# HPLT Analytics report

HPLTAnalytics

## General overview

| Corpus | Analytics date | Language |
|---|---|---|
| ckb_Arab.jsonl.tsv | 9/20/2024 | Central Kurdish (ckb) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Size | Characters |
|---|---|---|---|---|---|
| 273,745 | 5,225,884 | 2,963,957 (56.72 %) | 171M | 1.55 GB | 907,857,638 |

## Top 10 domains

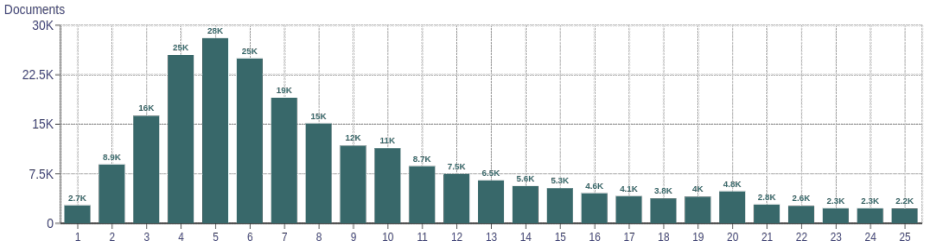
| Domain | Docs | % of total |
|---|---|---|
| dengiamerika.com | 38K | 13.82 |
| hawlati.co | 7.8K | 2.85 |
| blogfa.com | 6.2K | 2.28 |
| awene.com | 5.1K | 1.87 |
| penusakan.com | 4.6K | 1.70 |
| kurdistantv.net | 4.4K | 1.60 |
| dangiislam.org | 4.3K | 1.58 |
| blogspot.com | 3.9K | 1.42 |
| payam.tv | 3.8K | 1.41 |
| wishe.net | 3.7K | 1.35 |

## Top 10 TLDs

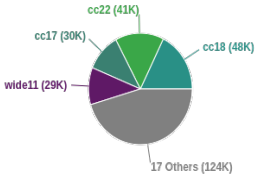
| Domain | Docs | % of total |
|---|---|---|
| com | 140K | 50.97 |
| net | 50K | 18.38 |
| org | 33K | 12.22 |
| co | 10K | 3.80 |
| info | 7.3K | 2.68 |
| krd | 6.8K | 2.47 |
| tv | 6.4K | 2.35 |
| ir | 3.7K | 1.33 |
| ca | 3.6K | 1.33 |
| se | 2.2K | 0.80 |

## Documents size (in segments)

**<= 25** segments **84.34%** (231K documents)
**> 25** segments **15.66%** (43K documents)

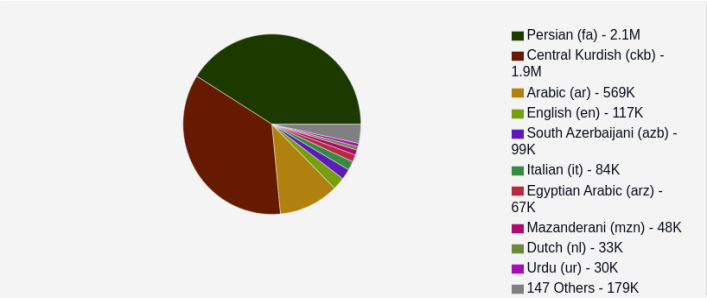



## Documents by collection

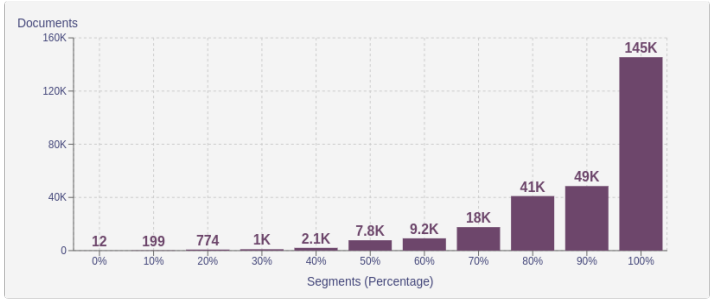



## Language Distribution

### Number of segments

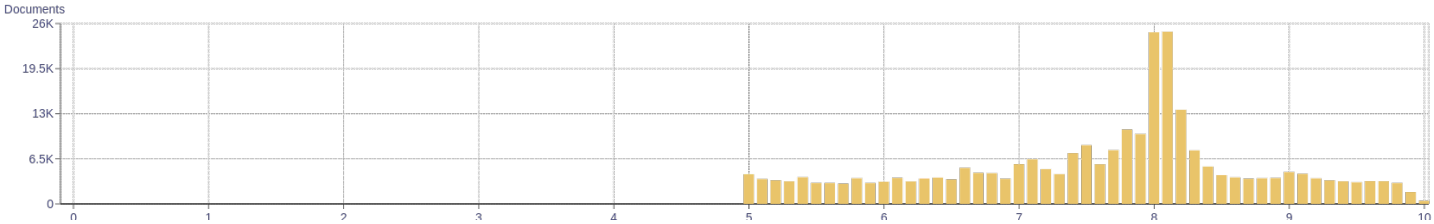



### Percentage of segments in Central Kurdish (ckb) inside documents



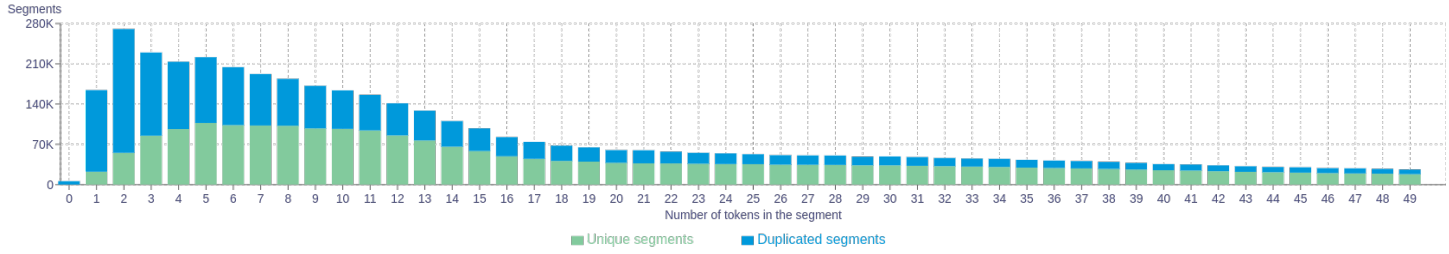

## Distribution of documents by document score

score <= 5 - **99.83%** (273K documents)
score > 5 - **0.17%** (457 documents)

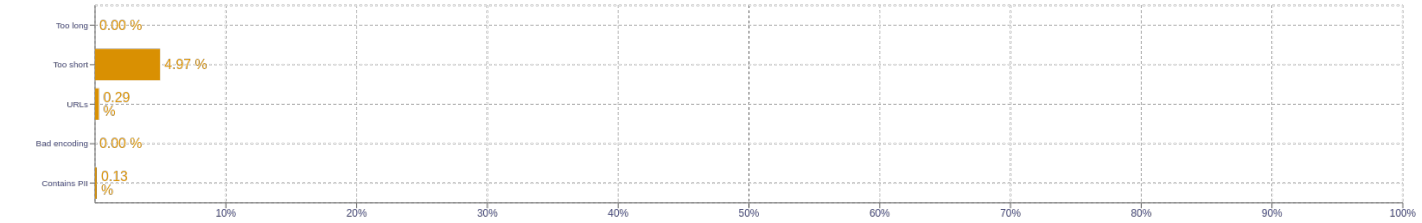



## Segment length distribution by token

**<= 49** tokens = **2.3M** segments | **1.9M** duplicates
**> 50** tokens = **994K** segments | **327K** duplicates




## Segment noise distribution



### Frequent n-grams

| Size | n-grams |
|---|---|
| 1 | 6089957 \| ی 4498158 \| لە 1204616 \| کە 1168762 \| ئەو 933437 \| کە |
| 2 | 181781 \| ان ی 169315 \| ێک ی 128083 \| ی ی 105117 \| ی ش 89001 \| کە لە |
| 3 | 43848 \| لە ساڵ ی 41560 \| وڵات ان ی 29221 \| خوا ی گەورە 26341 \| الله علیه وسلم 24269 \| صلی الله علیه |
| 4 | 22939 \| صلی الله علیه وسلم 18964 \| صلی الله علیه وسلم 7223 \| لە وڵات ان ی 5153 \| شا یان ی باسە 5148 \| لە ڕێگەی کلیـکـکردنی ئەو |
| 5 | 5113 \| لە ڕێگەی کلیـکـکردنی ئەو فایلە 3358 \| ڕێگەی کلیـکـکردنی ئەو فایلە دهنـگیـیانەی 3346 \| خوا ی لێ بێ ت 2494 \| خوا ی لەسەر بێ ت 2072 \| گفتوگۆکە بن لە ڕێگەی کلیـکـکردنی |

## 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