# HPLT Analytics report

HPLTAnalytics

## General overview

| Corpus | Analytics date | Language |
|---|---|---|
| HPLT-docslite.da.tsv | 6/9/2024 | Danish (da) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Size |
|---|---|---|---|---|
| 5,979,720 | 726,820,478 | 75,888 (0.01 %) | 8.3B | 43.17 GB |

## Top 10 domains

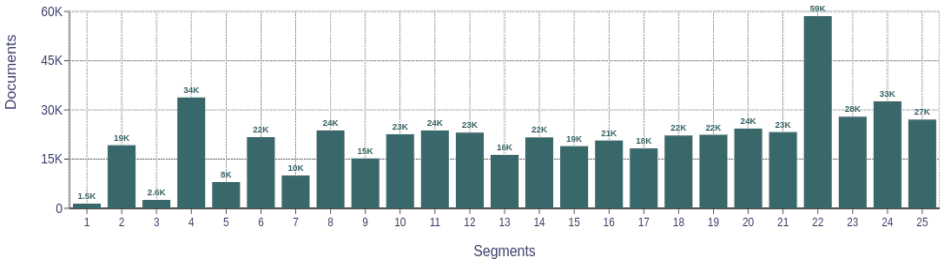
| Domain | Docs | % of total |
|---|---|---|
| blogspot.dk | 204K | 3.41 |
| docplayer.dk | 186K | 3.12 |
| blogspot.com | 41K | 0.69 |
| toppriser.dk | 36K | 0.61 |
| wikipedia.org | 22K | 0.36 |
| wordpress.com | 22K | 0.36 |
| blogspot.no | 19K | 0.32 |
| blogspot.de | 18K | 0.30 |
| skagensavis.dk | 18K | 0.30 |
| pris-billig.dk | 18K | 0.29 |

## Top 10 TLDs

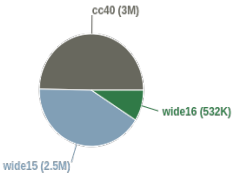
| Domain | Docs | % of total |
|---|---|---|
| dk | 4.9M | 82.20 |
| com | 594K | 9.93 |
| org | 79K | 1.32 |
| net | 55K | 0.92 |
| nu | 40K | 0.67 |
| eu | 38K | 0.64 |
| de | 25K | 0.43 |
| no | 24K | 0.40 |
| info | 20K | 0.34 |
| se | 17K | 0.28 |

## Documents size (in segments)

**<= 25** segments **9.02%** (539K documents)
**> 25** segments **90.98%** (5.4M documents)

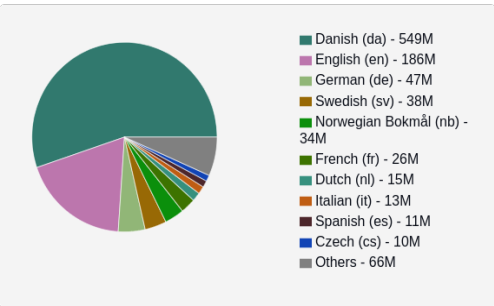



## Documents by collection

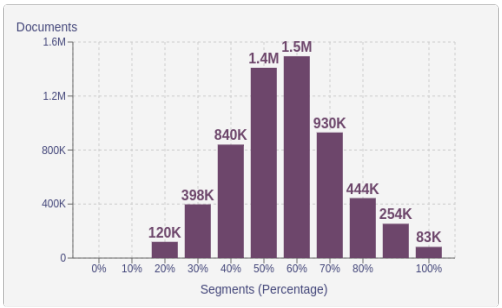



## Language Distribution

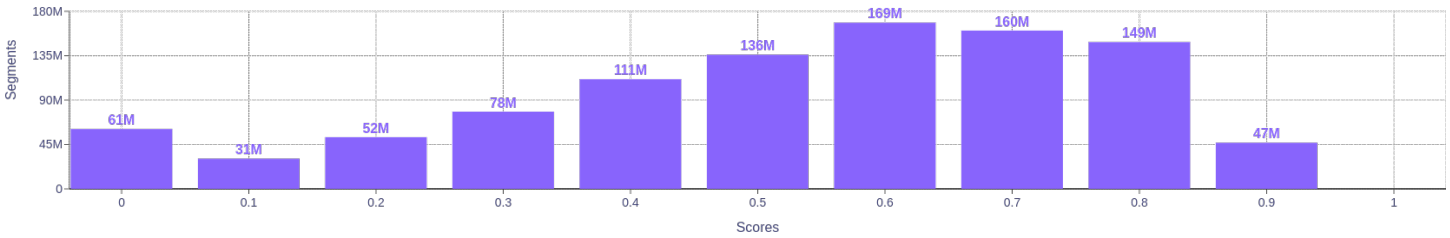
### Number of segments



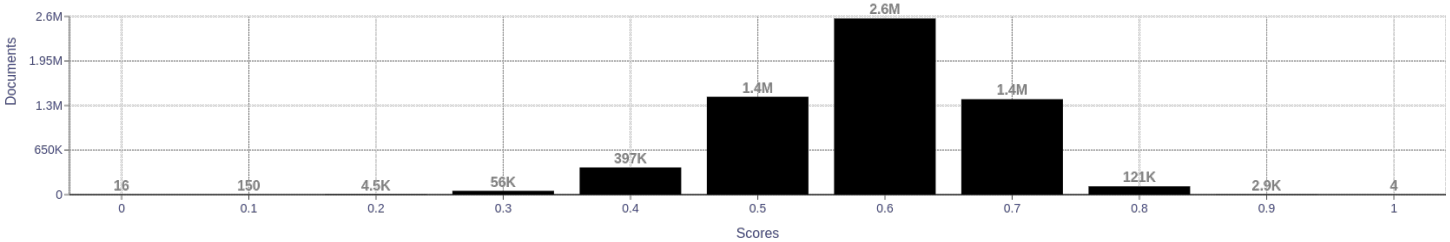

### Percentage of segments in Danish (da) inside documents



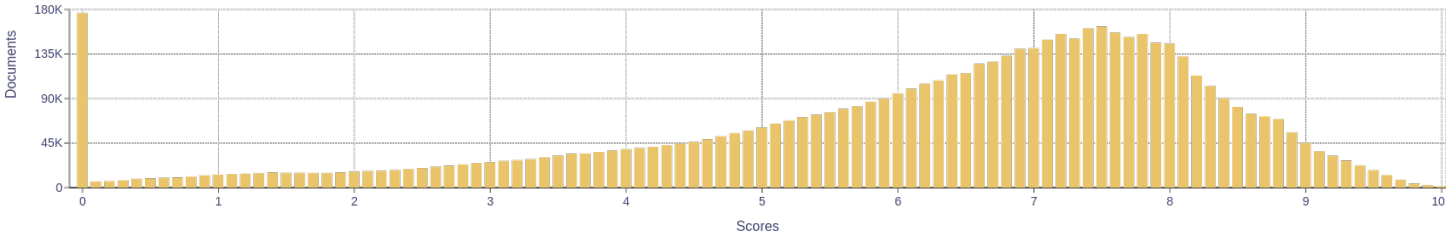

## Distribution of segments by fluency score




## Distribution of documents by average fluency score



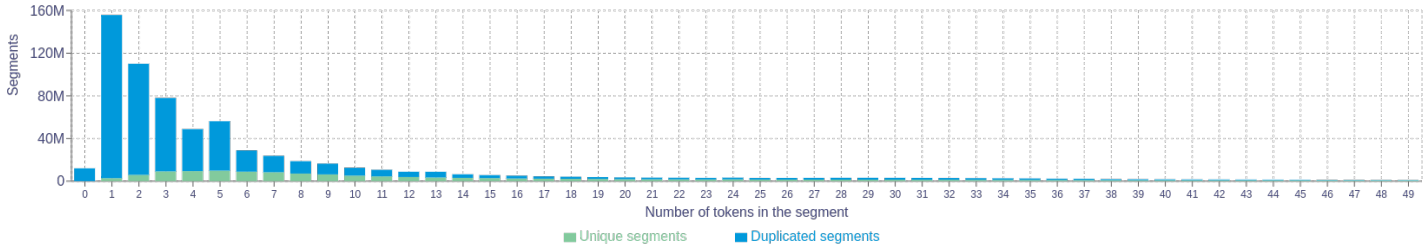

## Distribution of documents by document score



### Segment length distribution by token

**<= 49** tokens = **139M** segments | **554M** duplicates

**> 50** tokens = **34M** segments | **8.4M** duplicates




## 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