

## General overview

Corpus	Analytics date	Language
uz_1.jsonl.tsv	3/20/2024	Uzbek (uz)

## Volumes

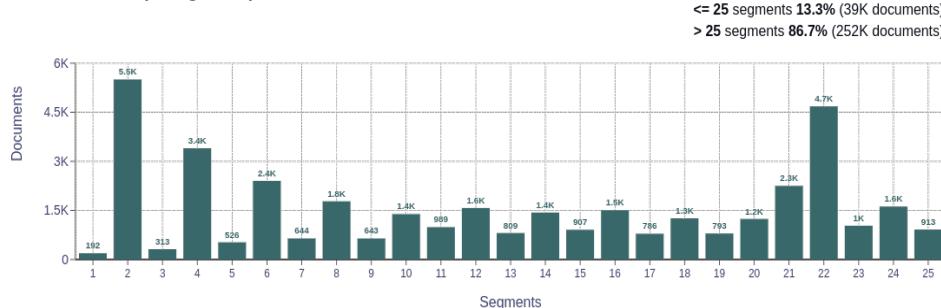
Docs	Segments	Unique segments	Tokens	Size
290,289	37,680,934	50,364 (0.13 %)	435M	3.83 GB

## Type-Token Ratio

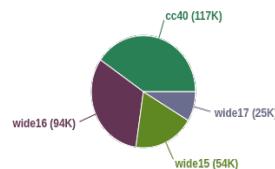
Uzbek (uz)
------------

0.03

## Documents size (in segments)

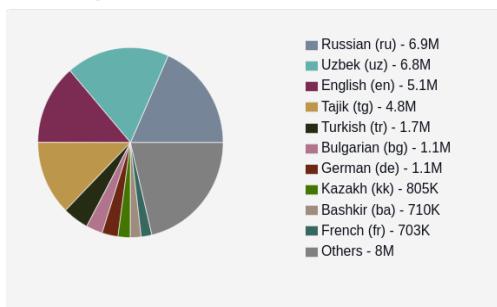


## Documents by collection

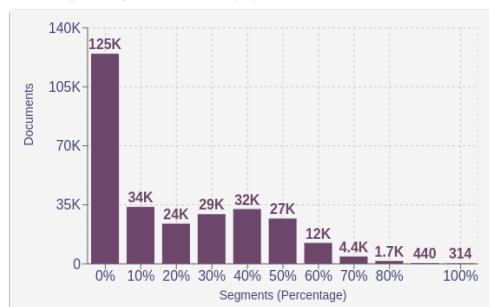


## Language Distribution

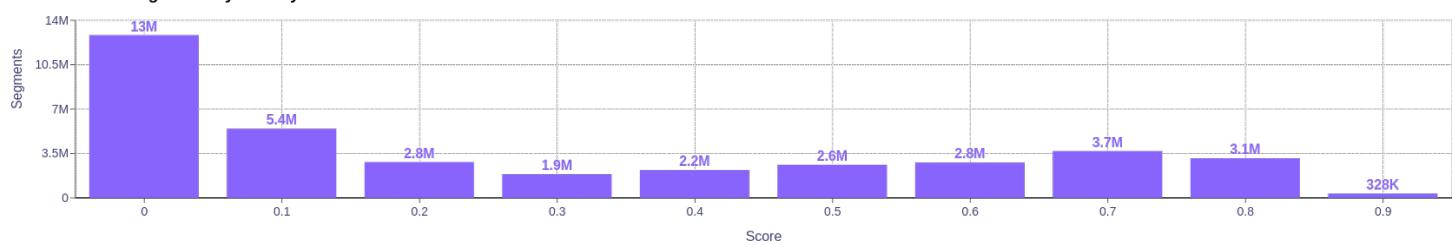
## Number of segments



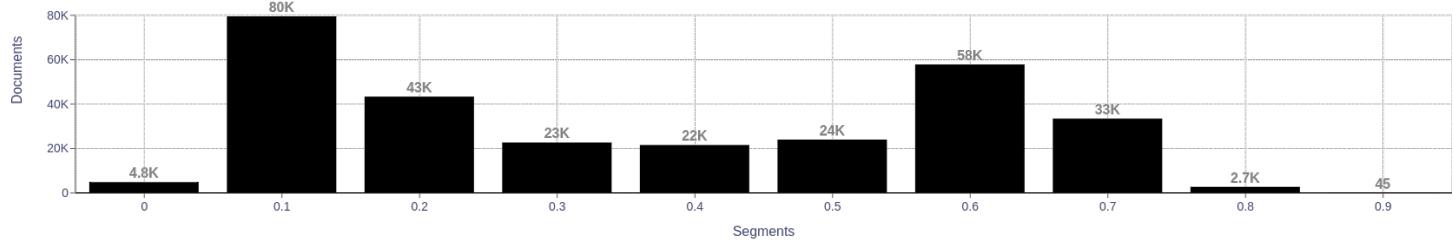
## Percentage of segments in Uzbek (uz) inside documents



## Distribution of segments by fluency score



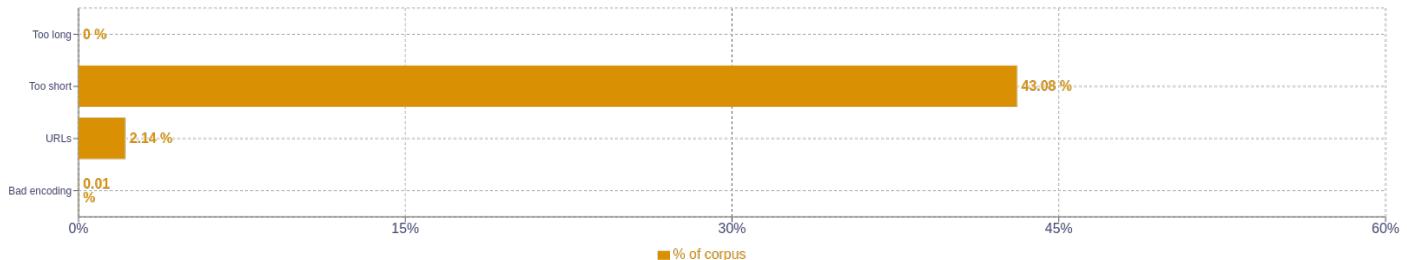
## Distribution of documents by average fluency score



## Segment length distribution by token



## Segment noise distribution



## Frequent n-grams

Size	n-grams
1	(ҳам   1014840) (бир   901803) (ўзбекистон   766960) (у   516526) (br   489531)
2	(ўзбекистон республикаси   351304) (марта ўқилган   140347) (o'zbekiston respublikasi   93956) (хар бир   93308) (bosh sahifa   86715)
3	(ўзбекистон республикаси олий   50967) (ўзбекистон республикаси вазирлар   41753) (ўзбекистон республикаси президентининг   40526) (ўзбекистон республикаси президенти   30614) (республикаси вазирлар маҳкамасининг   28368)
4	(ўзбекистон республикаси вазирлар маҳкамасининг   28192) (ўзбекистон республикаси олий мажлиси   22104) (соллаллоҳу алайхи ва саллам   18367) (бесплатные анимационные смайлики для   14972) (анимационные смайлики для одноклассники.ру   14947)
5	(бесплатные анимационные смайлики для одноклассники.ру   14946) (турецкие сериалы онлайн на русском   11705) (смотреть сериал великолепный век все   11444) (сериал великолепный век все серии   11444) (все серии на русском языке   11444)

## 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 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>, <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>).

### 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.

### 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>