# HPLT Analytics report

## General overview

| Corpus | Analytics date | Language |
|---|---|---|
| HPLT-docslite.bg.tsv | 6/10/2024 | Bulgarian (bg) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Size | Characters |
|---|---|---|---|---|---|
| 3,583,094 | 538,385,172 | 111,449,765 (20.70 %) | 5.8B | 50.14 GB | |

## Top 10 domains

| Domain | Docs | % of total |
|---|---|---|
| monitor.bg | 39K | 1.09 |
| blogspot.com | 38K | 1.07 |
| dtv-bg.com | 32K | 0.89 |
| pomagalo.com | 27K | 0.76 |
| chitanka.info | 26K | 0.74 |
| utre.bg | 26K | 0.73 |
| wikipedia.org | 23K | 0.65 |
| petardanov.com | 22K | 0.61 |
| europages.eu | 20K | 0.56 |
| dnevnik.bg | 20K | 0.55 |

## Top 10 TLDs

| Domain | Docs | % of total |
|---|---|---|
| bg | 1.5M | 41.35 |
| com | 1.3M | 37.05 |
| net | 220K | 6.14 |
| eu | 146K | 4.09 |
| org | 141K | 3.92 |
| info | 119K | 3.31 |
| news | 20K | 0.57 |
| pt | 13K | 0.37 |
| top | 8.7K | 0.24 |
| tv | 7K | 0.20 |

## Documents size (in segments)

<= 25 segments **4.15%** (148K documents)
> 25 segments **95.85%** (3.4M documents)
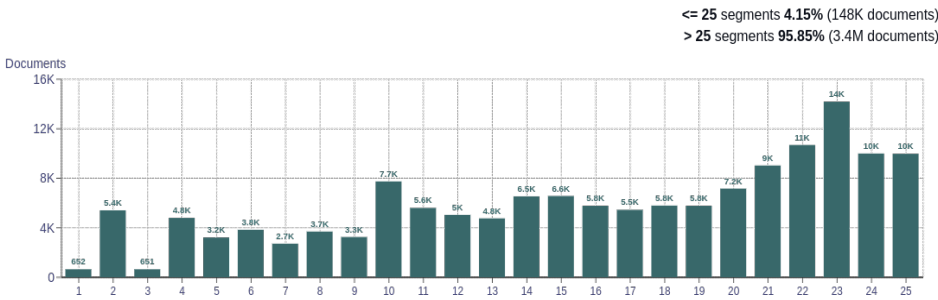


## Documents by collection



cc40 (2.2M)
wide16 (1.3M)
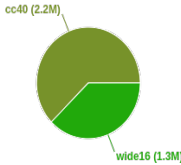
## Language Distribution

### Number of segments



- Bulgarian (bg) - 542M
- Russian (ru) - 134M
- English (en) - 95M
- Ukrainian (uk) - 21M
- Macedonian (mk) - 20M
- hbs-cyr - 19M
- French (fr) - 17M
- German (de) - 15M
- Spanish (es) - 6.9M
- Italian (it) - 4.2M
- 164 Others - 38M
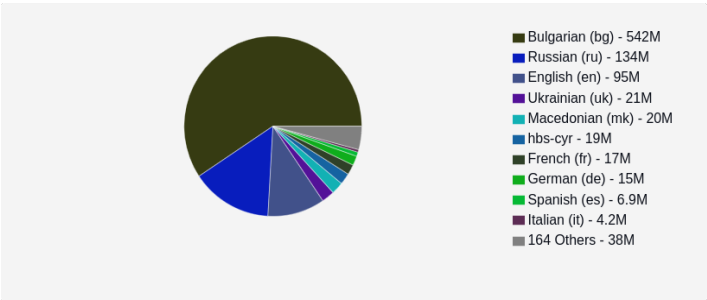
### Percentage of segments in Bulgarian (bg) inside documents
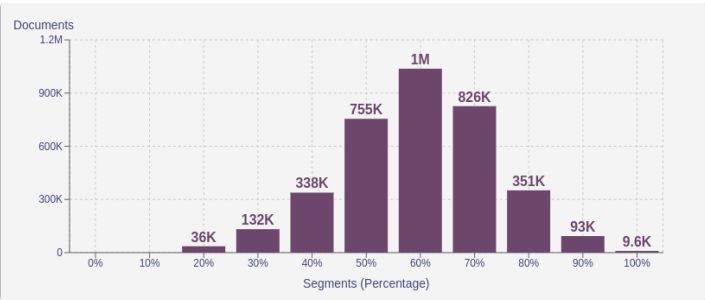


## Distribution of documents by document score

score <= 5 - **18.67%** (668K documents)
score > 5 - **81.33%** (2.9M documents)



## Segment length distribution by token

<= 49 tokens = **93M** segments | **421M** duplicates
> 50 tokens = **24M** segments | **6.2M** duplicates



- Unique segments
- Duplicated segments
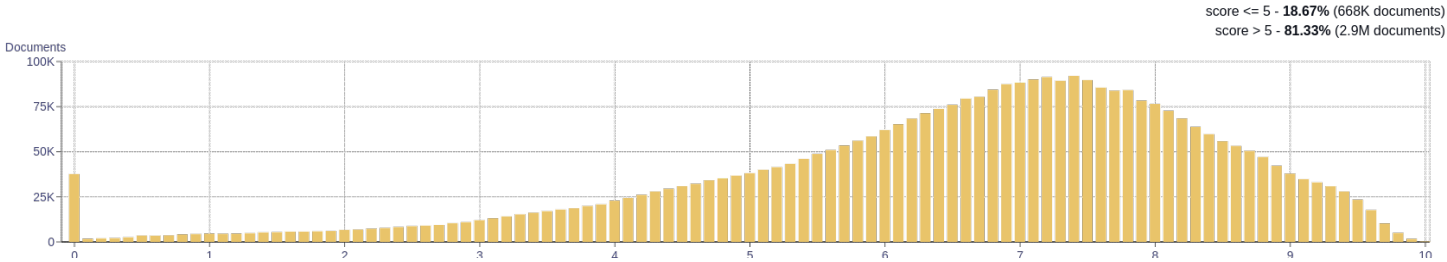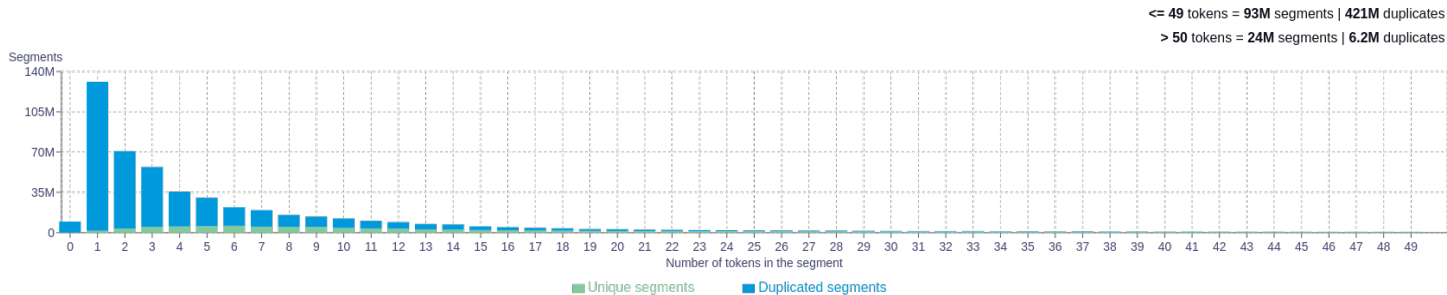
## About HPLT Analytics

**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