

## General overview

Corpus	Analytics date	Language
af_1.jsonl.tsv	3/21/2024	Afrikaans (af)

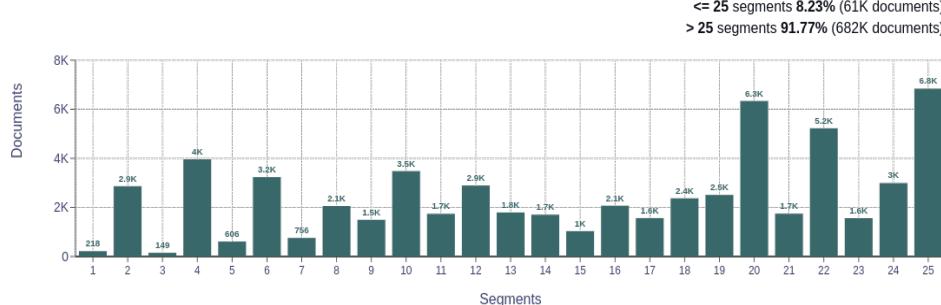
## Volumes

Docs	Segments	Unique segments	Tokens	Size
747,229	84,701,482	50,107 (0.06 %)	1B	4.99 GB

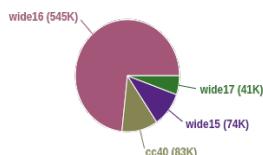
## Type-Token Ratio

Afrikaans (af)
0.01

## Documents size (in segments)

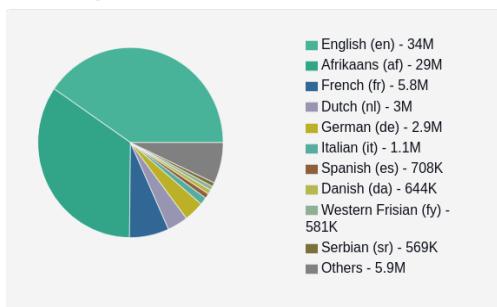


## Documents by collection

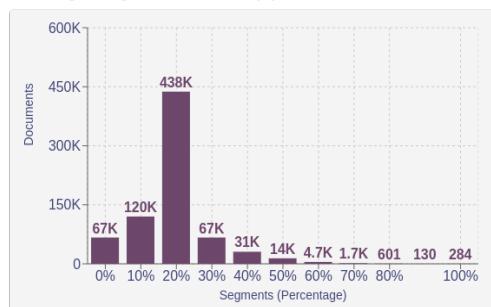


## Language Distribution

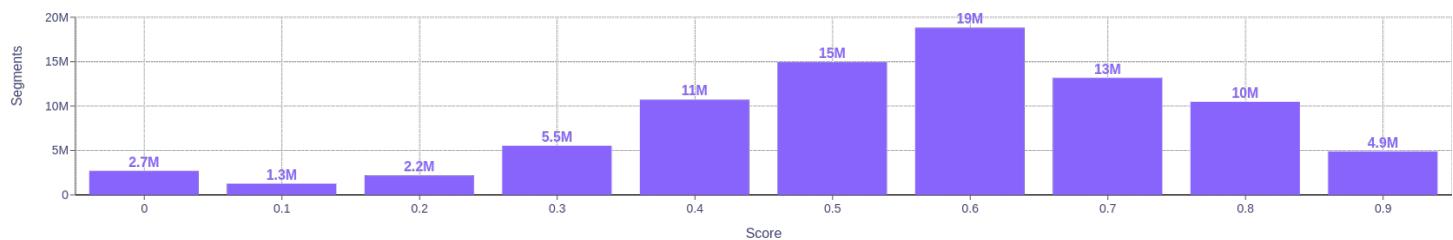
## Number of segments



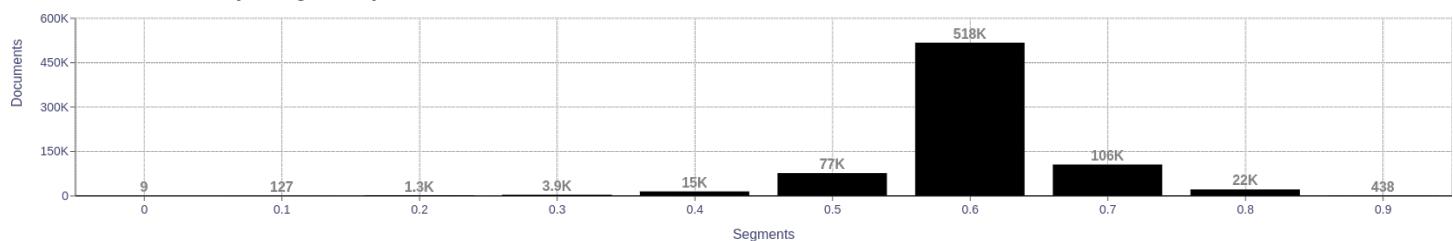
## Percentage of segments in Afrikaans (af) inside documents



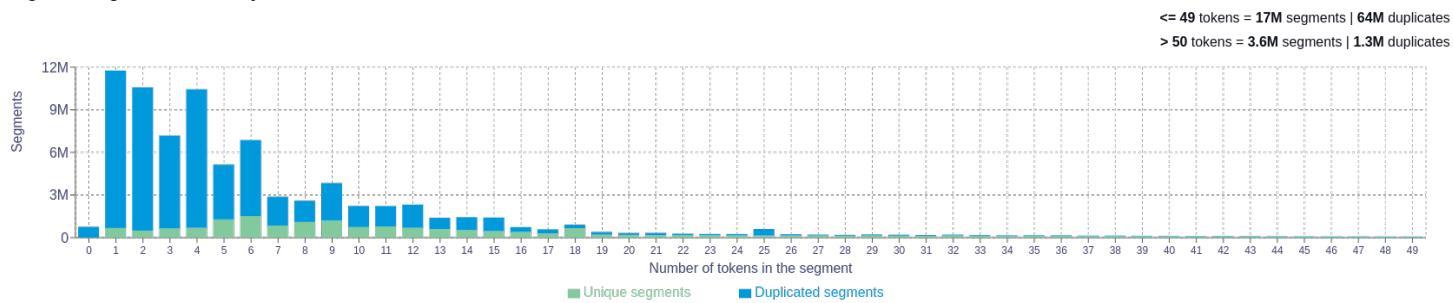
## Distribution of segments by fluency score



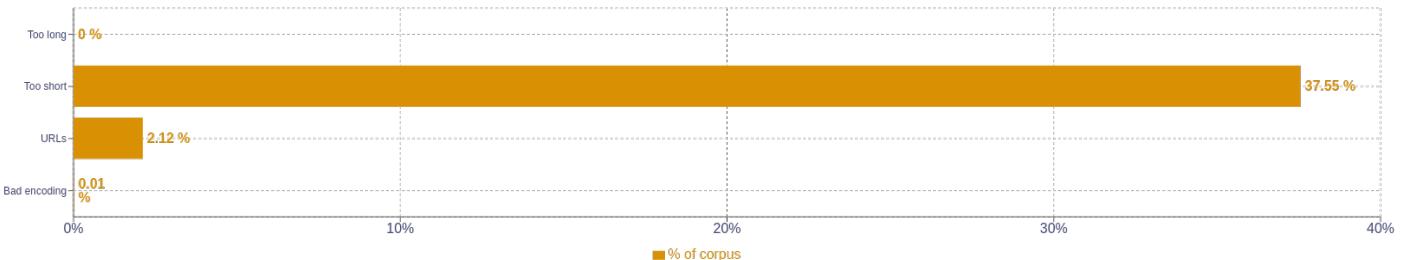
## Distribution of documents by average fluency score



## Segment length distribution by token



## Segment noise distribution



## Frequent n-grams

Size	n-grams
1	(the   13700283) (of   12108285) (and   8106124) (to   5890313) (a   4994182)
2	(of the   2027007) (meer besonderhede   1355889) (alternatiewe skryfwyses   1352488) (kyk boek   1346521) (stoer boek   1346517)
3	(to my favourites   410387) (made by freepik   408232) (freepik from www.flaticon.com   408232) (www.flaticon.com is licensed   408231) (licensed by cc   408231)
4	(add to my favourites   410386) (made by freepik from   408232) (icons made by freepik   408232) (from www.flaticon.com is licensed   408231) (oor die boek soek   404630)
5	(made by freepik from www.flaticon.com   408232) (icons made by freepik from   408232) (www.flaticon.com is licensed by cc   408231) (freepik from www.flaticon.com is licensed   408231) (united kingdom australia new zealand   387419)

## 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/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>