

General overview

Corpus	Date	Language
hplt-v3-bak_Cyrl	9/18/2025	Bashkir (ba)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
275,718	3,968,763	3,135,566 (79.01 %)	135M	800,026,131	1.35 GB

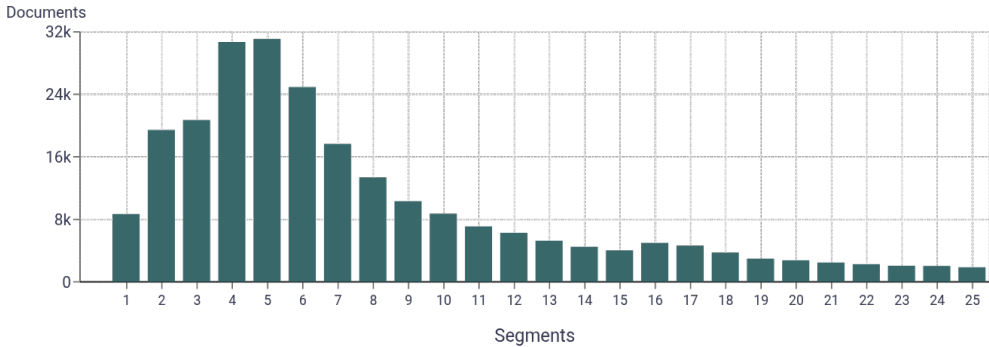
Top 10 domains

Domain	Docs	% of total
wikipedia.org	30K	10.85%
bashinform.ru	27K	9.93%
bashgazet.ru	16K	5.83%
ye02.ru	8.2K	2.96%
hakmar.ru	8K	2.90%
ye102.ru	7.8K	2.82%
башкирская-энци...	6.7K	2.44%
tv-rb.ru	6.4K	2.33%
gtrk.tv	5.4K	1.94%
ural-rb.ru	5K	1.80%

Top 10 TLDs

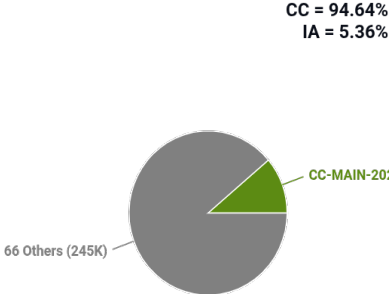
Domain	Docs	% of total
ru	169K	61.32%
org	38K	13.61%
com	24K	8.82%
info	10K	3.73%
рф	7.8K	2.81%
news	6.5K	2.37%
tv	5.4K	1.95%
eu	2.5K	0.91%
su	2.3K	0.84%
jp	1.3K	0.46%

Documents size (in segments) ⓘ



≤ 25 segments **88.01%** (243K documents)  
> 25 segments **11.99%** (33K documents)

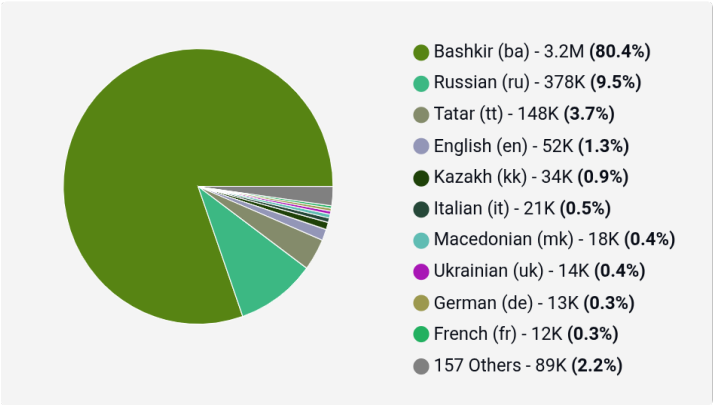
Document collections



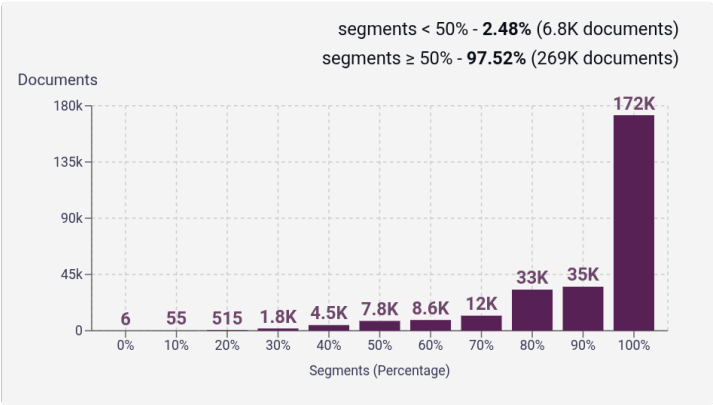
CC = 94.64%  
IA = 5.36%

Language Distribution

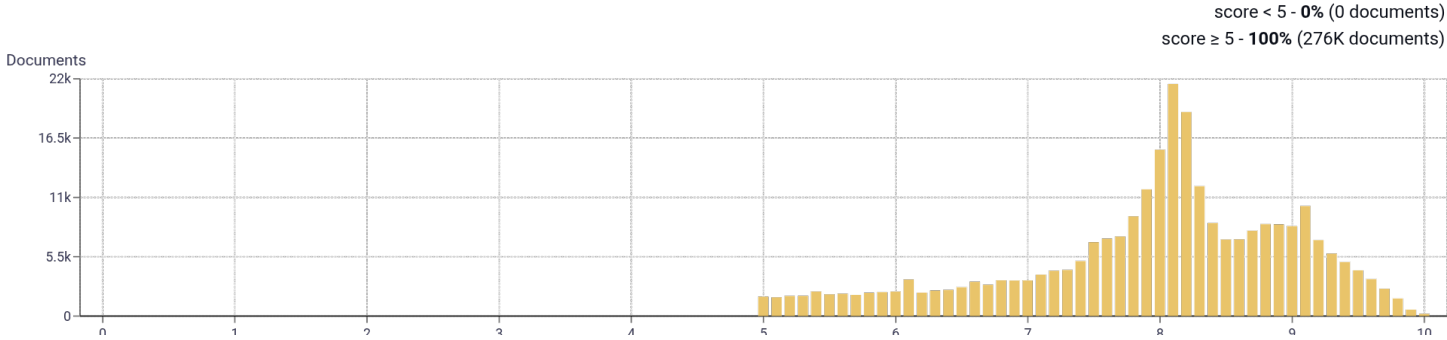
Number of segments in the Bashkir (ba) corpus



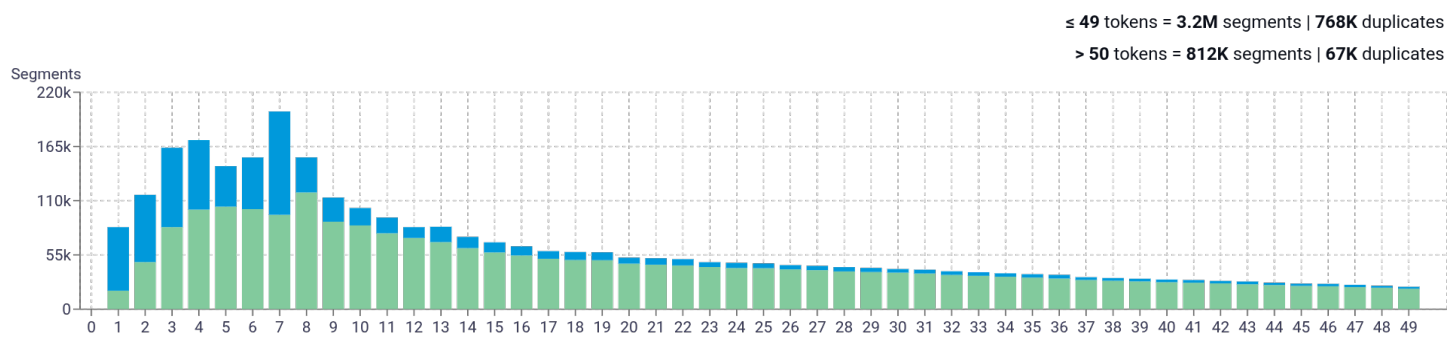
Percentage of segments in Bashkir (ba) inside documents



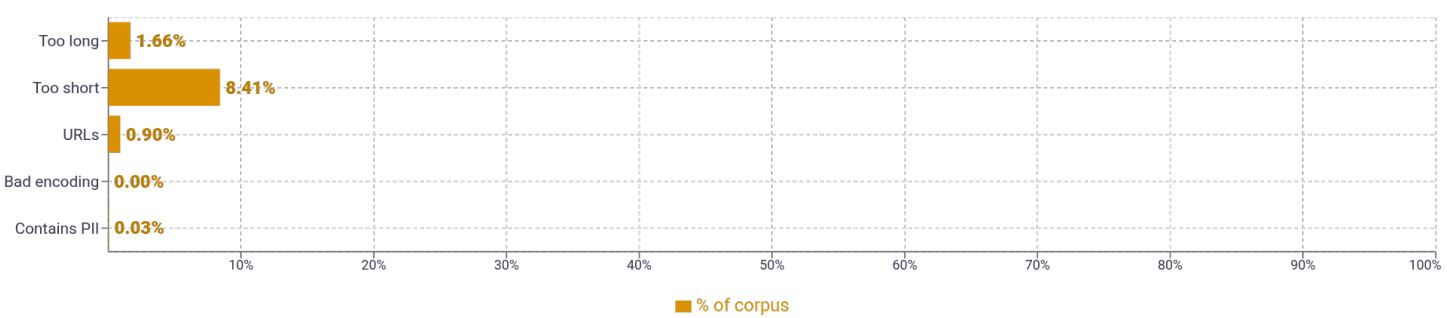
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	бер   611,491   үзгәртәргә   286,918   ук   241,392   йыл   224,647   башкорт   220,735	📄
2	вики-тексты үзгәртәргә   132,661   бер нисә   59,217   хайһы бер   36,493   ауыл хужалығы   35,817   башкортостан республикаһының   31,334	📄
3	бөйөк ватан һуғышы   11,357   бөйөк ватан һуғышында   9,076   башкортостан республикаһының атказанған   8,362   тәүге сығанактан архивланған   7,987   теле һәм әҙәбиәте   7,635	📄
4	башкорт теле һәм әҙәбиәте   5,757   һанына тамамланған йылдарҙа тыуғандар   4,331   башкортостан башлығы радиһы хәбирә   3,843   теле һәм әҙәбиәте уҡытыусыһы   3,062   башлығы вазиһаһын ваҡытлыса башкарыусы   2,898	📄
5	хеҙмәт һәм халықты социаль яклау   2,576   башлығы вазиһаһын ваҡытлыса башкарыусы радиһы   2,394   сығышы менән хәҙерге башкортостан республикаһының   2,204   башкорт теле һәм әҙәбиәте уҡытыусыһы   2,148   салауат юлаев исемендәге дәүләт премияһы   2,104	📄

# 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  $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$ , 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	ntp
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				