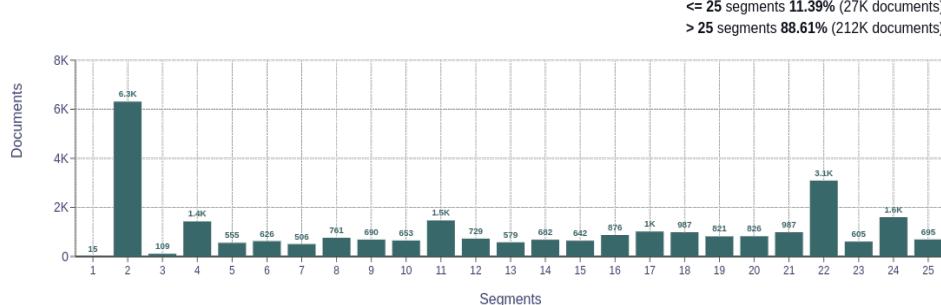
## Volumes

Docs	Segments	Unique segments	Tokens	Size
239,473	47,772,618	68,061 (0.14 %)	501M	8.0 GB

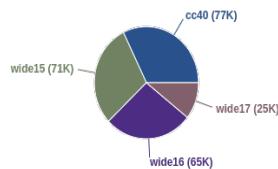
## Type-Token Ratio

Burmese (my)
0.07

## Documents size (in segments)

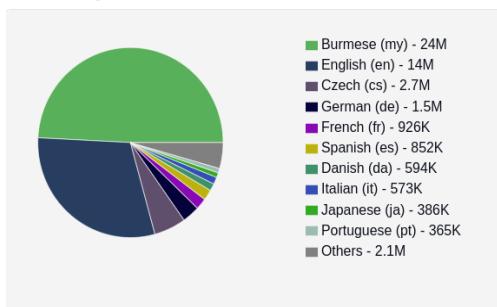


## Documents by collection

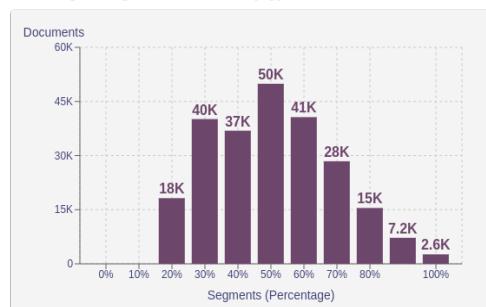


## Language Distribution

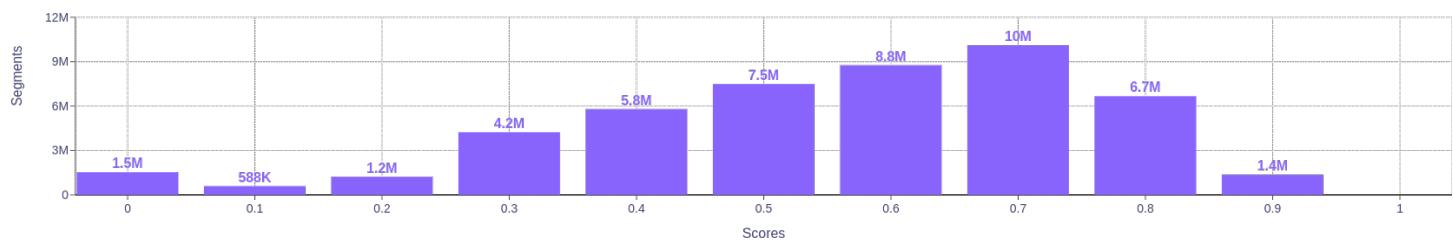
## Number of segments



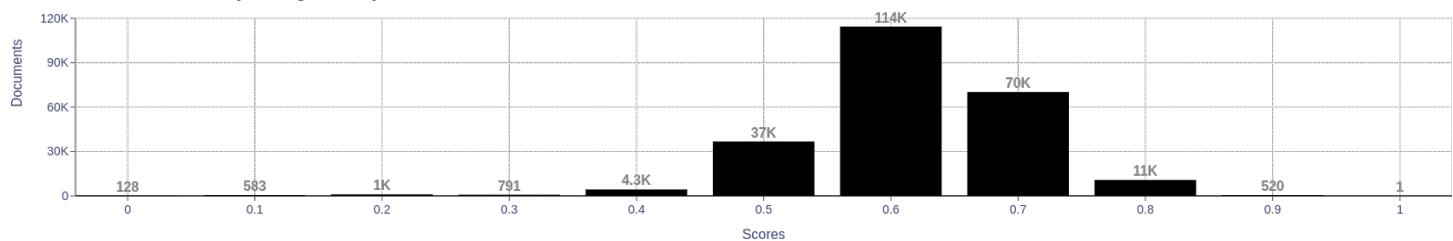
## Percentage of segments in Burmese (my) inside documents



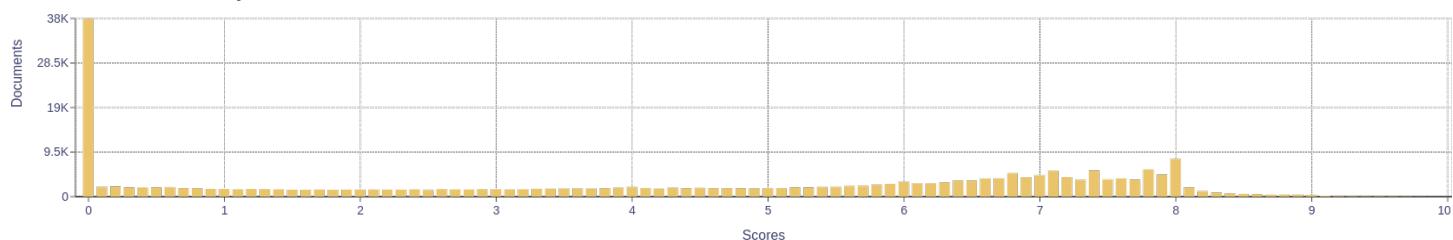
## Distribution of segments by fluency score



## Distribution of documents by average fluency score

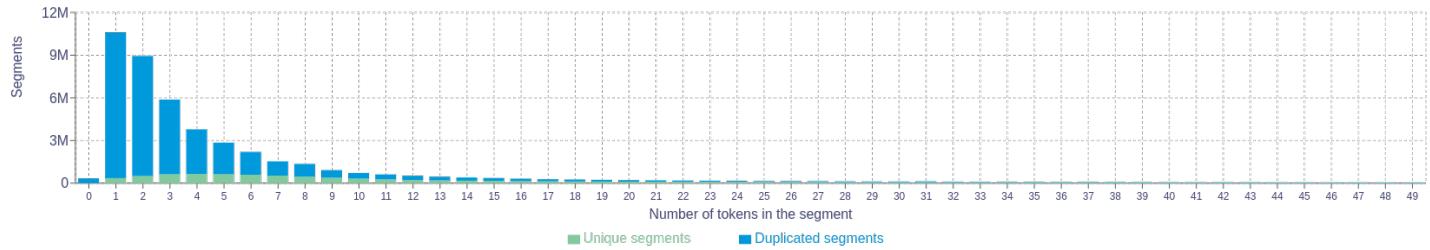


## Distribution of documents by document score

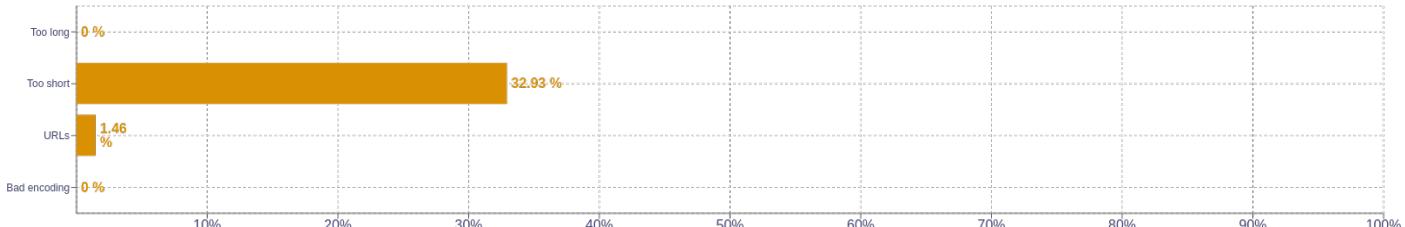


## Segment length distribution by token

<= 49 tokens = 8.6M segments | 37M duplicates  
> 50 tokens = 1.8M segments | 568K duplicates



## Segment noise distribution



## Frequent n-grams

Size	n-grams
1	ମୁଁ   8333926 କାନ୍ତି   8297226 ପାଇଁ   7883766 କାନ୍ତି   3850051 କାନ୍ତି   3547059
2	ପାଇଁ କାନ୍ତି   1665682 କାନ୍ତି କାନ୍ତି   656538 କାନ୍ତି କାନ୍ତି   540256 କାନ୍ତି କାନ୍ତି   422091 କାନ୍ତି କାନ୍ତି   415393
3	lsd exception locked   270833 this blog this!   219870 blog this! sharetotwitter   219860 this! sharetotwitter shareoffacebook   219825 email this blog   219644
4	this blog this! sharetotwitter   219846 blog this! sharetotwitter shareoffacebook   219825 email this blog this!   219644 this! sharetotwitter shareoffacebook sharetop   197889 sharetotwitter shareoffacebook sharetop interest   197888
5	this blog this! sharetotwitter shareoffacebook   219811 email this blog this! sharetotwitter   219642 blog this! sharetotwitter shareoffacebook sharetop   197889 this! sharetotwitter shareoffacebook sharetop interest   197888 newer post older post home   41638

## 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/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>).

### 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>