# HPLT Analytics report

### General overview

| Corpus | Date | Language |
|---|---|---|
| uzn_Latn.jsonl.tsv | 9/7/2024 | Uzbek (uzn) |

### Volumes

| Docs | Segments | Unique segments | Tokens | Characters | Size |
|---|---|---|---|---|---|
| 706,922 | 14,800,770 | 8,877,672 (59.98 %) | 405M | 2,831,493,777 | 2.72 GB |

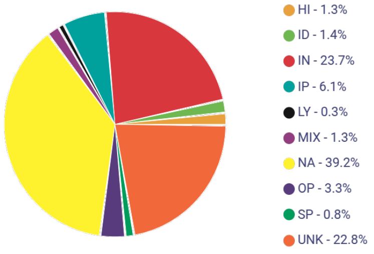
### Top 10 domains

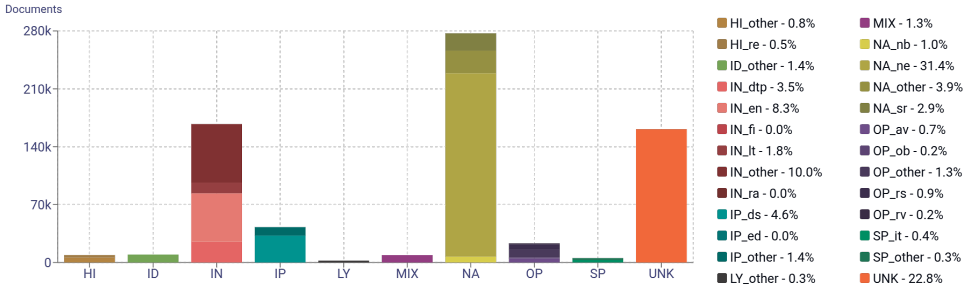
| Domain | Docs | % of total |
|---|---|---|
| wikipedia.org | 61K | 8.63% |
| amerikaovozi.com | 44K | 6.26% |
| daryo.uz | 24K | 3.38% |
| ozodlik.org | 16K | 2.20% |
| ello.uz | 14K | 2.01% |
| ziyouz.com | 10K | 1.44% |
| xit.uz | 8.9K | 1.26% |
| infocom.uz | 8.3K | 1.18% |
| gazeta.uz | 7.1K | 1.01% |
| bbc.com | 6.4K | 0.91% |

### Top 10 TLDs

| Domain | Docs | % of total |
|---|---|---|
| uz | 341K | 48.22% |
| com | 161K | 22.72% |
| org | 99K | 13.96% |
| net | 30K | 4.27% |
| ru | 22K | 3.07% |
| info | 4.3K | 0.61% |
| de | 3.3K | 0.47% |
| biz | 3.3K | 0.47% |
| net.tr | 3.1K | 0.43% |
| su | 2.7K | 0.39% |

## Register labels

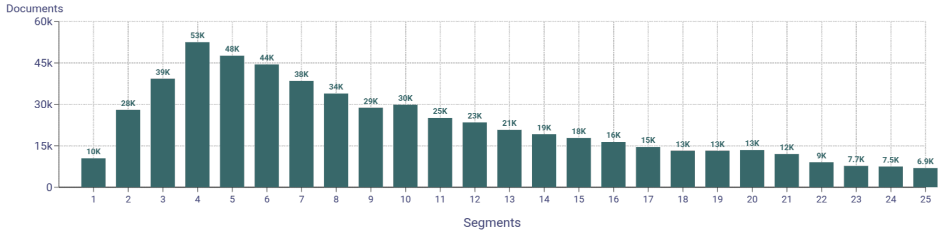




**MT**:18.2% | 128K Documents

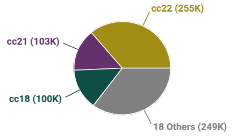
### Documents size (in segments)

<= 25 segments **81.18%** (574K documents)
> 25 segments **18.82%** (133K documents)



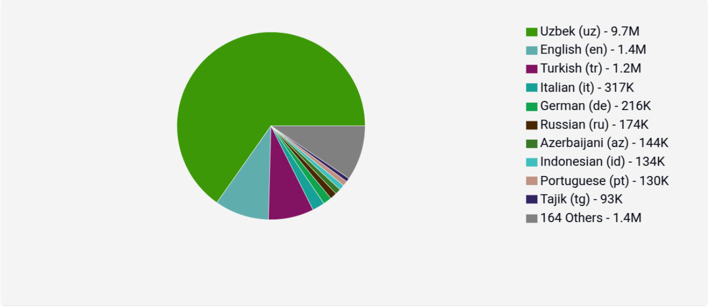
### Documents by collection

**CC = 79.30%**
**IA = 20.70%**

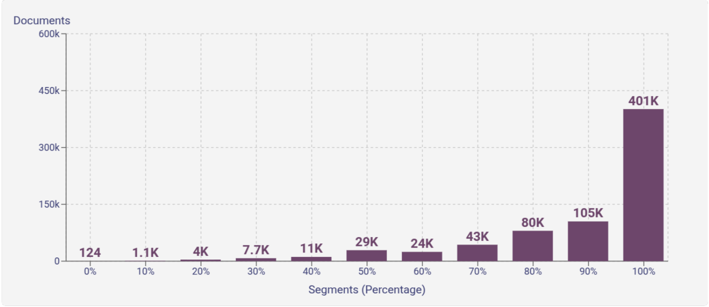


### Language Distribution

**Number of segments in the Uzbek (uzn) corpus**

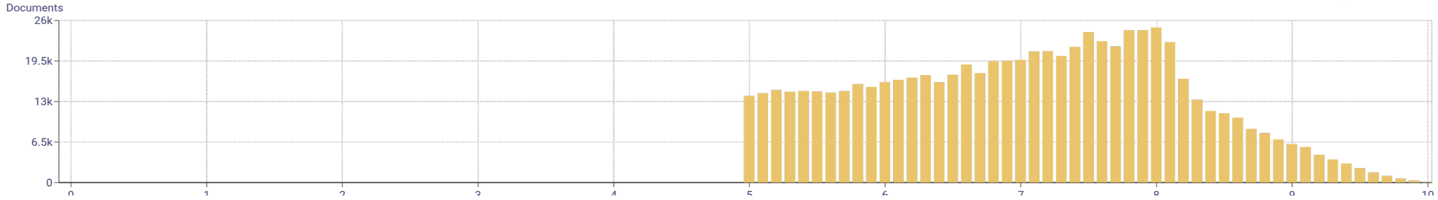


**Percentage of segments in Uzbek (uzn) inside documents**



### Distribution of documents by document score

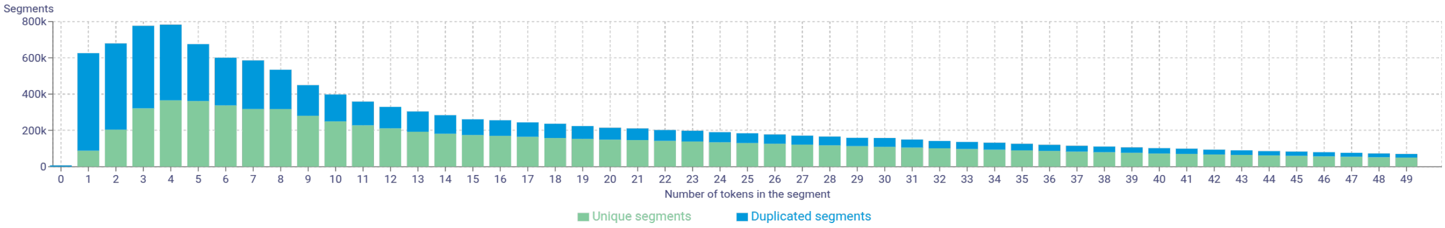
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (707K documents)



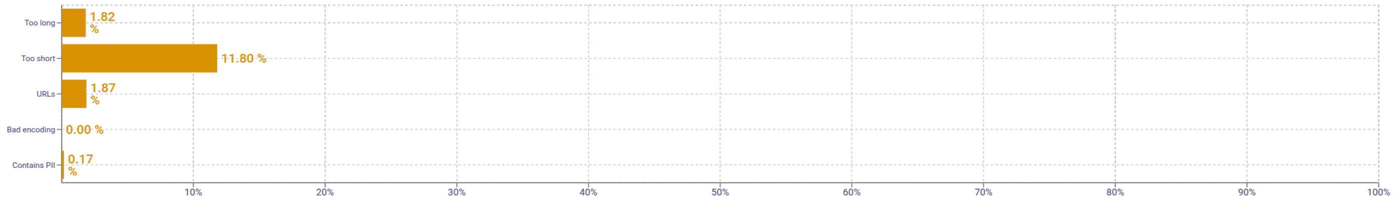
### Segment length distribution by token

≤ **49** tokens = **7.3M** segments | **5.3M** duplicates

> **50** tokens = **2.2M** segments | **627K** duplicates



### Segment noise distribution



### Frequent n-grams

| Size | n-grams |
|---|---|
| 1 | davlat \| 611092<br>yil \| 564781<br>tashkil \| 457537<br>katta \| 434801<br>o‘zbekiston \| 426718 |
| 2 | batafsil ma \| 213103<br>o‘zbekiston respublikasi \| 193166<br>uzbek tilida \| 132712<br>amalga oshirish \| 90683<br>tosh maydalagich \| 76326 |
| 3 | uzbek tilida o'zbekcha \| 51620<br>o'zbekcha tarjima kino \| 46825<br>tilida o'zbekcha tarjima \| 45069<br>hd tas-ix skachat \| 40235<br>respublikasi vazirlar mahkamasining \| 29686 |
| 4 | uzbek tilida o'zbekcha tarjima \| 44988<br>tilida o'zbekcha tarjima kino \| 41929<br>full hd tas-ix skachat \| 25373<br>o‘zbekiston respublikasi vazirlar mahkamasining \| 15168<br>o‘zme. birinchi jild. toshkent \| 11307 |
| 5 | uzbek tilida o'zbekcha tarjima kino \| 41860<br>oliy va o‘rta maxsus ta’lim \| 8357<br>we are searching data for \| 7936<br>searching data for your request \| 7936<br>are searching data for your \| 7936 |

## 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. |
|---|---|
| **Machine-translated** | MT |
| **Lyrical** | LY |
| **Spoken** | SP |
| Interview | it |
| **Interactive discussion** | ID |
| **Narrative** | NA |
| News report | ne |
| Sports report | sr |
| Narrative blog | nb |

| Name | Abbr. |
|---|---|
| **How-to or instructions** | **HI** |
| Recipe | re |
| **Informational persuasion** | **IP** |
| Description with intent to sell | ds |
| News & opinion blog or editorial | ed |
| **Informational description** | **IN** |
| Enciclopedia article | en |
| Research article | ra |

| Name | Abbr. |
|---|---|
| Description of a thing or person | dtp |
| FAQ | fi |
| Legal terms & conditions | lt |
| **Opinion** | **OP** |
| Review | rv |
| Opinion blog | ob |
| Denominational religious blog or sermon | rs |
| Advice | av |