

## General overview

| Corpus         | Analytics date | Language    |
|----------------|----------------|-------------|
| so_1.jsonl.tsv | 3/17/2024      | Somali (so) |

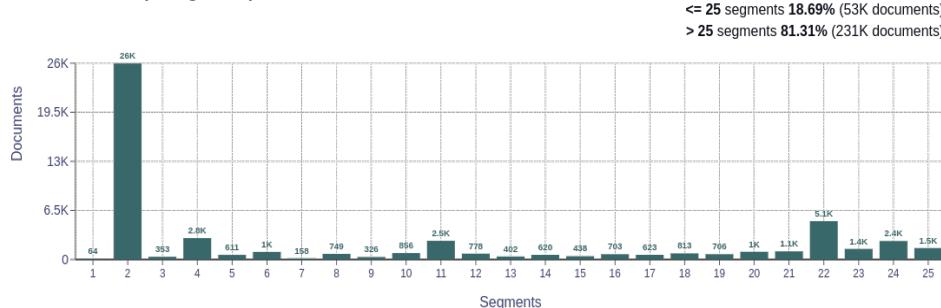
## Volumes

| Docs    | Segments   | Unique segments | Tokens | Size    |
|---------|------------|-----------------|--------|---------|
| 283,712 | 23,606,211 | 22,713 (0.10 %) | 249M   | 1.32 GB |

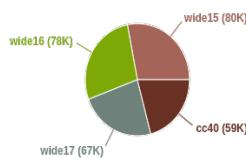
## Type-Token Ratio

|             |
|-------------|
| Somali (so) |
| 0.01        |

## Documents size (in segments)

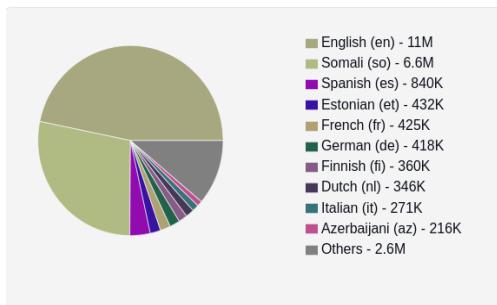


## Documents by collection

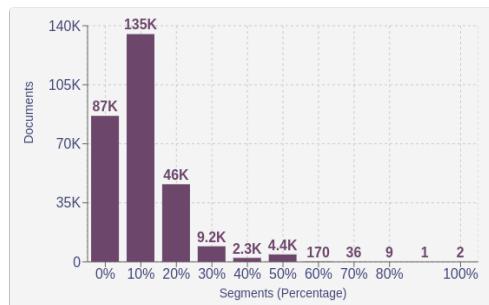


## Language Distribution

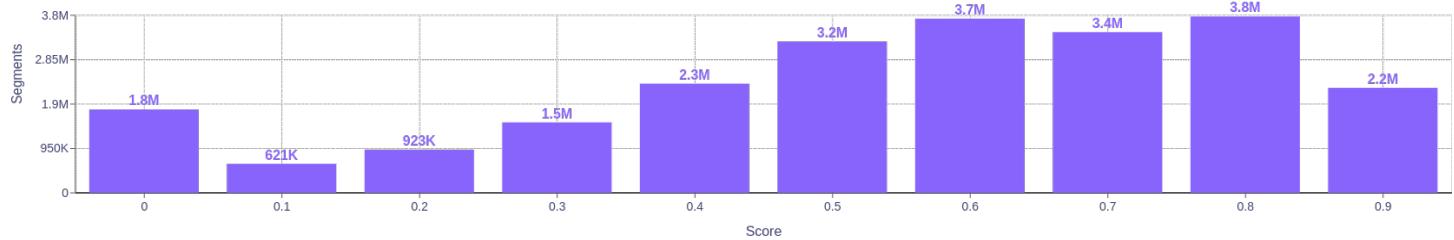
## Number of segments



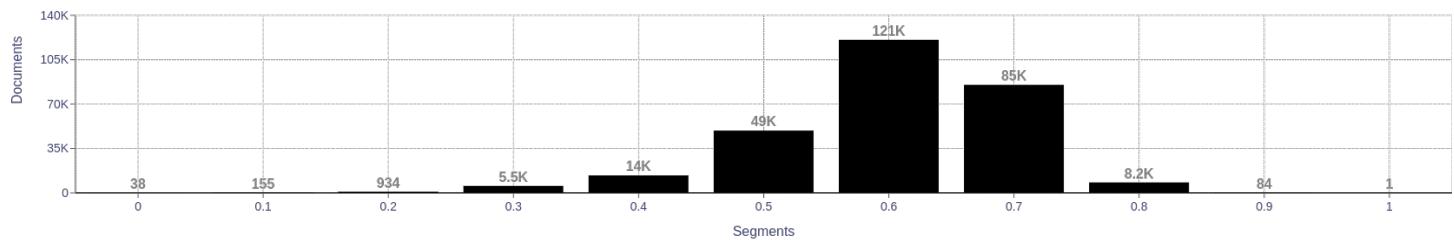
## Percentage of segments in Somali (so) inside documents



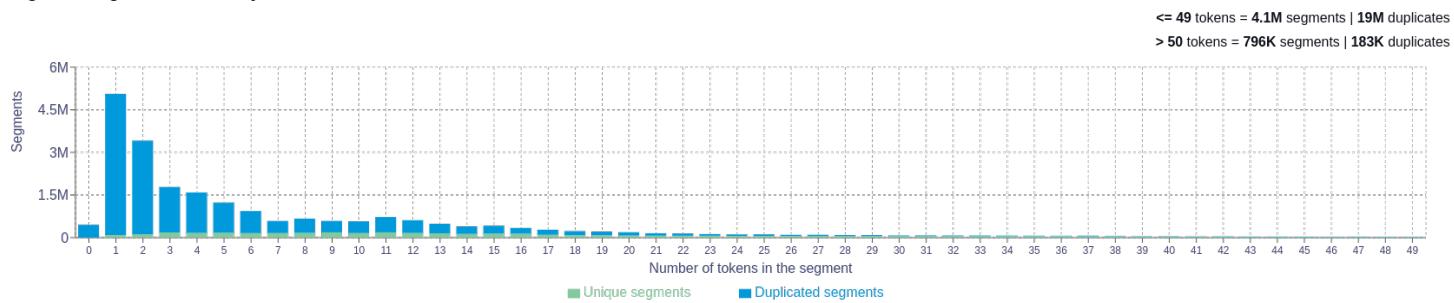
## Distribution of segments by fluency score



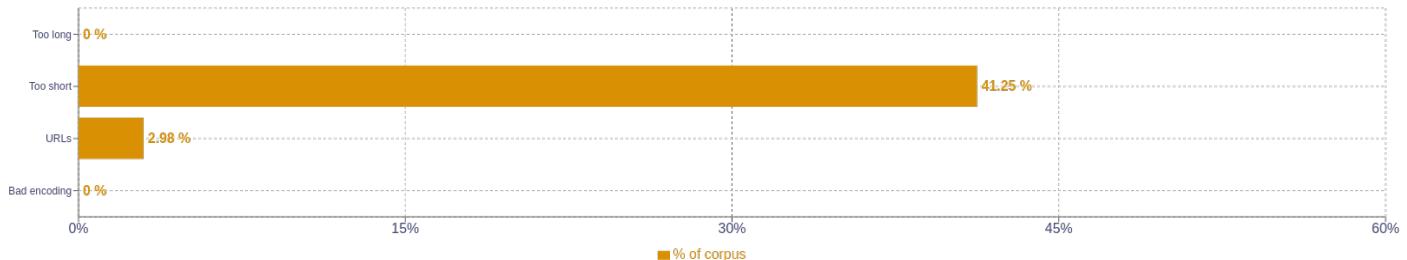
## Distribution of documents by average fluency score



## Segment length distribution by token



## Segment noise distribution



## Frequent n-grams

| Size | n-grams  |
|------|--|
| 1    | iyo   2763688 ee   2524139 ah   2253410 u   1953608 la   1517808   |
| 2    | read more   186488 ah ee   180934 mid ah   163683 contact us   149776 of the   149575  |
| 3    | all rights reserved   133041 opens in new   102073 click to share   91413 to share on   91405 news in english   83799  |
| 4    | opens in new window   102059 click to share on   91401 log into your account   33422 leave a reply cancel   30785 a reply cancel reply   30576               |
| 5    | leave a reply cancel reply   30569 of new posts by email   24673 click to share on twitter   24433 notify me of new posts   24417 me of new posts by   24266 |

## 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>