# HPLT Analytics report

HPLTAnalytics

## General overview

| Corpus | Date | Language |
|---|---|---|
| hplt-v3-kir_Cyrl | 9/18/2025 | Kyrgyz (ky) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Characters | Size |
|---|---|---|---|---|---|
| 1,489,065 | 20,264,647 | 15,365,796 (75.83 %) | 607M | 3,782,448,224 | 6.45 GB |

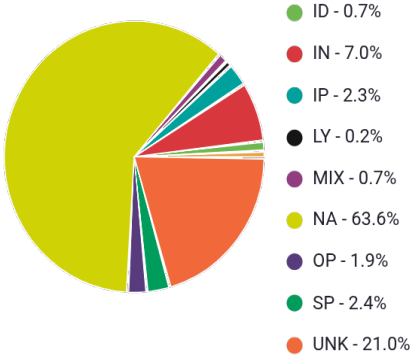
## Top 10 domains

| Domain | Docs | % of total |
|---|---|---|
| azattyk.org | 245K | 16.46% |
| sputnik.kg | 138K | 9.25% |
| kabar.kg | 59K | 3.95% |
| super.kg | 40K | 2.67% |
| kloop.asia | 31K | 2.09% |
| wikipedia.org | 26K | 1.74% |
| turmush.kg | 24K | 1.63% |
| kyrgyztoday.org | 23K | 1.52% |
| pk.kg | 21K | 1.41% |
| trt.net.tr | 17K | 1.14% |

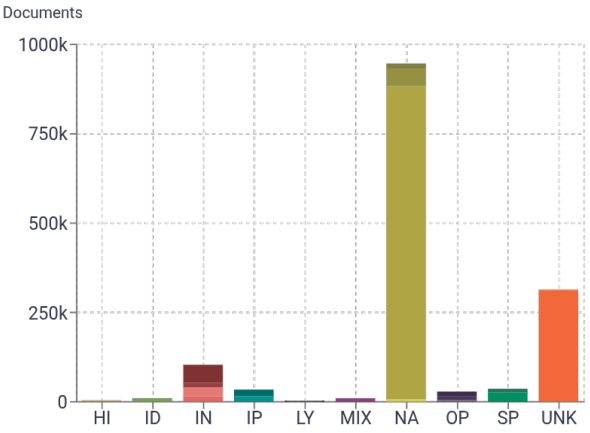
## Top 10 TLDs

| Domain | Docs | % of total |
|---|---|---|
| kg | 674K | 45.23% |
| org | 344K | 23.13% |
| com | 173K | 11.63% |
| ru | 98K | 6.60% |
| asia | 32K | 2.14% |
| gov.kg | 25K | 1.68% |
| net | 19K | 1.26% |
| media | 18K | 1.22% |
| net.tr | 17K | 1.14% |
| tv | 9.7K | 0.65% |

## Register labels



HI - 0.3%
ID - 0.7%
IN - 7.0%
IP - 2.3%
LY - 0.2%
MIX - 0.7%
NA - 63.6%
OP - 1.9%
SP - 2.4%
UNK - 21.0%

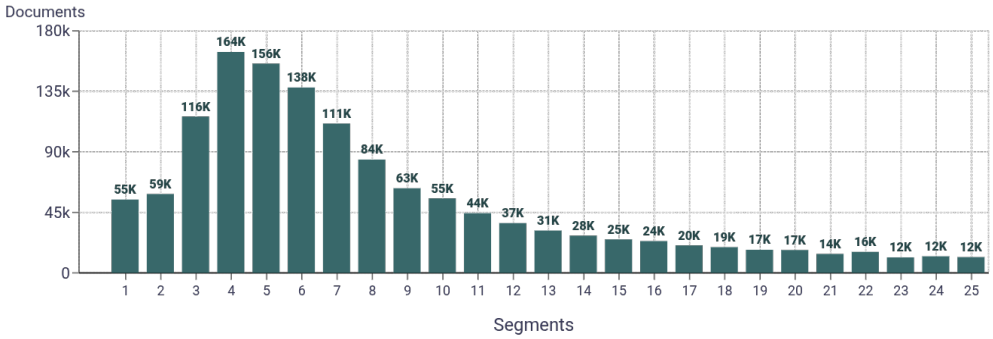



HI_other - 0.2%
HI_re - 0.1%
ID_other - 0.7%
IN_dtp - 1.0%
IN_en - 1.7%
IN_lt - 0.8%
IN_other - 3.4%
IN_ra - 0.0%
IP_ds - 1.1%
IP_ed - 0.0%
IP_other - 1.2%
LY_other - 0.2%
MIX - 0.7%
NA_nb - 0.5%
NA_ne - 58.8%
NA_other - 3.3%
NA_sr - 0.9%
OP_av - 0.2%
OP_ob - 0.1%
OP_other - 0.7%
OP_rs - 0.9%
OP_rv - 0.1%
SP_it - 1.8%
SP_other - 0.6%
UNK - 21.0%

**MT**:16.0% | 238K Documents

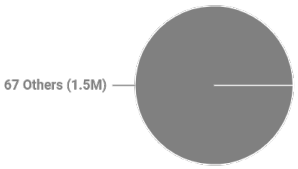
## Documents size (in segments)

**≤ 25** segments **89.29%** (1.3M documents)
**> 25** segments **10.71%** (160K documents)



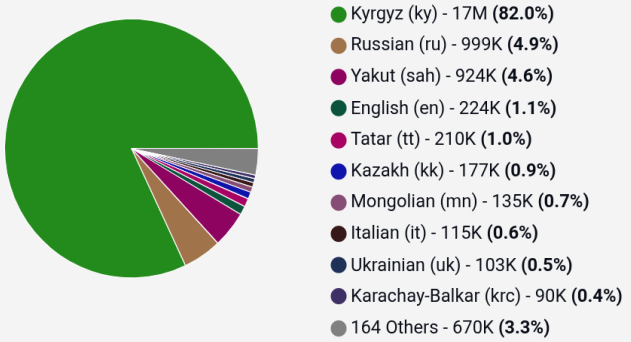

## Document collections

**CC = 96.29%**
**IA = 3.71%**

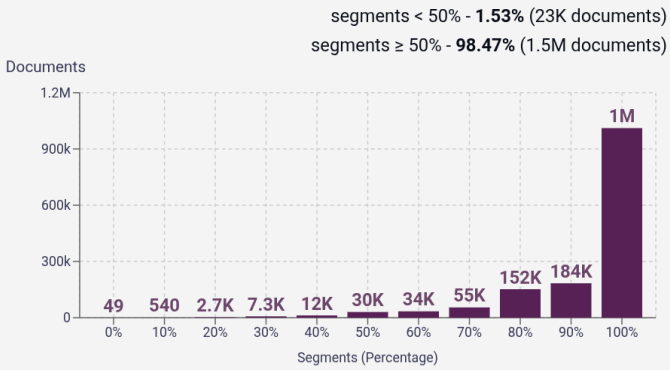


## Language Distribution

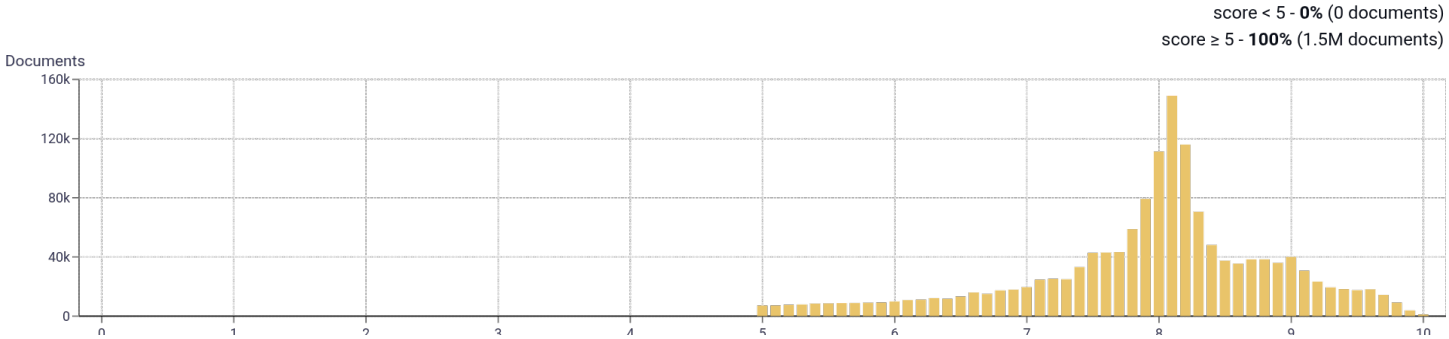
### Number of segments in the Kyrgyz (ky) corpus



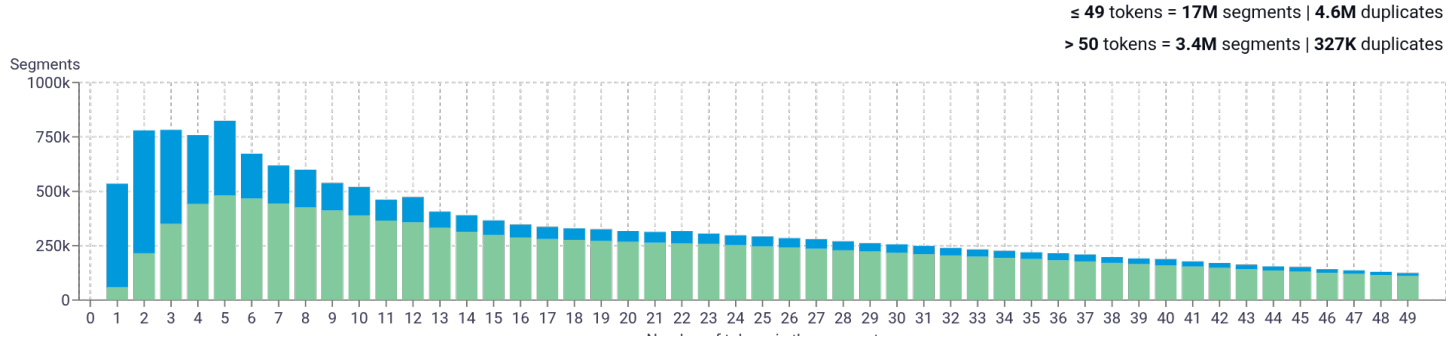

### Percentage of segments in Kyrgyz (ky) inside documents



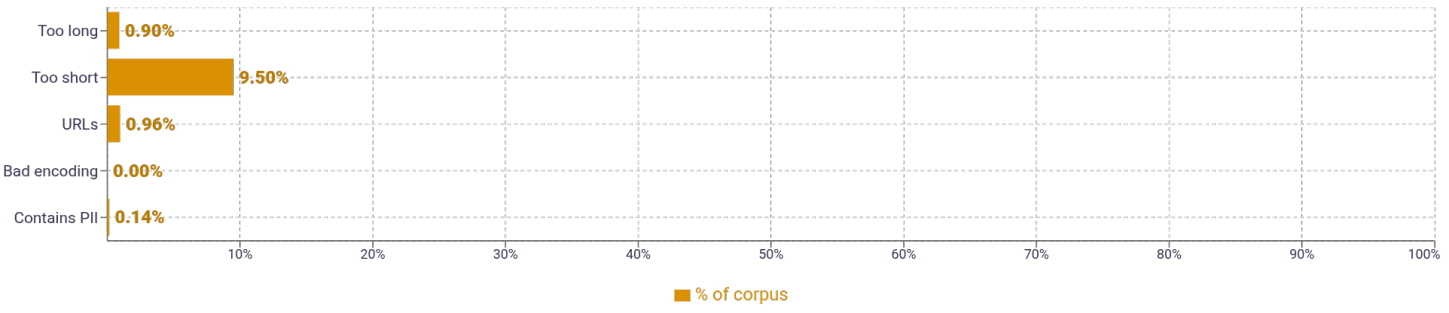

## Distribution of documents by document score




## Segment length distribution by token




## Segment noise distribution



### Frequent n-grams

| SIZE | N-GRAMS |
|---|---|
| 1 | иш \| 760,911 · жылы \| 691,441 · жакшы \| 615,575 · жылдын \| 603,654 · жөнүндө \| 594,368 |
| 2 | билим берүү \| 197,352 · орун басары \| 145,833 · мындан тышкары \| 110,938 · ички иштер \| 107,907 · жылдан бери \| 106,986 |
| 3 | жергиликтүү өз алдынча \| 33,115 · берүү жана илим \| 33,057 · президент садыр жапаров \| 30,405 · тышкы иштер министри \| 28,710 · министринин орун басары \| 27,536 |
| 4 | билим берүү жана илим \| 32,976 · жергиликтүү өз алдынча башкаруу \| 27,581 · дат баспас болоттон жасалган \| 12,827 · дене тарбия жана спорт \| 10,646 · министрлигинин басма сөз кызматы \| 10,601 |
| 5 | билим берүү жана илим министрлиги \| 9,353 · билим берүү жана илим министрлигинин \| 9,154 · жергиликтүү өз алдынча башкаруу органдарынын \| 8,479 · министрлер кабинетинин төрагасы акылбек жапаров \| 7,658 · который будет отправлен нашим редакторам \| 6,389 |

## 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

**Register labels**

| Name | Abbr. | Name | Abbr. | Name | Abbr. |
|---|---|---|---|---|---|
| **Machine-translated** | MT | **How-to or instructions** | **HI** | Description of a thing or person | dtp |
| **Lyrical** | LY | Recipe | re | FAQ | fi |
| **Spoken** | SP | **Informational persuasion** | **IP** | Legal terms & conditions | lt |
| Interview | it | Description with intent to sell | ds | **Opinion** | **OP** |
| **Interactive discussion** | ID | News & opinion blog or editorial | ed | Review | rv |
| **Narrative** | NA | **Informational description** | **IN** | Opinion blog | ob |
| News report | ne | Enciclopedia article | en | Denominational religious blog or sermon | rs |
| Sports report | sr | Research article | ra | Advice | av |
| Narrative blog | nb | | | | |