

General overview

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

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

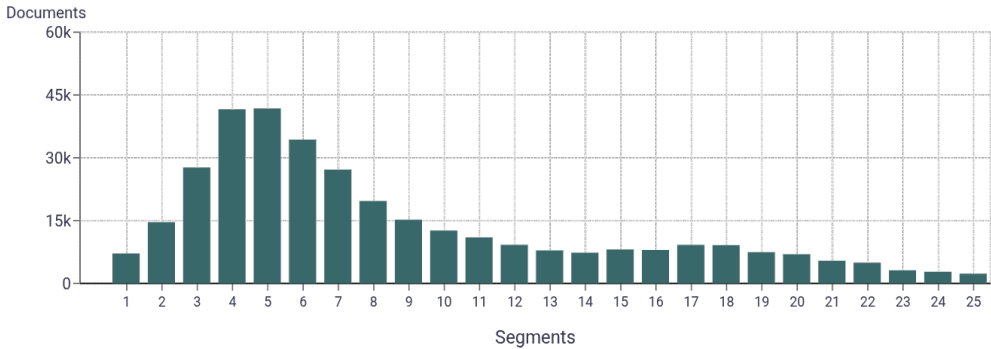
Top 10 domains

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%

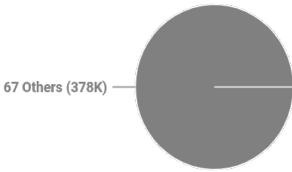
Documents size (in segments) ⓘ



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

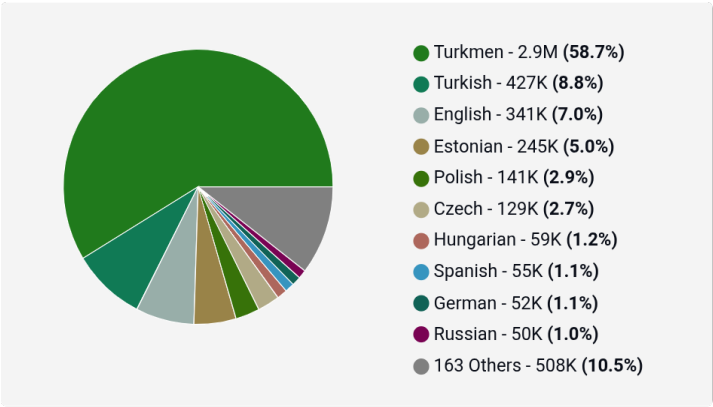
Document collections

CC = **95.42%**
IA = **4.58%**

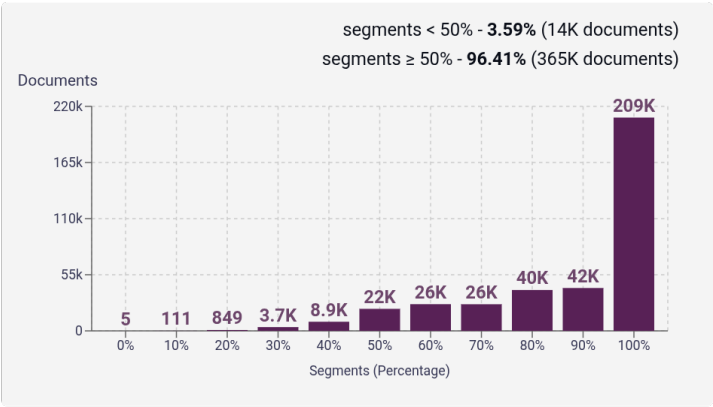


Language Distribution

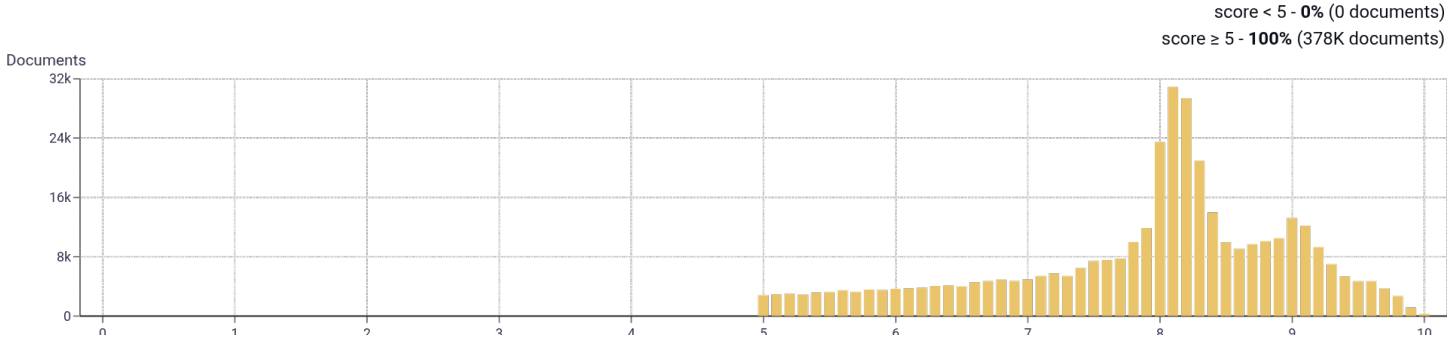
Number of segments in the Turkmen corpus



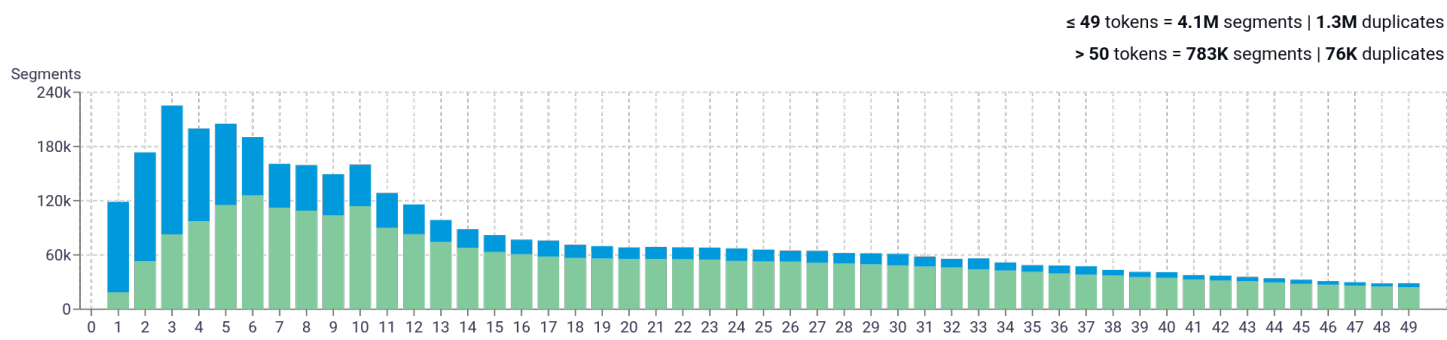
Percentage of segments in Turkmen inside documents



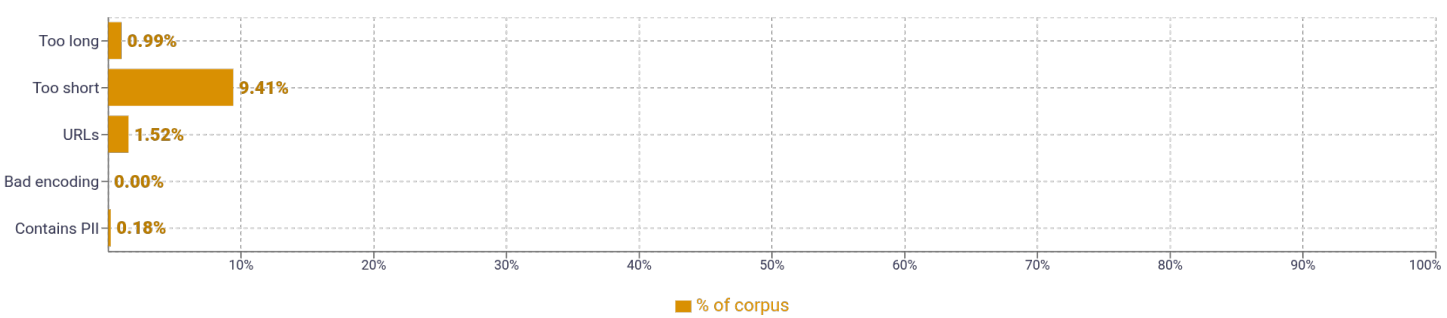
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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>, , , 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				