

General overview

| Corpus | Date | Language |
|------------------|-----------|----------|
| hplt-v3-kir_Cyrl | 9/18/2025 | Kirghiz |

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 |

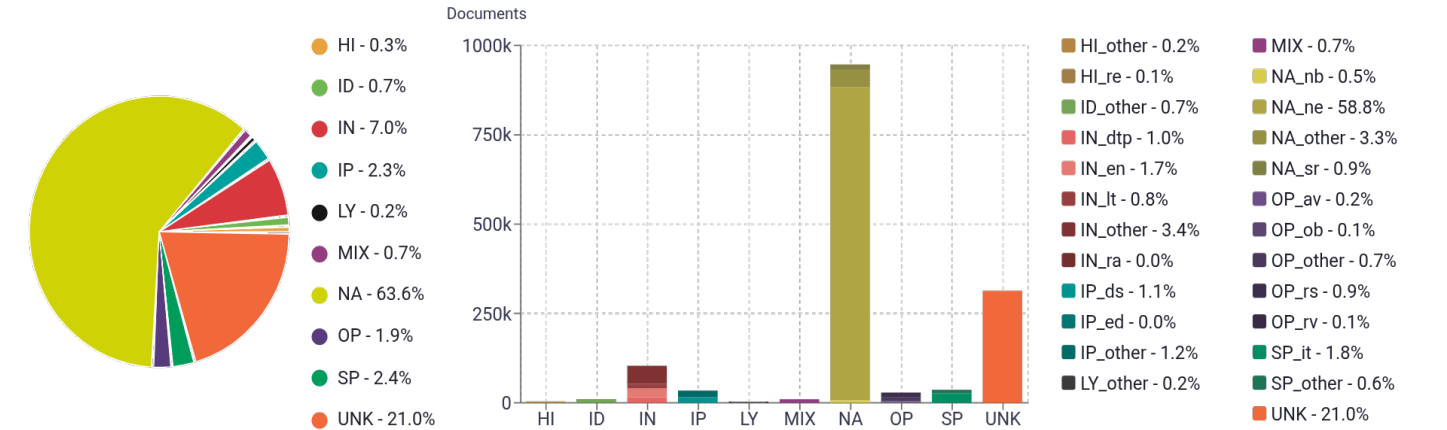
Top 10 domains

| 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

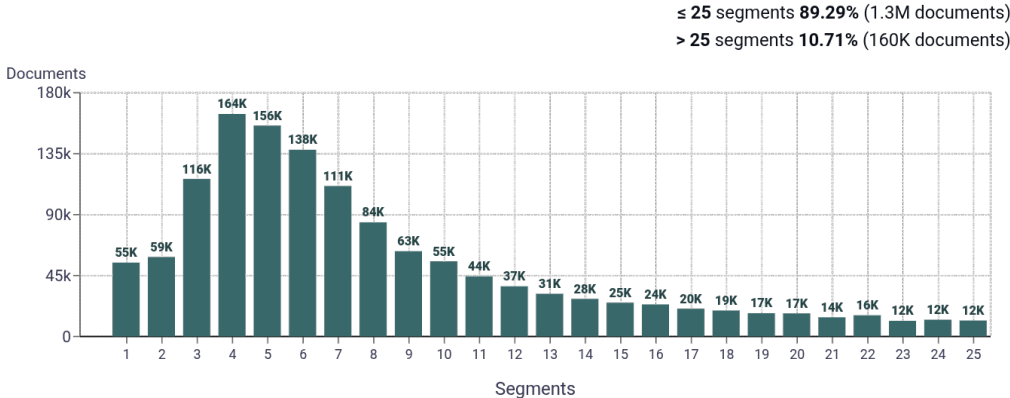
| Domain | Docs | % of total |
|--------|------|------------|
| kg | 674K | 45.23% |
| org | 344K | 23.13% |
| com | 173K | 11.63% |
| ru | 98K | 6.60% |
| asia | 32K | 2.14% |
| gov.kg | 25K | 1.68% |
| net | 19K | 1.26% |
| media | 18K | 1.22% |
| net.tr | 17K | 1.14% |
| tv | 9.7K | 0.65% |

Register labels

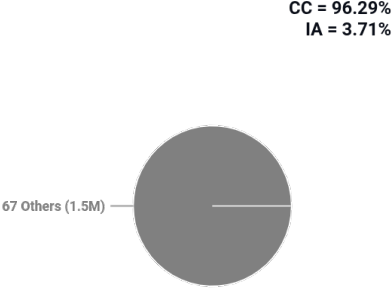


MT:16.0% | 238K Documents

Documents size (in segments) ⓘ

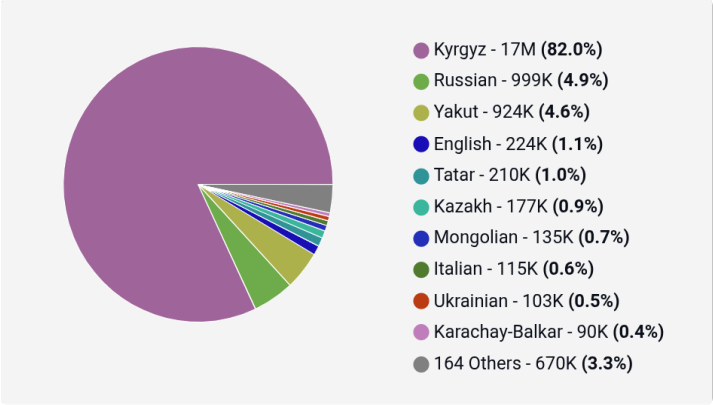


Document collections

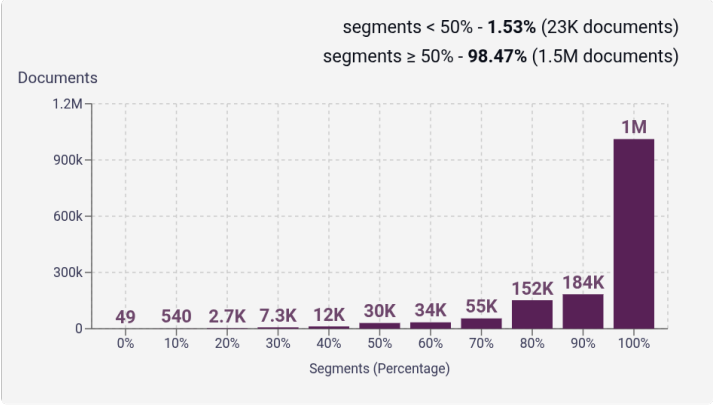


Language Distribution

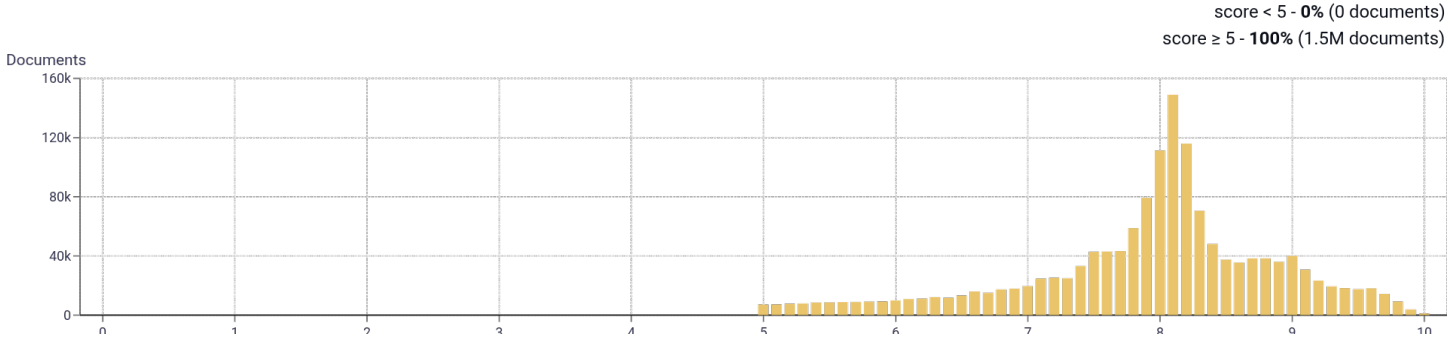
Number of segments in the Kirghiz corpus



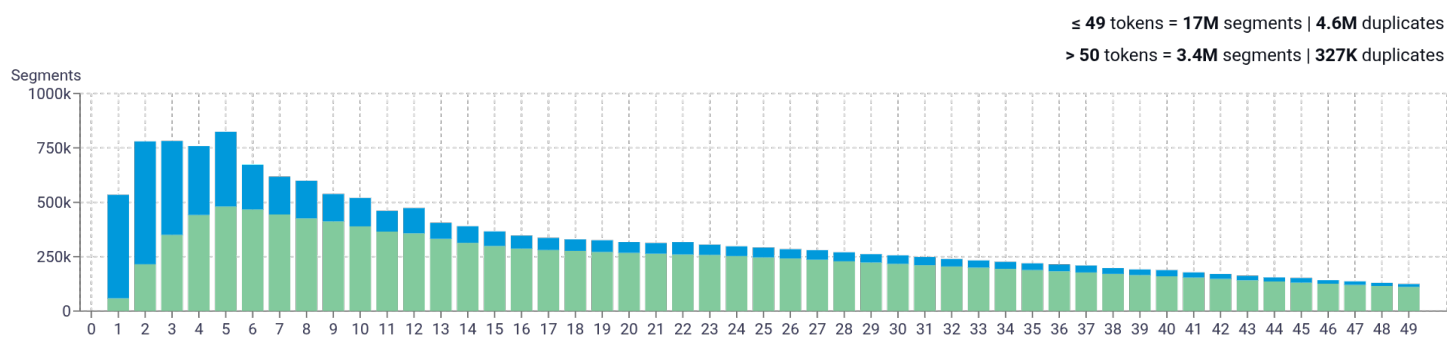
Percentage of segments in Kirghiz inside documents



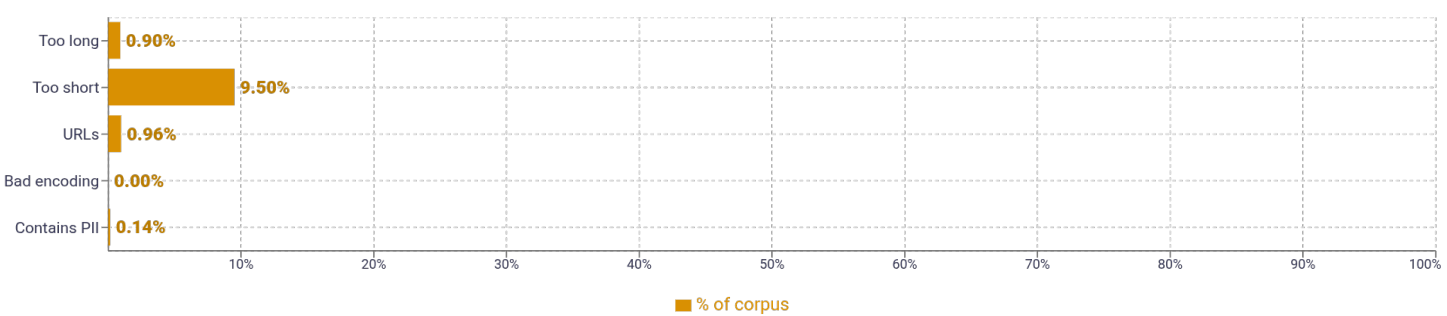
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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

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/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 | | | | |