

General overview

| Corpus | Analytics date | Language |
|---------------------|----------------|-------------|
| HPLT-docsite.he.tsv | 6/9/2024 | Hebrew (he) |

Volumes

| Docs | Segments | Unique segments | Tokens | Size |
|-----------|-------------|------------------|--------|----------|
| 4,979,580 | 838,116,375 | 114,769 (0.01 %) | 9B | 65.29 GB |

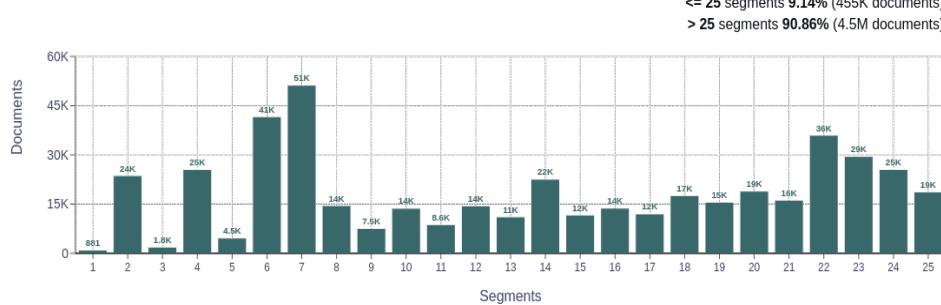
Top 10 domains

| Domain | Docs | % of total |
|-------------------|------|------------|
| alibaba.com | 543K | 10.91 |
| blogspot.co.il | 237K | 4.75 |
| eip.co.il | 170K | 3.42 |
| aliexpress.com | 150K | 3.01 |
| diebuchsueche.com | 66K | 1.32 |
| wikipedia.org | 65K | 1.30 |
| informed.co.il | 50K | 1.01 |
| blogspot.com | 45K | 0.90 |
| lightinthebox.com | 44K | 0.89 |
| saloona.co.il | 31K | 0.63 |

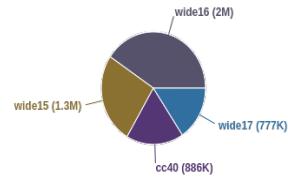
Top 10 TLDs

| Domain | Docs | % of total |
|--------|------|------------|
| co.il | 2M | 40.58 |
| com | 2M | 39.44 |
| org.il | 289K | 5.81 |
| org | 251K | 5.04 |
| net | 143K | 2.87 |
| ac.il | 42K | 0.84 |
| info | 31K | 0.63 |
| xyz | 14K | 0.27 |
| biz | 11K | 0.23 |
| co | 11K | 0.22 |

Documents size (in segments)

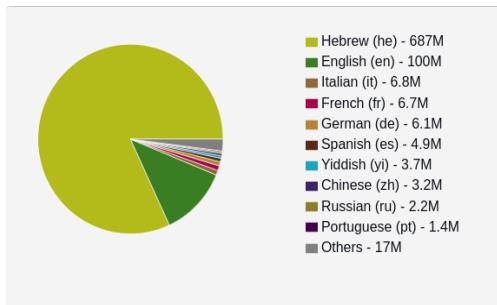


Documents by collection

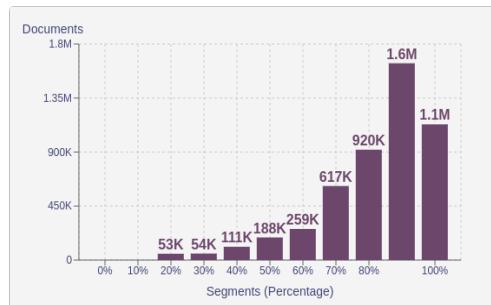


Language Distribution

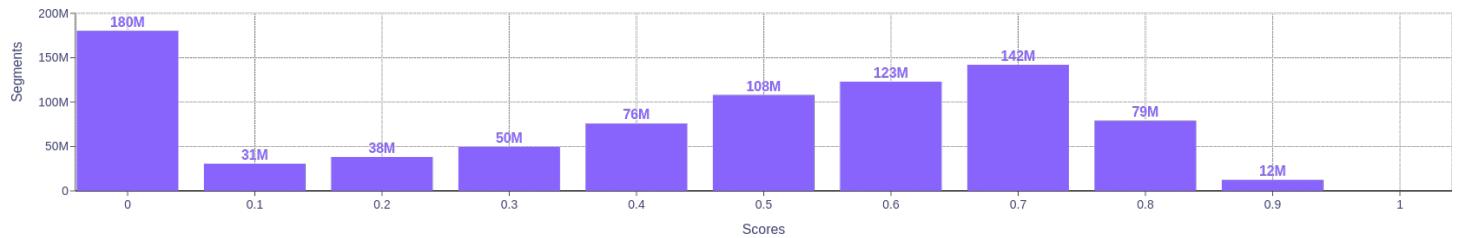
Number of segments



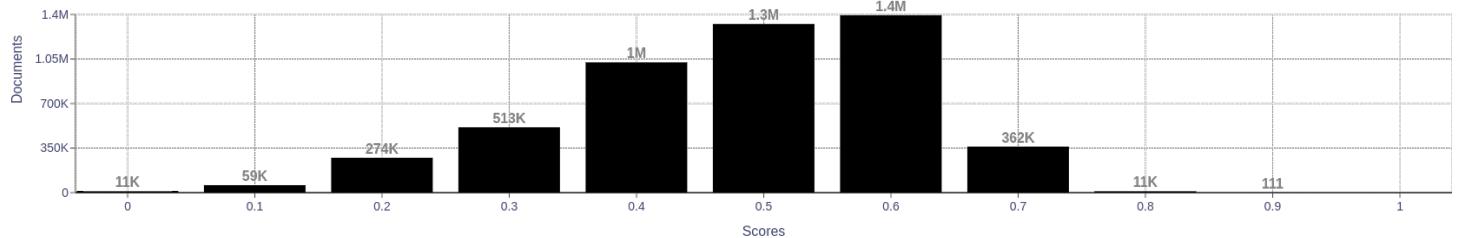
Percentage of segments in Hebrew (he) inside documents



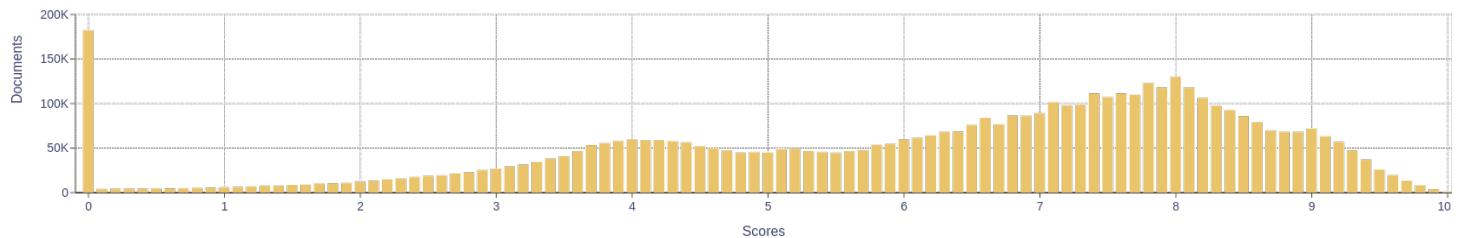
Distribution of segments by fluency score



Distribution of documents by average fluency score



Distribution of documents by document score



Segment length distribution by token

<= 49 tokens = 150M segments | 661M duplicates

> 50 tokens = 28M segments | 8.5M duplicates



About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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>, , , 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>