HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-tat_Cyrl	9/18/2025	Tatar (tt)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,259,059	22,923,640	12,418,183 (54.17 %)	636M	3,669,676,634	6.21 GB

Top 10 domains

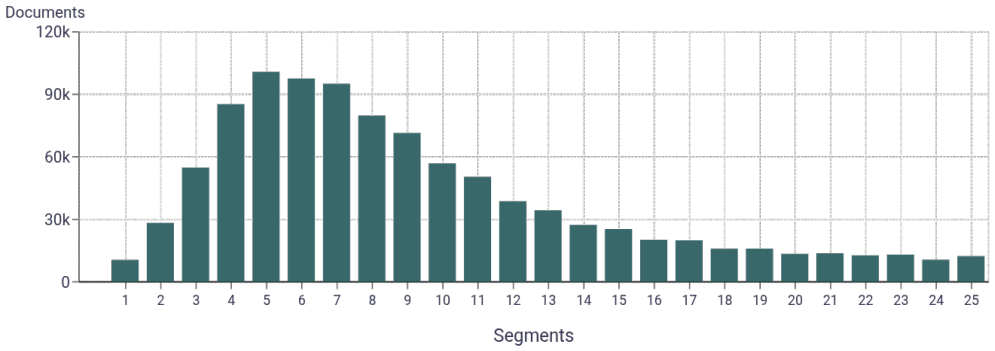
Domain	Docs	% of total
azatliq.org	60K	4.74%
tatar-inform.tatar	51K	4.07%
wikipedia.org	30K	2.38%
shahrikazan.ru	29K	2.28%
syuyumbike.ru	29K	2.27%
alabuganury.ru	28K	2.25%
yalkyn.ru	26K	2.04%
baltaci.ru	25K	1.98%
muslumirc.ru	22K	1.78%
shahrichalli.ru	22K	1.77%

Top 10 TLDs

Domain	Docs	% of total
ru	939K	74.60%
org	125K	9.96%
tatar	83K	6.56%
com	70K	5.57%
info	12K	0.93%
рф	5.9K	0.47%
net.tr	5.7K	0.45%
net	3.1K	0.25%
su	2.1K	0.17%
biz	1.3K	0.10%

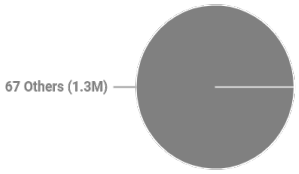
Documents size (in segments)

≤ 25 segments **79.8%** (1M documents)
> 25 segments **20.2%** (254K documents)



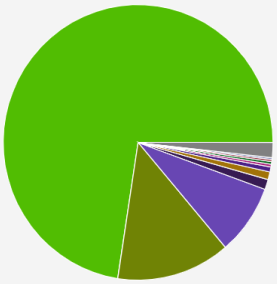
Document collections

CC = 97.31%
IA = 2.69%



Language Distribution

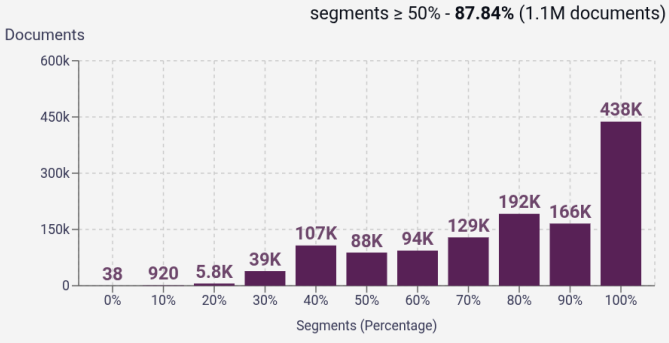
Number of segments in the Tatar (tt) corpus



- Tatar (tt) - 17M **(72.6%)**
- Russian (ru) - 3.1M **(13.6%)**
- Italian (it) - 1.9M **(8.3%)**
- Chuvash (cv) - 270K **(1.2%)**
- Bashkir (ba) - 209K **(0.9%)**
- English (en) - 134K **(0.6%)**
- German (de) - 72K **(0.3%)**
- Ukrainian (uk) - 71K **(0.3%)**
- Kazakh (kk) - 54K **(0.2%)**
- Bulgarian (bg) - 49K **(0.2%)**
- 160 Others - 408K **(1.8%)**

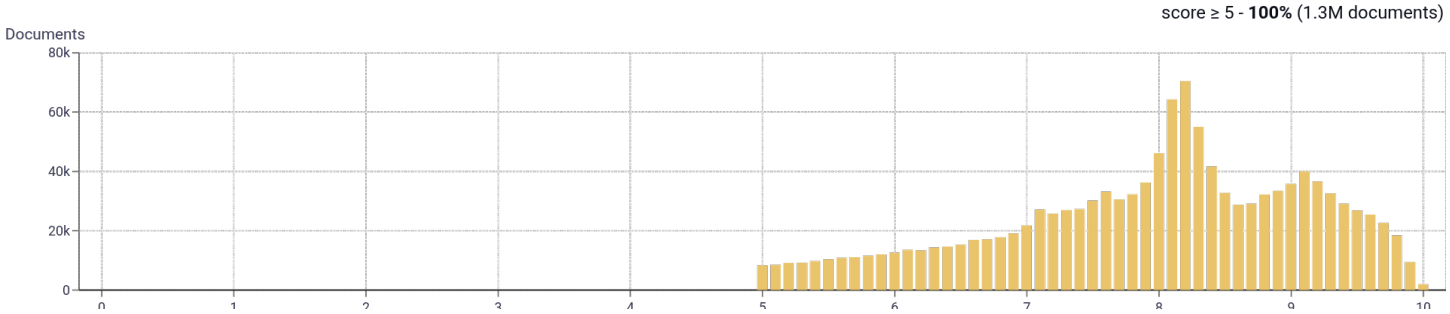
Percentage of segments in Tatar (tt) inside documents

segments < 50% - **12.16%** (153K documents)
segments ≥ 50% - **87.84%** (1.1M documents)



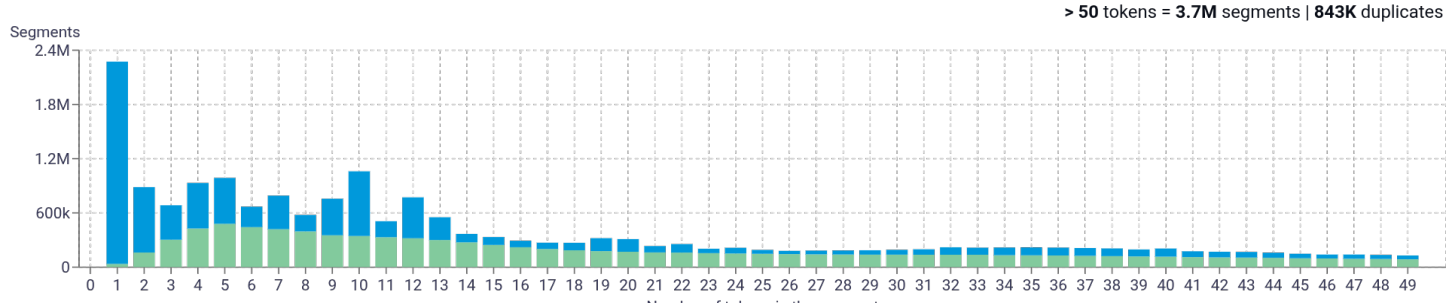
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (1.3M documents)

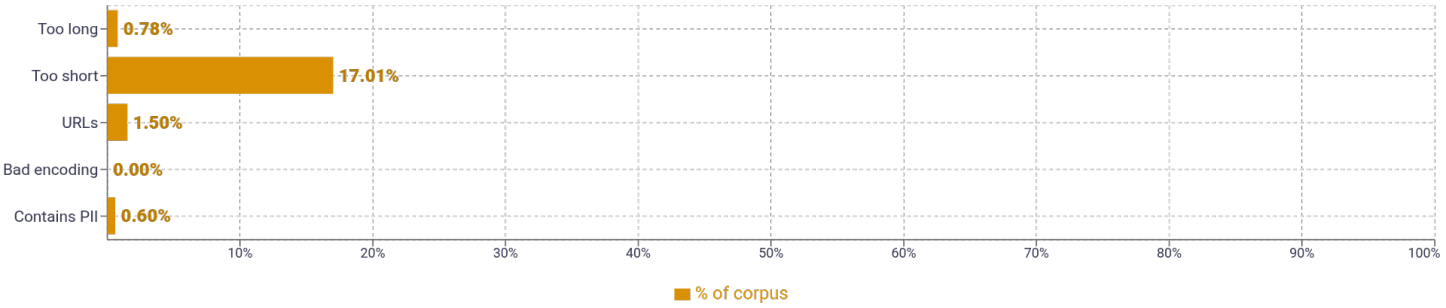


Segment length distribution by token

≤ 49 tokens = **19M** segments | **9.7M** duplicates
> 50 tokens = **3.7M** segments | **843K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	татар | 1,297,844 · и | 1,106,538 · татарстан | 976,007 · в | 964,096 · булган | 918,907
2	на сайте | 294,047 · кызыклы язмаларны | 252,007 · каналында укыгыз | 251,193 · татмедиа telegram | 233,408 · язмаларны татмедиа | 233,404
3	мөһим һәм кызыклы | 233,513 · язмаларны татмедиа telegram | 233,404 · кызыклы язмаларны татмедиа | 233,404 · самым важным и | 217,704 · интересным в telegram | 217,704
4	мөһим һәм кызыклы язмаларны | 233,405 · кызыклы язмаларны татмедиа telegram | 233,404 · самым важным и интересным | 217,704 · и интересным в telegram | 217,704 · за самым важным и | 217,704
5	мөһим һәм кызыклы язмаларны татмедиа | 233,404 · самым важным и интересным в | 217,704 · за самым важным и интересным | 217,704 · важным и интересным в telegram | 217,704 · следите за самым важным и | 217,696

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				