HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-crh_Latn	9/17/2025	Crimean Tatar

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
120,315	1,528,212	1,270,190 (83.12 %)	52M	314,517,891	341.87 MB

Top 10 domains

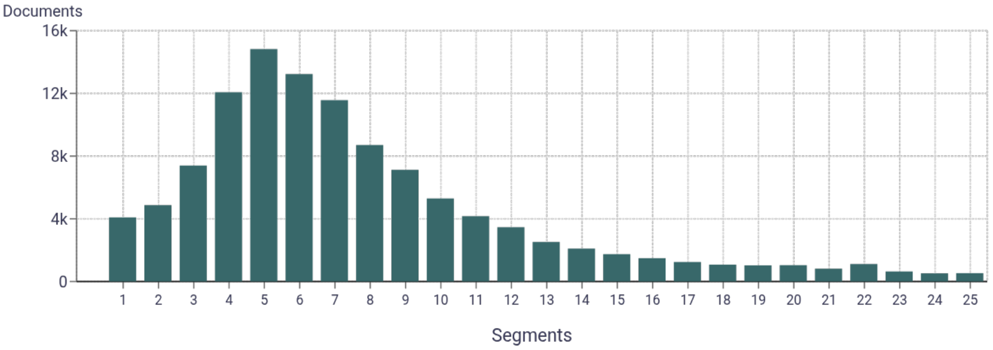
Domain	Docs	% of total
krymr.com	29K	24.33%
azatliq.org	22K	18.67%
trt.net.tr	10K	8.68%
qazaqtimes.com	8.6K	7.13%
inform.kz	5.5K	4.58%
wikipedia.org	4.5K	3.71%
abai.kz	4.4K	3.69%
minre.gov.ua	3.1K	2.61%
kazgazeta.kz	2.7K	2.26%
avdet.org	2.1K	1.75%

Top 10 TLDs

Domain	Docs	% of total
com	41K	34.10%
org	32K	26.95%
kz	24K	19.54%
net.tr	10K	8.68%
gov.ua	3.9K	3.23%
uz	3.7K	3.11%
ru	931	0.77%
media	865	0.72%
ua	794	0.66%
tatar	638	0.53%

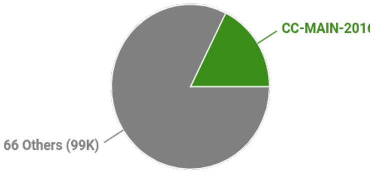
Documents size (in segments) ⓘ

≤ **25** segments **93.66%** (113K documents)
> **25** segments **6.34%** (7.6K documents)



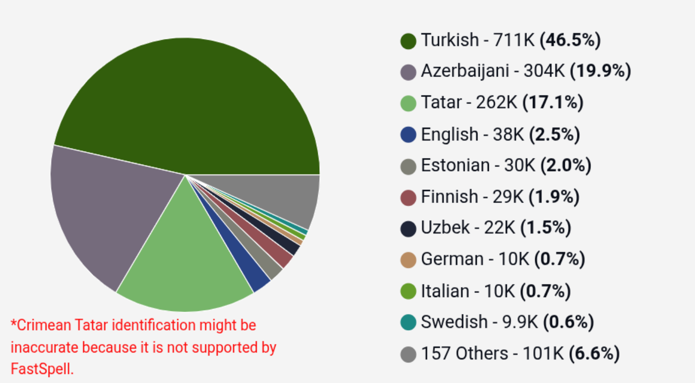
Document collections

CC = 90.28%
IA = 9.72%



Language Distribution

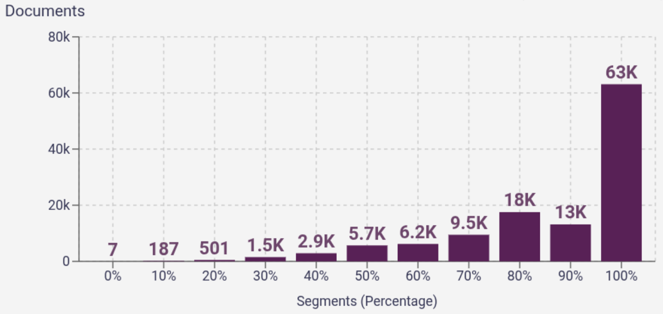
Number of segments in the Crimean Tatar corpus



*Crimean Tatar identification might be inaccurate because it is not supported by FastSpell.

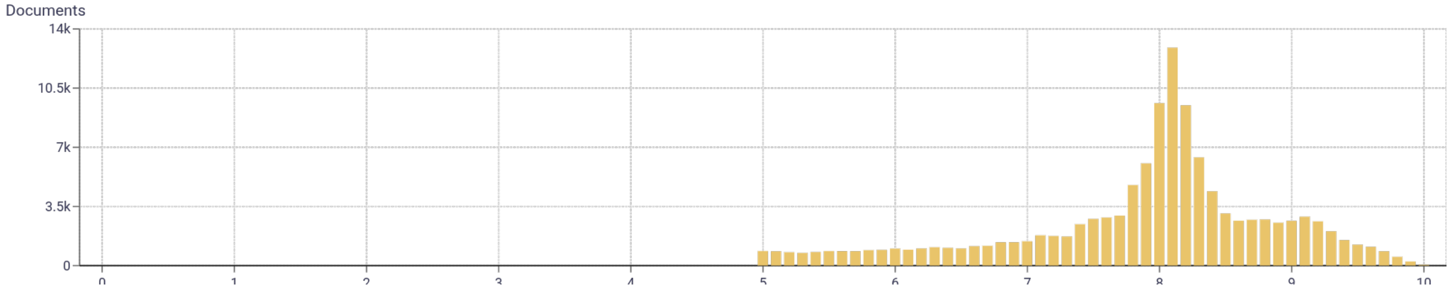
Percentage of segments in Crimean Tatar inside documents

segments < 50% - **4.25%** (5.1K documents)
segments ≥ 50% - **95.75%** (115K documents)



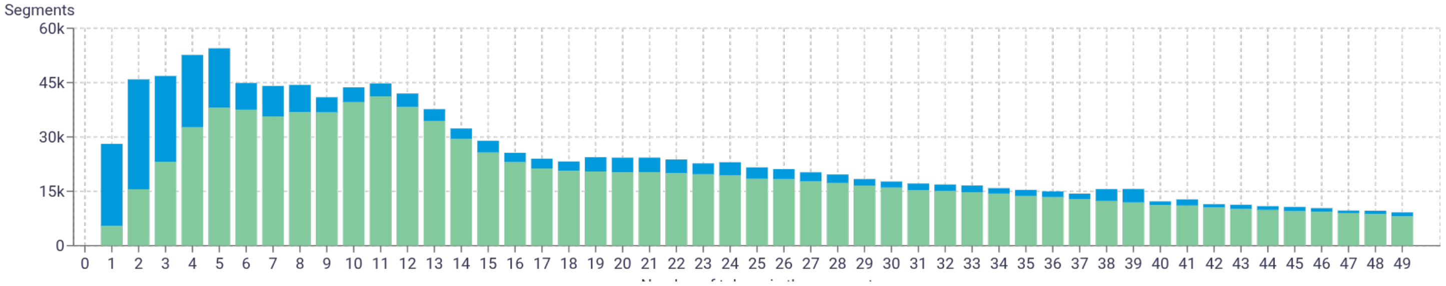
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (120K documents)

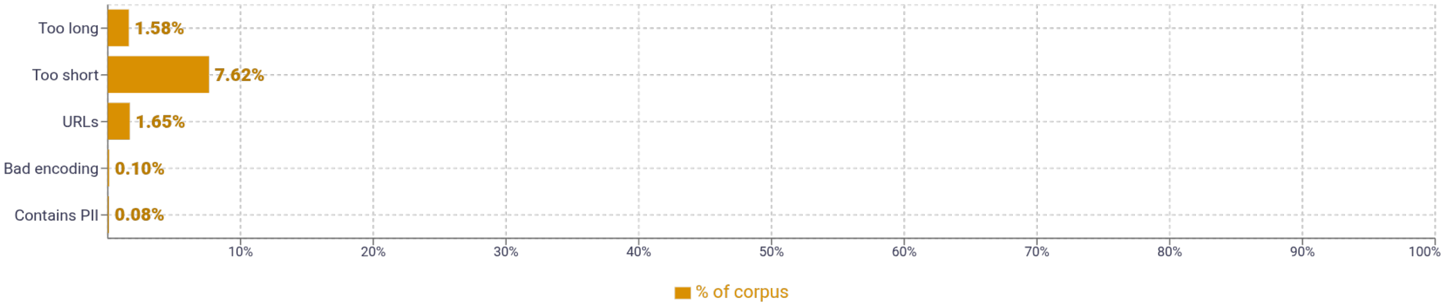


Segment length distribution by token

≤ **49** tokens = **1.2M** segments | **235K** duplicates
> **50** tokens = **304K** segments | **24K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	häm | 163,189 dep | 155,887 men | 118,055 belän | 117,852 qırım | 114,720
2	işğal etilgen | 14,780 qırımtatar milliy | 11,857 hizb ut | 10,991 qayd etti | 10,613 şulay uq | 8,191
3	aqiqat saytını blok | 4,860 vastasınen oqumaq mümkün | 4,859 saytını blok etti | 4,858 saytı vastasınen oqumaq | 4,858 saifelerinden taqip etiñiz | 4,848
4	saytı vastasınen oqumaq mümkün | 4,858 aqiqat saytını blok etti | 4,858 instagram saifelerinden taqip etiñiz | 4,848 telegram ve instagram saifelerinden | 4,745 aqiqatnıñ telegram ve instagram | 4,745
5	telegram ve instagram saifelerinden taqip | 4,745 aqiqatnıñ telegram ve instagram saifelerinden | 4,745 küzgü saytı vastasınen oqumaq mümkün | 2,729 aqiqatnı küzgü saytı vastasınen oqumaq | 2,729 aqiqatnıküzgü saytı vastasınen oqumaq mümkün | 2,129

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				