

## General overview

| Corpus               | Analytics date | Language     |
|----------------------|----------------|--------------|
| HPLT-docslite.lv.tsv | 6/8/2024       | Latvian (lv) |

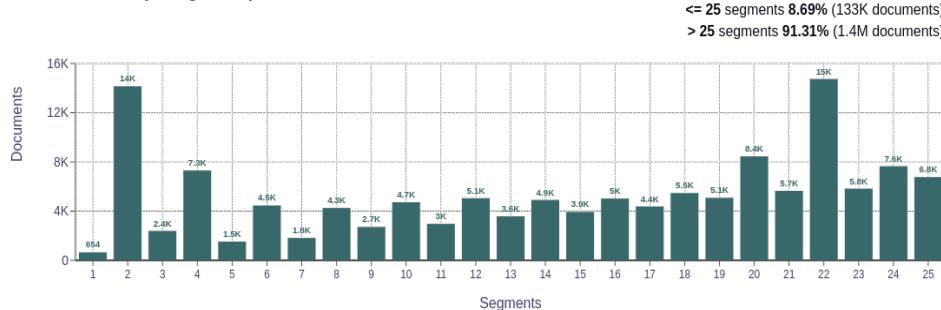
## Volumes

| Docs      | Segments    | Unique segments | Tokens | Size     |
|-----------|-------------|-----------------|--------|----------|
| 1,537,254 | 185,494,115 | 60,689 (0.03 %) | 2B     | 11.55 GB |

## Type-Token Ratio

|              |
|--------------|
| Latvian (lv) |
|--------------|

## Documents size (in segments)



## Top 10 domains

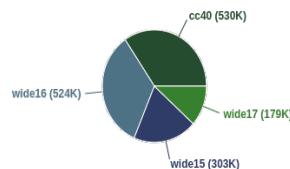
## Top 10 domains

| Domain            | Docs | % of total |
|-------------------|------|------------|
| diebuchsueche.com | 112K | 7.29       |
| viss.lv           | 107K | 6.96       |
| delfi.lv          | 40K  | 2.58       |
| europages.lv      | 31K  | 2.02       |
| tvnet.lv          | 17K  | 1.09       |
| lsm.lv            | 17K  | 1.08       |
| agoda.com         | 12K  | 0.78       |
| maminuklubs.lv    | 12K  | 0.75       |
| wikipedia.org     | 11K  | 0.72       |
| la.lv             | 11K  | 0.71       |

## Top 10 TLDs

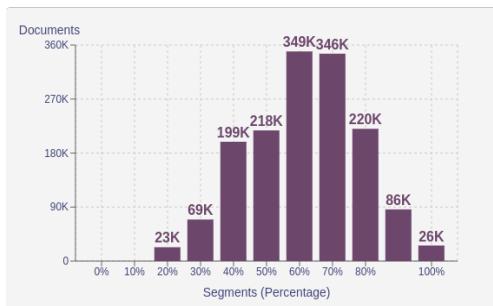
| Domain | Docs | % of total |
|--------|------|------------|
| lv     | 1M   | 67.83      |
| com    | 315K | 20.48      |
| eu     | 31K  | 2.03       |
| org    | 26K  | 1.67       |
| net    | 17K  | 1.13       |
| gov.lv | 16K  | 1.03       |
| lt     | 14K  | 0.90       |
| info   | 12K  | 0.79       |
| io     | 6K   | 0.39       |
| co.uk  | 4.8K | 0.31       |

## Documents by collection

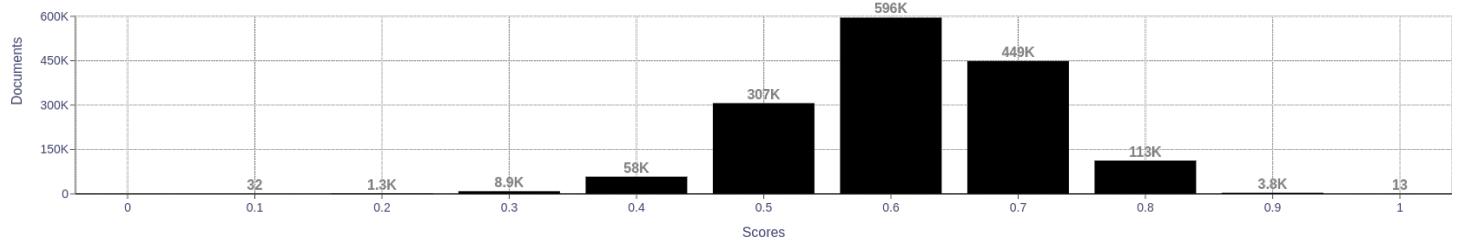


## Language Distribution

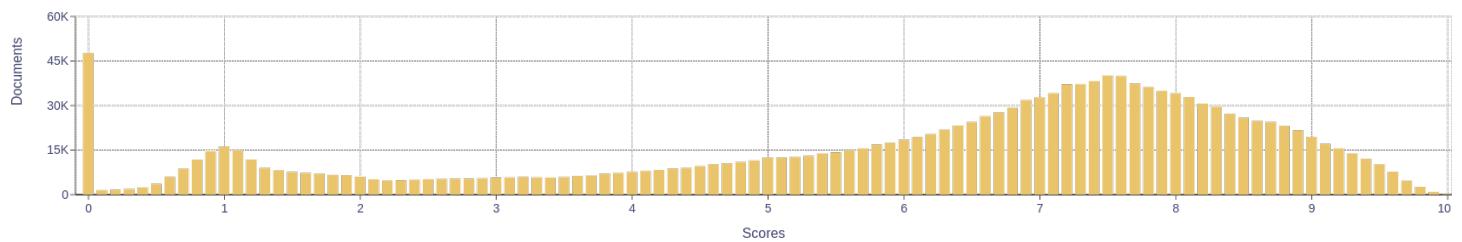
## Percentage of segments in Latvian (lv) inside documents



## Distribution of documents by average fluency score



## Distribution of documents by document score



## Segment length distribution by token



## 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>