

## General overview

Corpus	Analytics date	Language
pa_1.jsonl.tsv	3/17/2024	Punjabi (pa)

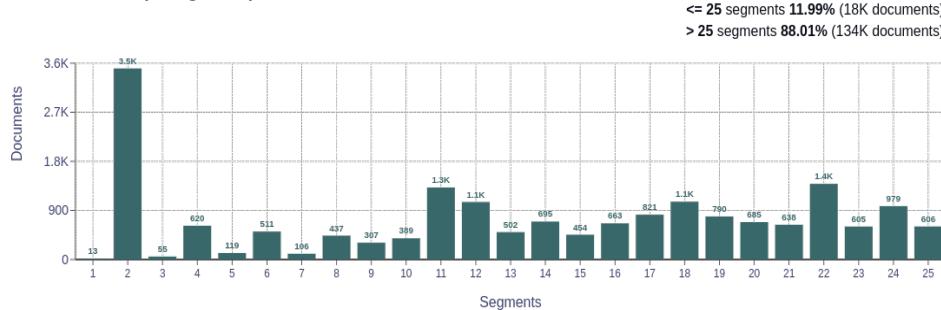
## Volumes

Docs	Segments	Unique segments	Tokens	Size
152,775	17,180,972	23,512 (0.14 %)	219M	2.03 GB

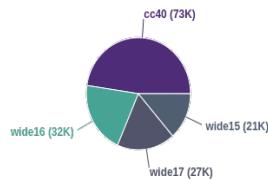
## Type-Token Ratio

Punjabi (pa)
0.01

## Documents size (in segments)

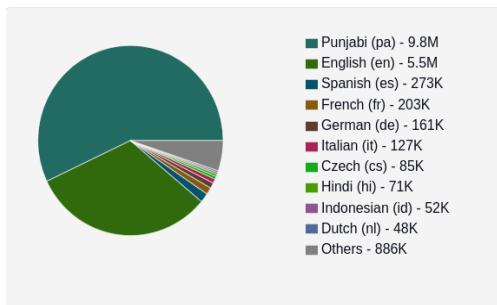


## Documents by collection

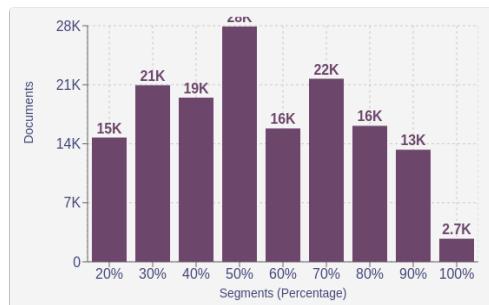


## Language Distribution

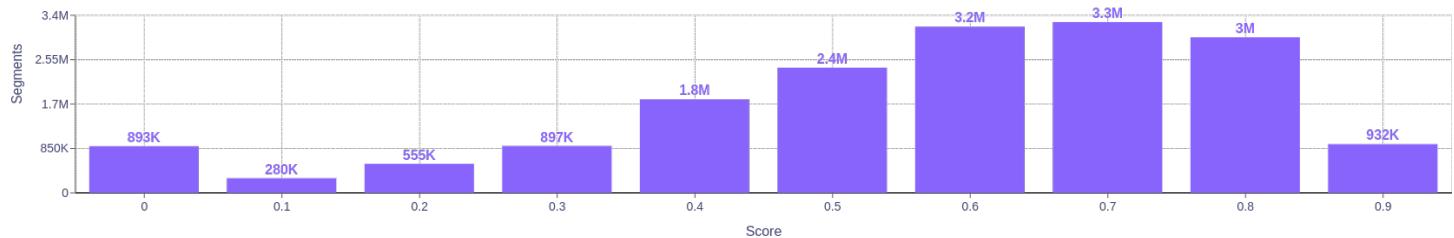
## Number of segments



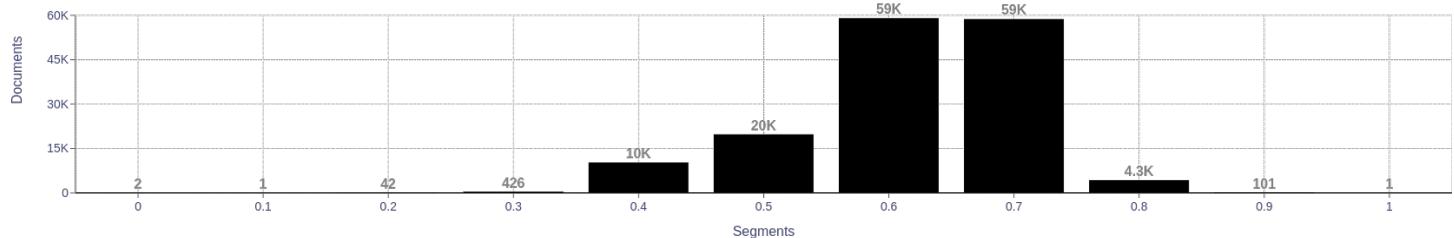
## Percentage of segments in Punjabi (pa) inside documents



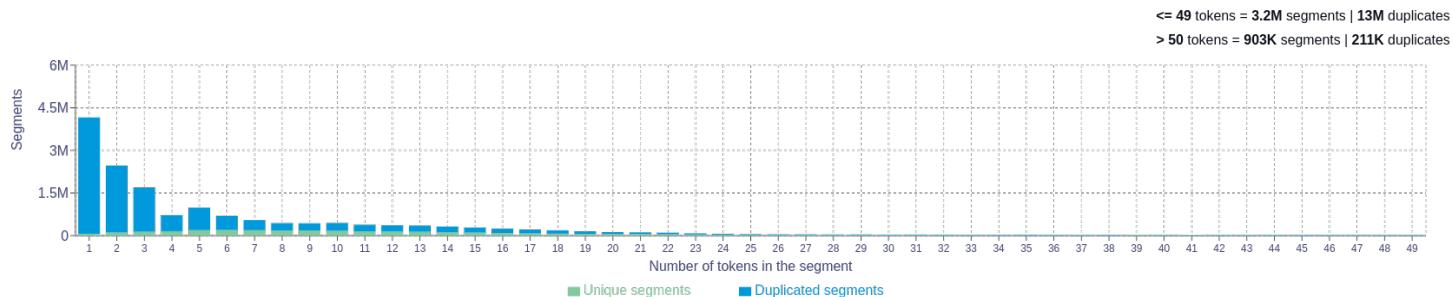
## Distribution of segments by fluency score



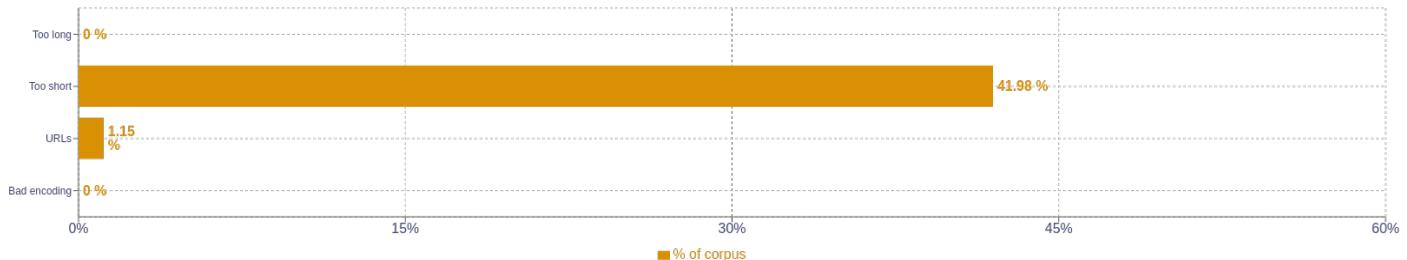
## Distribution of documents by average fluency score



## Segment length distribution by token



## Segment noise distribution



## Frequent n-grams

Size	n-grams
1	(the   770337) (to   535359) (of   442946) (and   424906) (news   417666)
2	(of the   101145) (all rights   70140) (rights reserved   69793) (contact us   67417) (read more   61786)
3	(all rights reserved   69688) (to twittershare to   27188) (share to twittershare   27188) (twittershare to facebookshare   26828) (to facebookshare to   26828)
4	(share to twittershare to   27188) (twittershare to facebookshare to   26828) (to twittershare to facebookshare   26828) (to facebookshare to pinterest   26828) (leave a reply cancel   19968)
5	(twittershare to facebookshare to pinterest   26828) (to twittershare to facebookshare to   26828) (share to twittershare to facebookshare   26828) (leave a reply cancel reply   19949) (your email address will not   16384)

## 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>).

### 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>