HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-tuk_Latn	9/18/2025	Turkmen (tk)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
378,448	4,863,457	3,515,564 (72.29 %)	140M	898,266,128	941.32 MB

Top 10 domains

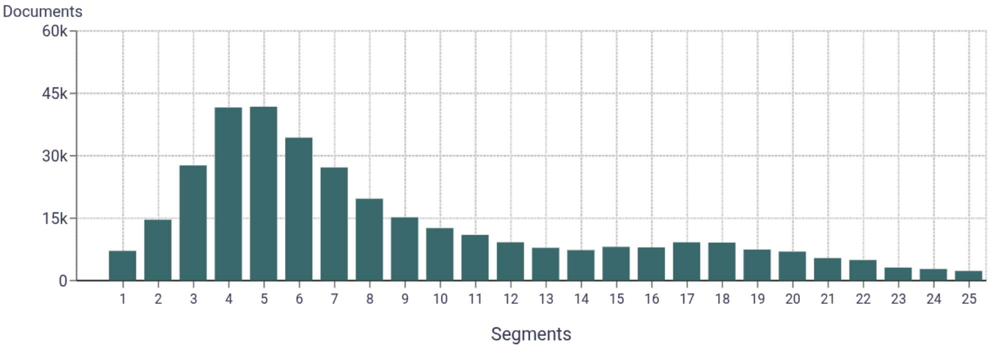
Domain	Docs	% of total
azathabar.com	65K	17.22%
inform.kz	33K	8.66%
egemen.kz	29K	7.79%
atavatan-turkme...	11K	2.95%
trt.net.tr	11K	2.87%
turkmenportal.com	11K	2.79%
tmcars.info	11K	2.79%
ertir.com	9K	2.38%
talyplar.com	8.6K	2.27%
business.com.tm	8.3K	2.19%

Top 10 TLDs

Domain	Docs	% of total
com	165K	43.51%
kz	90K	23.75%
gov.tm	48K	12.71%
com.tm	19K	5.04%
info	12K	3.22%
net.tr	11K	2.87%
org	8.5K	2.24%
news	4.8K	1.28%
tm	4K	1.06%
edu.tm	3.9K	1.04%

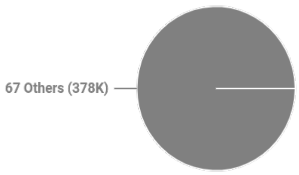
Documents size (in segments) ⓘ

≤ **25** segments **91.05%** (345K documents)
> 25 segments **8.95%** (34K documents)



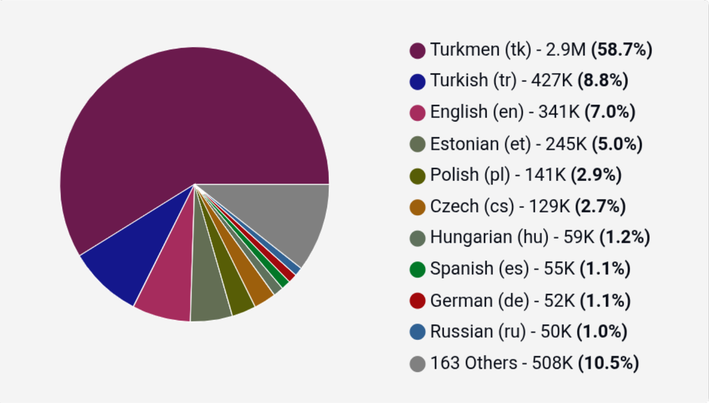
Document collections

CC = 95.42%
IA = 4.58%



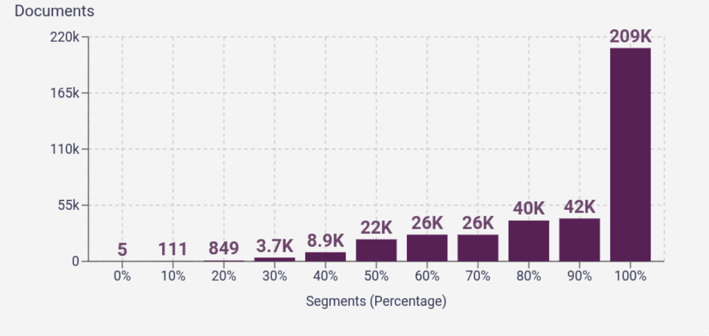
Language Distribution

Number of segments in the Turkmen (tk) corpus



Percentage of segments in Turkmen (tk) inside documents

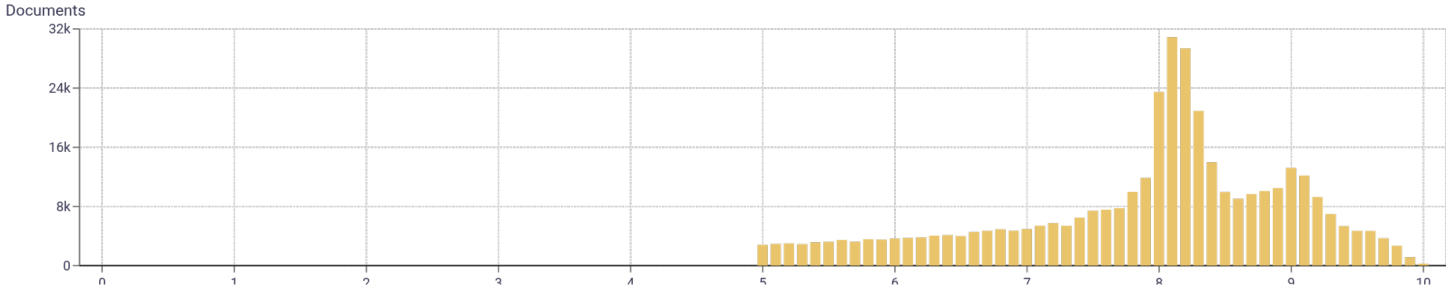
segments < 50% - **3.59%** (14K documents)
segments ≥ 50% - **96.41%** (365K documents)



Distribution of documents by document score

score < 5 - **0%** (0 documents)

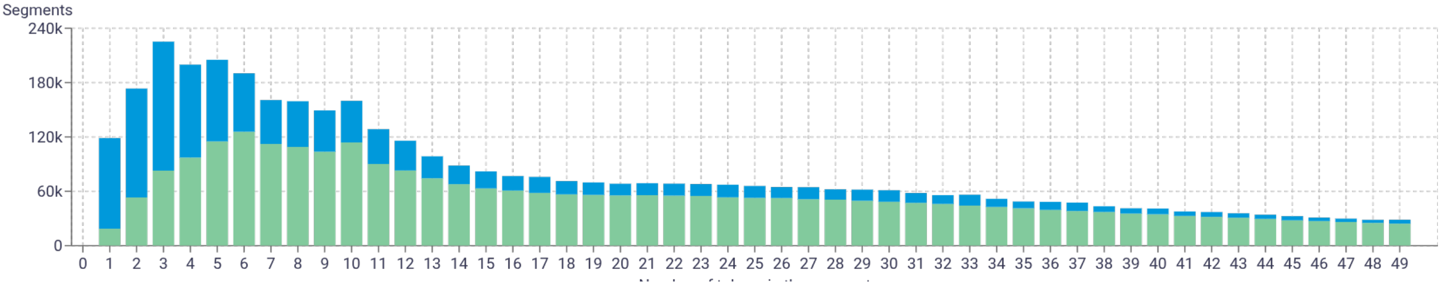
score ≥ 5 - **100%** (378K documents)



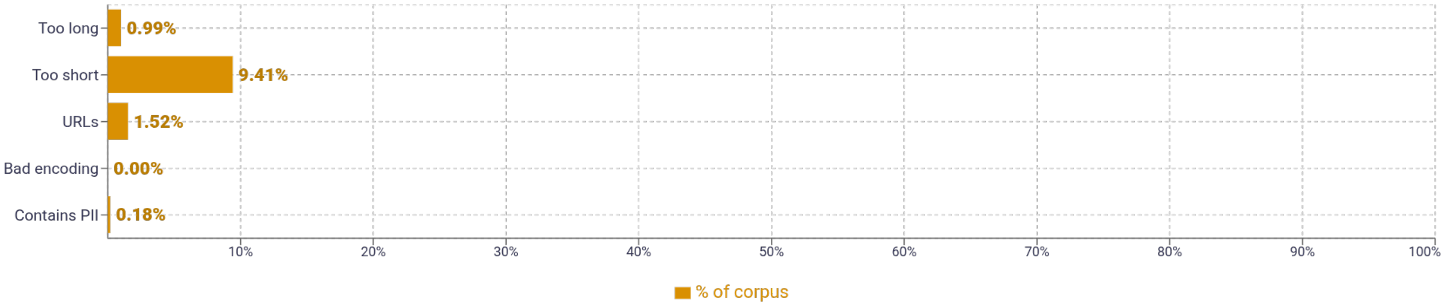
Segment length distribution by token

≤ **49** tokens = **4.1M** segments | **1.3M** duplicates

> 50 tokens = **783K** segments | **76K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	we | 1,645,540 bilen | 943,826 bu | 640,523 üçin | 595,520 hem | 588,371
2	nji ýylyň | 76,763 şeýle hem | 76,362 nji ýylda | 61,267 dep habarlaıdy | 50,433 hormatly prezidentimiz | 43,309
3	dep habarlaıdy qazaqparat | 32,696 habarlaıdy qazaqparat tilshisi | 21,712 hormatly prezidentimiz gurbanguly | 19,454 joldaý júktegen mindetter | 19,057 álemdik baq qazaqstan | 19,036
4	dep habarlaıdy qazaqparat tilshisi | 21,667 álemdik baq qazaqstan týraly | 19,036 hormatly prezidentimiz gurbanguly berdimuhamedow | 10,876 şol bir wagtyň özünde | 8,916 türkmen halkynyň milli lideri | 8,598
5	berkarar döwletiň täze eýýamynyň galkynyşy | 7,186 döwletiň täze eýýamynyň galkynyşy döwründe | 5,397 adatdan daşary we doly ygtyýarly | 4,386 türkmenistanyň halk maslahatynyň başlygy gurbanguly | 4,201 türkmenistanyň daşary işler ministrliginiň halkara | 4,088

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				