# HPLT Analytics report

HPLTAnalytics

## General overview

| Corpus | Analytics date | Language |
|---|---|---|
| ydd_Hebr.jsonl.tsv | 12/5/2024 | Yiddish (ydd) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Size | Characters |
|---|---|---|---|---|---|
| 128,265 | 2,940,163 | 1,420,745 (48.32 %) | 89M | 774.19 MB | 455,684,296 |

## Top 10 domains

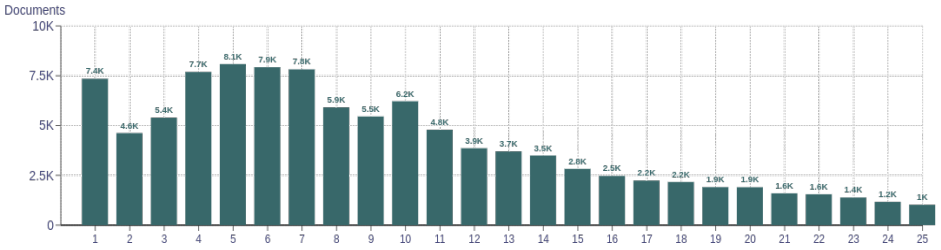
| Domain | Docs | % of total |
|---|---|---|
| wikipedia.org | 45K | 34.77 |
| kaveshtiebel.com | 15K | 11.49 |
| ivelt.com | 10K | 8.06 |
| yiddish.news | 4.1K | 3.22 |
| soft-free-download.com | 3.7K | 2.87 |
| sgames.org | 2.5K | 1.93 |
| freeplayonlinegame.net | 1.8K | 1.41 |
| actualidadgadget.com | 1.4K | 1.12 |
| itsmygame.org | 1.2K | 0.95 |
| creativosonline.org | 1.2K | 0.92 |

## Top 10 TLDs

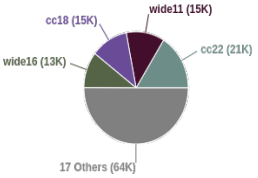
| Domain | Docs | % of total |
|---|---|---|
| com | 58K | 45.11 |
| org | 54K | 42.22 |
| news | 4.3K | 3.34 |
| net | 3.7K | 2.91 |
| ru | 1.1K | 0.84 |
| co.il | 775 | 0.60 |
| zone | 667 | 0.52 |
| gov | 609 | 0.47 |
| de | 549 | 0.43 |
| co | 513 | 0.40 |

## Documents size (in segments)

**<= 25** segments **79.98%** (103K documents)
**> 25** segments **20.02%** (26K documents)

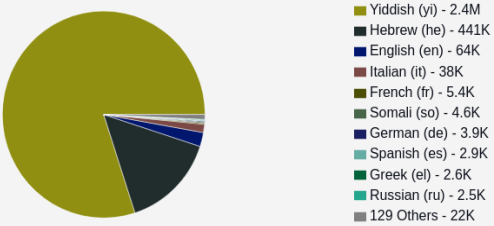



## Documents by collection



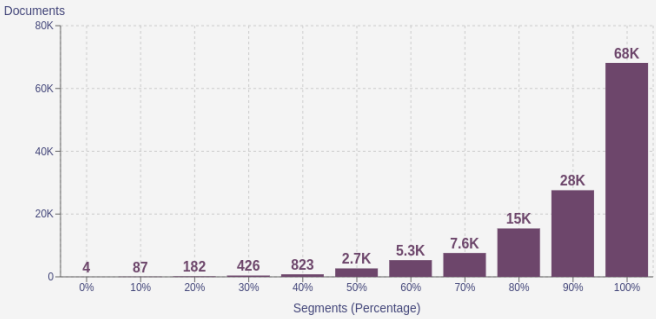

## Language Distribution

### Number of segments



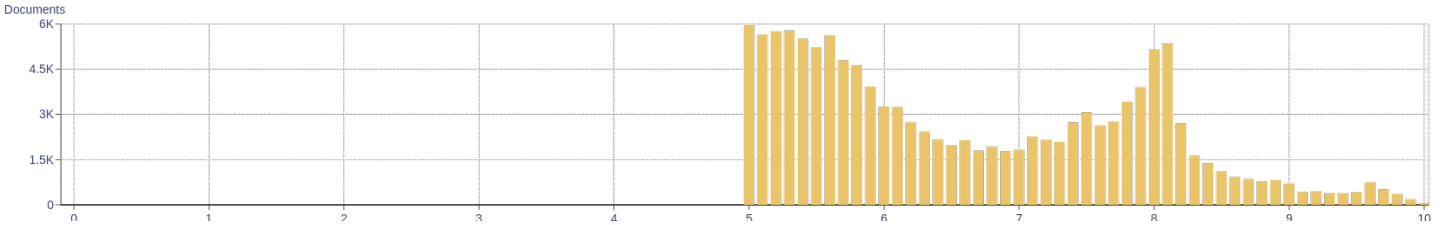

*Yiddish (ydd) identification might be inaccurate because language is not supported by Fasttext

### Percentage of segments in Yiddish (ydd) inside documents



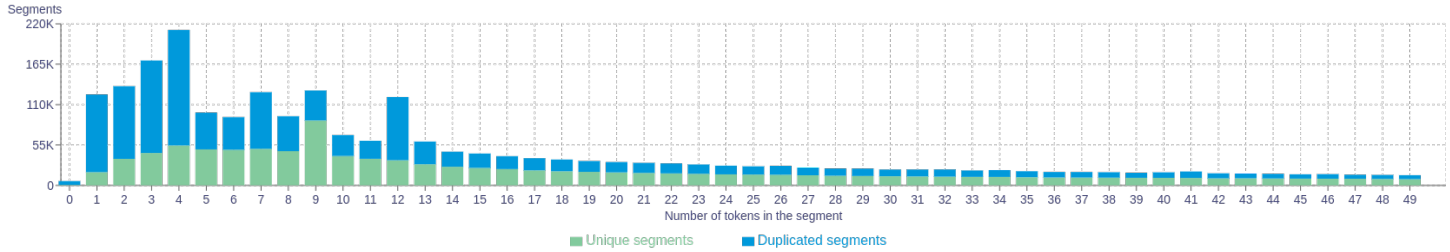

## Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (128K documents)

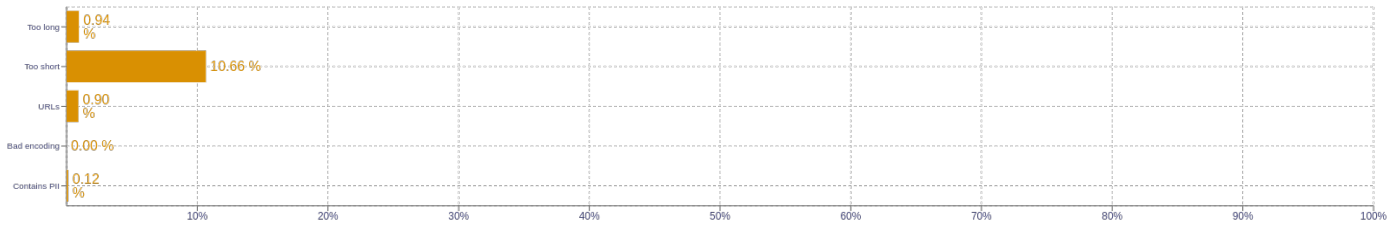



## Segment length distribution by token

**<= 49** tokens = **1.1M** segments | **1.3M** duplicates
**> 50** tokens = **579K** segments | **223K** duplicates




## Segment noise distribution



### Frequent n-grams

| Size | n-grams |
|---|---|
| 1 | 1847266 \| פון · 1600808 \| אין · 1355274 \| צו · 853755 \| האט · 680390 \| נישט |
| 2 | 206425 \| עס איז · 129219 \| איז געווען · 127409 \| האט געשריבן · 96478 \| ער האט · 72789 \| וואס איז |
| 3 | 63357 \| זיך איינגעשריבען אום · 15325 \| עס איז נישט · 15249 \| עס איז אַ · 15004 \| עס איז געווען · 14271 \| אז עס איז |
| 4 | 6270 \| באנוצערס וואס דרייען זיך · 6176 \| וואס לייענען דעם פארום · 5839 \| וואס דרייען זיך דא · 5468 \| איינער פון די מערסט · 5110 \| נישטא קיין אנליין באניצער |
| 5 | 6176 \| באניצער וואס לייענען דעם פארום · 5814 \| באנוצערס וואס דרייען זיך דא · 2529 \| טן יאנואר ליטן יוליאנישן קאלענדאר · 2527 \| יאנואר ליטן יוליאנישן קאלענדאר מיט · 2527 \| דאטעס דא זענען ליטן גרעגאריאנישן |

## 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