# HPLT Analytics report

HPLTAnalytics

## General overview

| Corpus | Analytics date | Language |
|---|---|---|
| is_1.jsonl.tsv | 3/22/2024 | Icelandic (is) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Size |
|---|---|---|---|---|
| 481,328 | 62,190,599 | 40,366 (0.06 %) | 662M | 3.66 GB |

## Type-Token Ratio

| Icelandic (is) |
|---|
| 0.01 |

## Top 10 domains

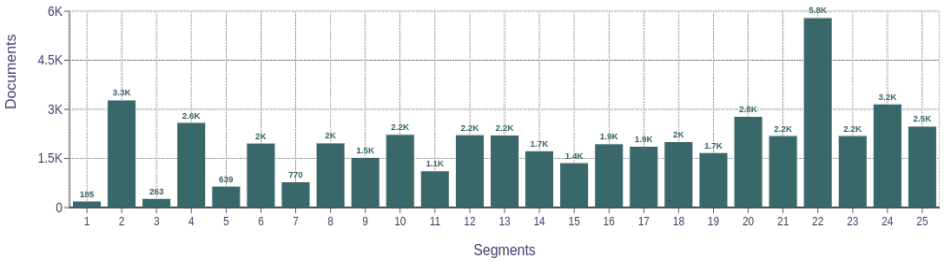
| Domain | Docs | % of total |
|---|---|---|
| blogspot.com | 12K | 2.47 |
| booking.com | 9.5K | 1.97 |
| wikipedia.org | 7.3K | 1.51 |
| kvennabladid.is | 6.4K | 1.34 |
| blogspot.is | 6.3K | 1.30 |
| blog.is | 5.4K | 1.13 |
| hotels.com | 5.4K | 1.11 |
| visindavefur.is | 4.4K | 0.91 |
| mbl.is | 4.3K | 0.89 |
| ruv.is | 4K | 0.83 |

## Top 10 TLDs

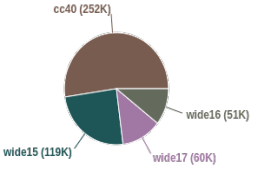
| Domain | Docs | % of total |
|---|---|---|
| is | 350K | 72.71 |
| com | 77K | 16.05 |
| org | 18K | 3.76 |
| net | 12K | 2.41 |
| nl | 1.8K | 0.37 |
| cc | 1.6K | 0.34 |
| fr | 1.6K | 0.34 |
| eu | 1.6K | 0.33 |
| fi | 1.2K | 0.24 |
| se | 975 | 0.20 |

## Documents size (in segments)

**<= 25** segments **10.38%** (50K documents)
**> 25** segments **89.62%** (431K documents)

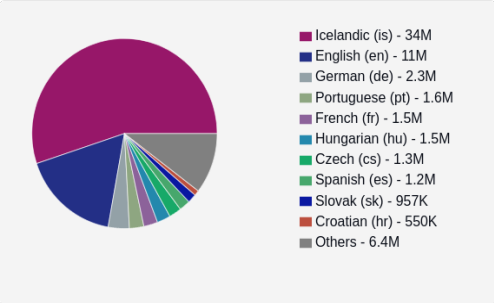



## Documents by collection

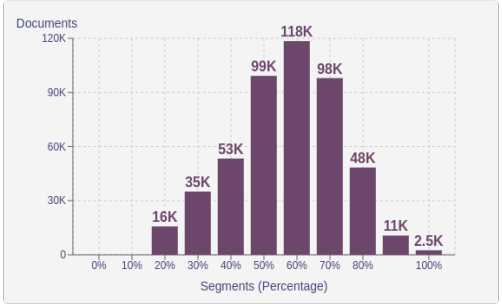



## Language Distribution

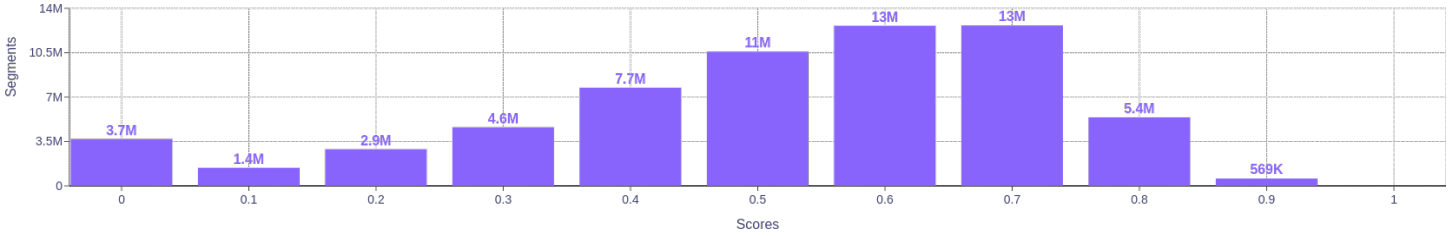
### Number of segments



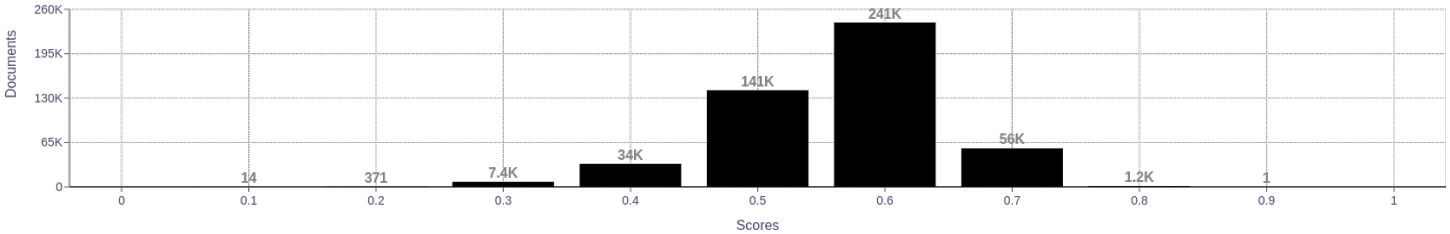

### Percentage of segments in Icelandic (is) inside documents



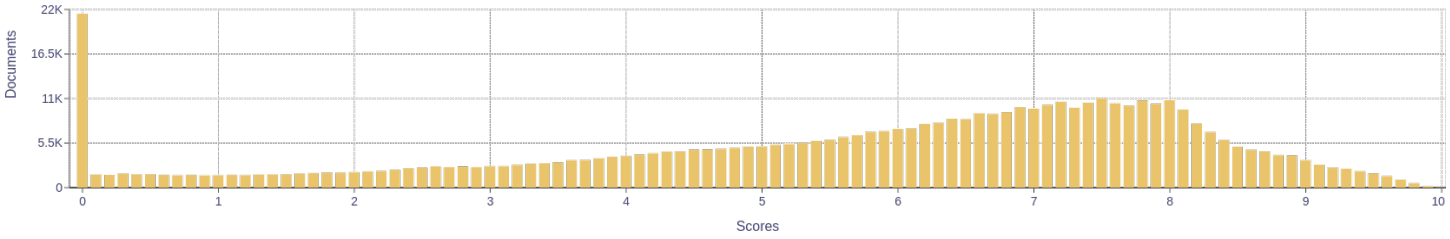

## Distribution of segments by fluency score




## Distribution of documents by average fluency score



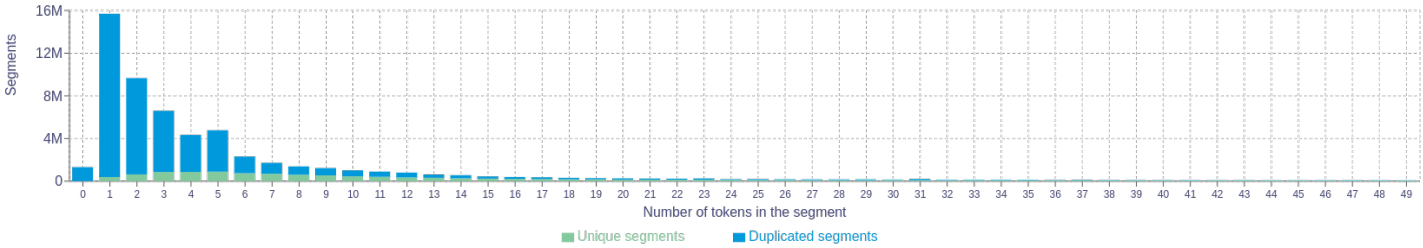

## Distribution of documents by document score



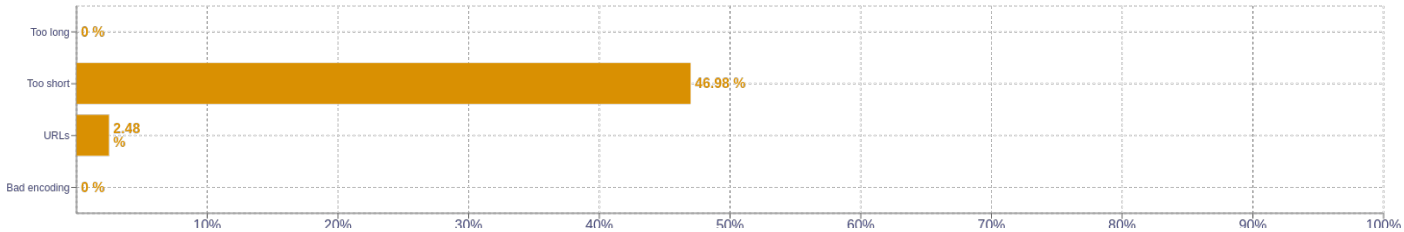
### Segment length distribution by token

**<= 49** tokens = **12M** segments | **47M** duplicates

**> 50** tokens = **3M** segments | **708K** duplicates




### Segment noise distribution




### Frequent n-grams

| Size | n-grams |
|---|---|
| 1 | var \| 2408989 · the \| 1902422 · a \| 1596154 · hafa \| 1271463 · and \| 1097022 |
| 2 | hafa samband \| 193006 · gististaðnum mynd \| 172266 · in the \| 167965 · lesa meira \| 160579 · eyða breyta \| 112195 |
| 3 | mynd af gististaðnum \| 185686 · sýna meira sýna \| 79671 · fær góða einkunn \| 78014 · meðalverð á nótt \| 67449 · hér á landi \| 62376 |
| 4 | mynd af gististaðnum mynd \| 172266 · gististaðnum mynd af gististaðnum \| 172266 · cookie is set by \| 34405 · opnast í nýjum glugga \| 33409 · the cookies in the \| 29360 |
| 5 | gististaðnum mynd af gististaðnum mynd \| 167795 · twitterdeila á facebookdeila á pinterest \| 41666 · deila á twitterdeila á facebookdeila \| 41666 · user consent for the cookies \| 29359 · the user consent for the \| 29359 |

## 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