# HPLT Analytics report

HPLTAnalytics

## General overview

| Corpus | Analytics date | Language |
|---|---|---|
| mn_1.jsonl.tsv | 3/26/2024 | Mongolian (mn) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Size |
|---|---|---|---|---|
| 594,905 | 80,448,202 | 42,312 (0.05 %) | 977M | 8.73 GB |

## Type-Token Ratio

| Mongolian (mn) |
|---|
| 0.01 |

## Top 10 domains

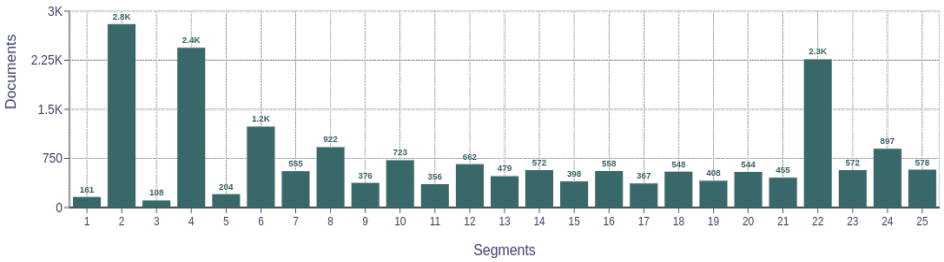
| Domain | Docs | % of total |
|---|---|---|
| blogspot.kr | 9.8K | 1.65 |
| fact.mn | 8.6K | 1.45 |
| miss.mn | 4.4K | 0.75 |
| wikipedia.org | 4.3K | 0.72 |
| leaders.mn | 4.1K | 0.69 |
| cekc.mn | 3.9K | 0.66 |
| emegteichuud.mn | 3.9K | 0.66 |
| blogspot.com | 3.7K | 0.62 |
| ikon.mn | 3.3K | 0.55 |
| goolingoo.mn | 3.1K | 0.52 |

## Top 10 TLDs

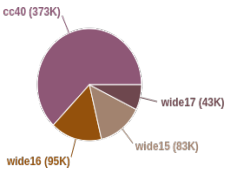
| Domain | Docs | % of total |
|---|---|---|
| mn | 204K | 34.25 |
| pl | 107K | 17.98 |
| nl | 55K | 9.31 |
| com | 38K | 6.47 |
| be | 32K | 5.38 |
| fr | 24K | 4.08 |
| de | 22K | 3.76 |
| es | 18K | 3.11 |
| gov.mn | 14K | 2.35 |
| org | 11K | 1.82 |

## Documents size (in segments)

**<= 25** segments **3.23%** (19K documents)
**> 25** segments **96.77%** (575K documents)

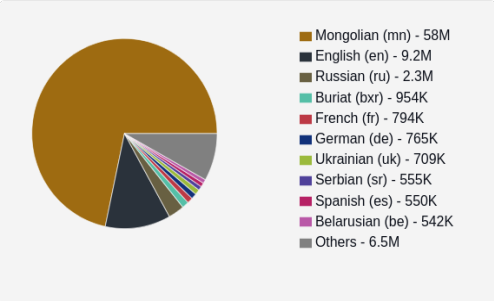



## Documents by collection

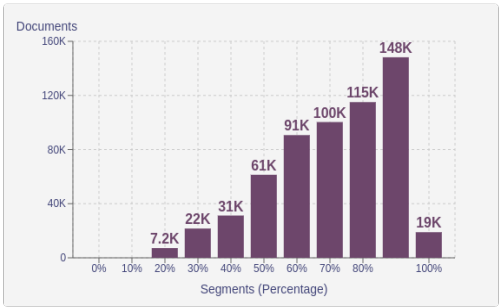



## Language Distribution

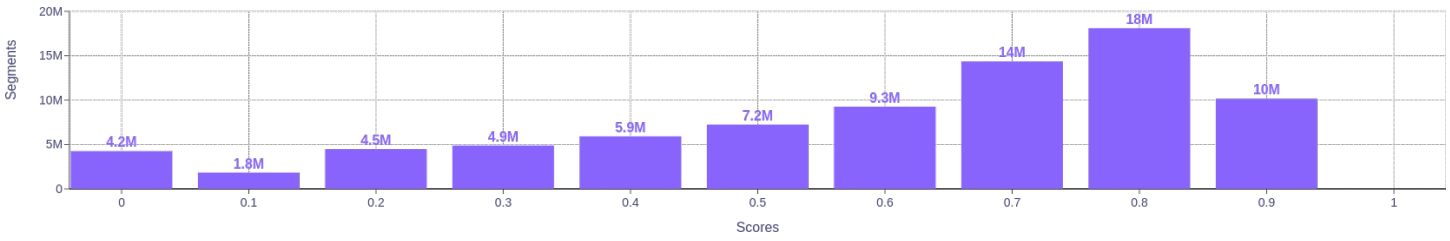
### Number of segments



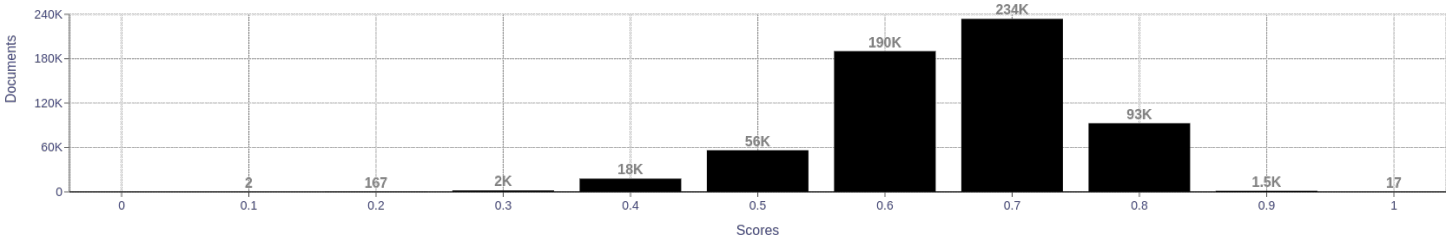

### Percentage of segments in Mongolian (mn) inside documents



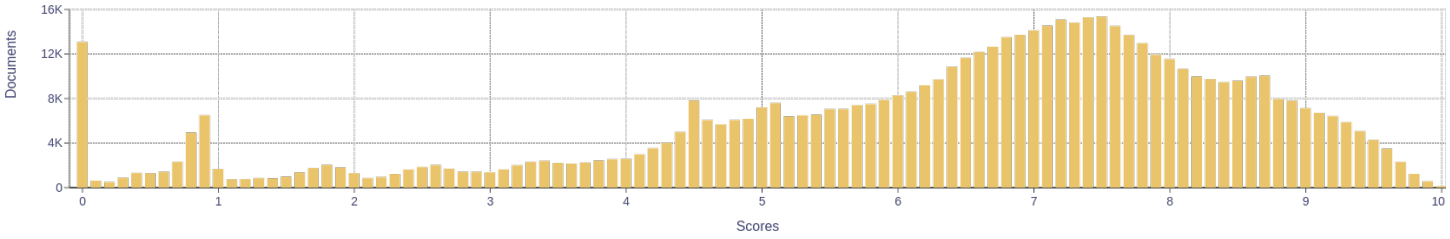

## Distribution of segments by fluency score




## Distribution of documents by average fluency score

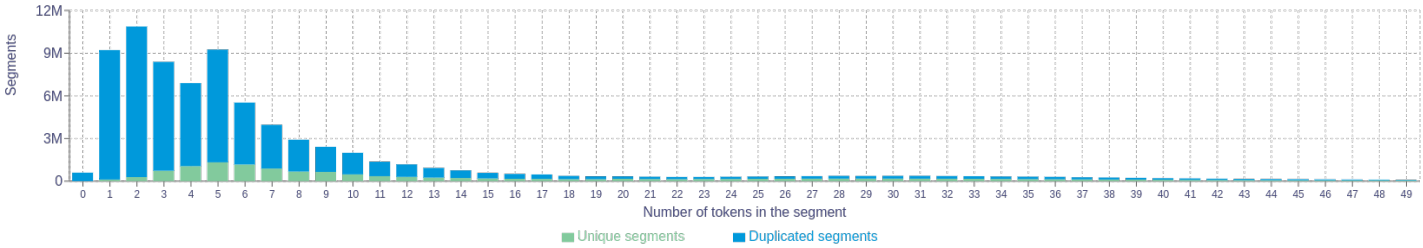



## Distribution of documents by document score

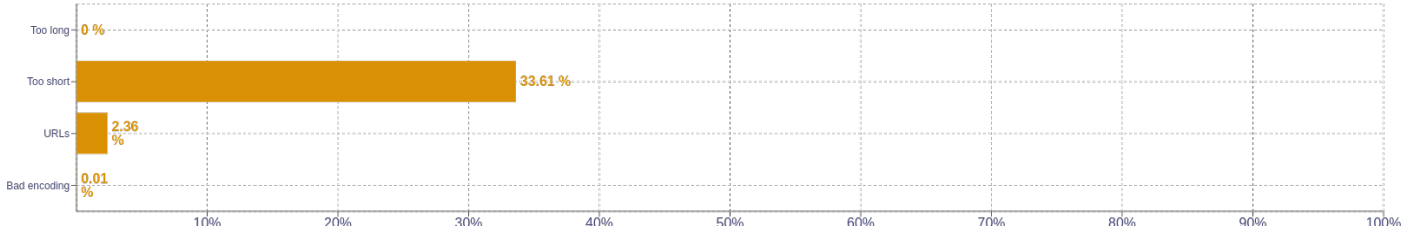


### Segment length distribution by token

**<= 49** tokens = **13M** segments | **63M** duplicates
**> 50** tokens = **3.7M** segments | **848K** duplicates




### Segment noise distribution




### Frequent n-grams

| Size | n-grams |
|---|---|
| 1 | бутлуур \| 12542558 · чулуу \| 6178063 · бутлуурын \| 4837091 · машин \| 4818551 · үнэ \| 4051676 |
| 2 | чулуу бутлуур \| 2249495 · тоног төхөөрөмж \| 2081635 · үнэ авах \| 1931351 · хацарт бутлуур \| 1782974 · уул уурхайн \| 1601339 |
| 3 | бидэнтэй холбоо барина \| 428512 · хоёр дахь гар \| 398470 · уул уурхайн тоног \| 381526 · уурхайн тоног төхөөрөмж \| 326418 · чулуу бутлах машин \| 286444 |
| 4 | уул уурхайн тоног төхөөрөмж \| 268677 · яг одоо бидэнтэй нэгдээрэй \| 218968 · худалдах хоёр дахь гар \| 166970 · 144398 \| كسارة الحجر كسارة الحجر · бутлуур нь шохойн чулуу \| 124304 |
| 5 | 109455 \| الحجر كسارة الحجر كسارة الحجر · 109413 \| كسارة الحجر كسارة الحجر كسارة · бутлах машин хийх элс машин \| 105919 · чулуу бутлах машин хийх элс \| 104051 · машин хийх элс машин чулуу \| 83600 |

## 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