HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-uig_Arab	9/18/2025	Uyghur (ug)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
645,397	10,466,635	6,763,305 (64.62 %)	347M	2,148,294,034	3.7 GB

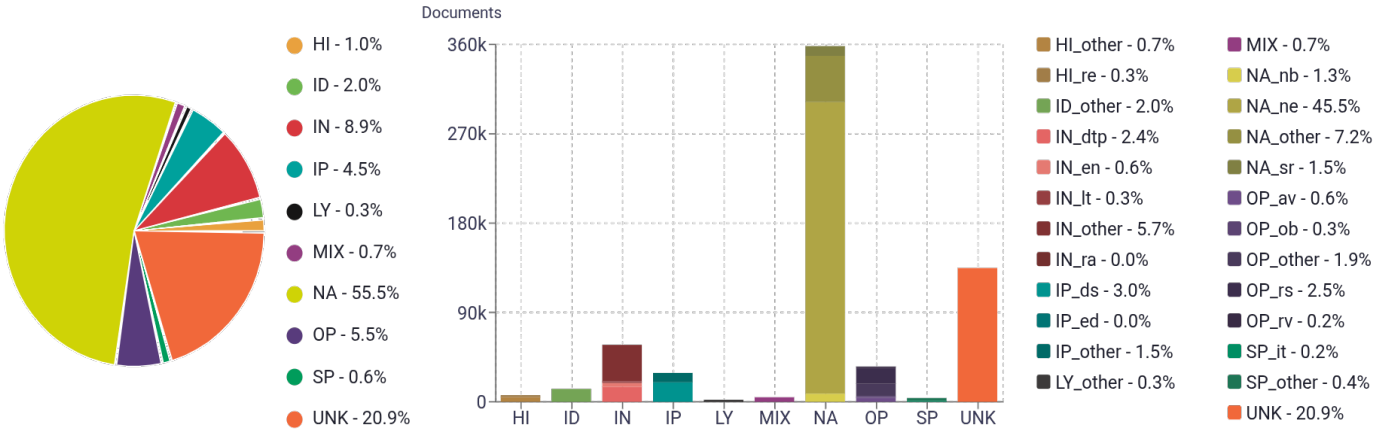
Top 10 domains

Domain	Docs	% of total
people.com.cn	66K	10.18%
inform.kz	50K	7.70%
egemen.kz	33K	5.04%
ts.cn	31K	4.80%
rfa.org	27K	4.20%
trt.net.tr	15K	2.38%
nur.cn	14K	2.12%
rfaweb.org	12K	1.83%
abai.kz	12K	1.79%
chinabroadcast.cn	11K	1.69%

Top 10 TLDs

Domain	Docs	% of total
com	213K	33.00%
kz	127K	19.63%
cn	121K	18.80%
com.cn	73K	11.38%
org	52K	8.07%
net.tr	15K	2.38%
net	14K	2.18%
biz	8.1K	1.26%
cc	6.9K	1.06%
gov.cn	4K	0.62%

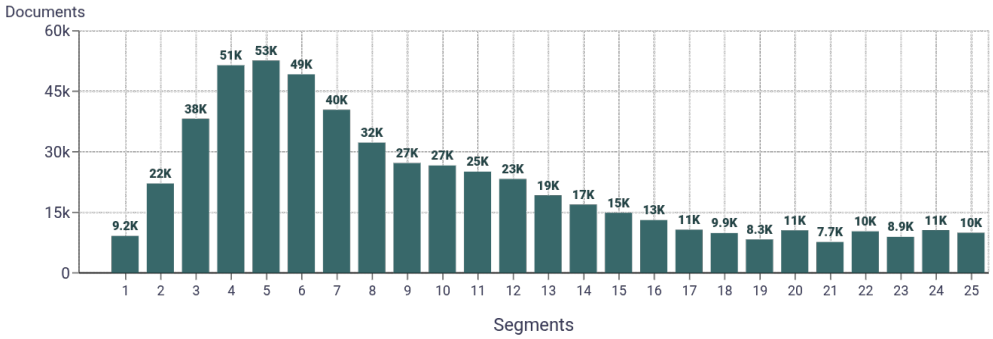
Register labels



MT:8.7% | 56K Documents

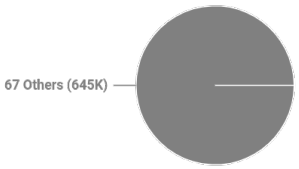
Documents size (in segments)

≤ 25 segments **85.04%** (549K documents)
> 25 segments **14.96%** (97K documents)



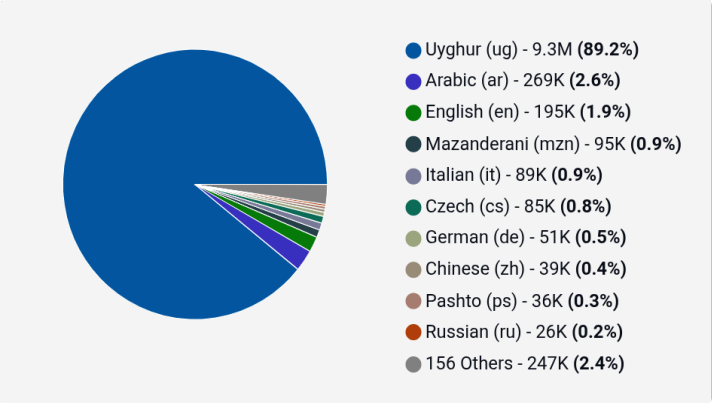
Document collections

CC = 62.76%
IA = 37.24%

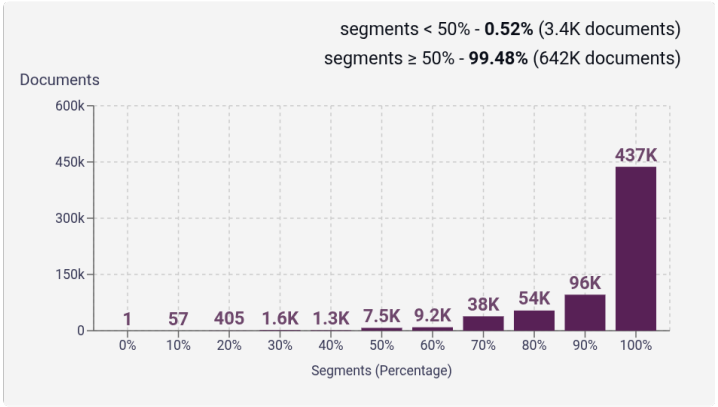


Language Distribution

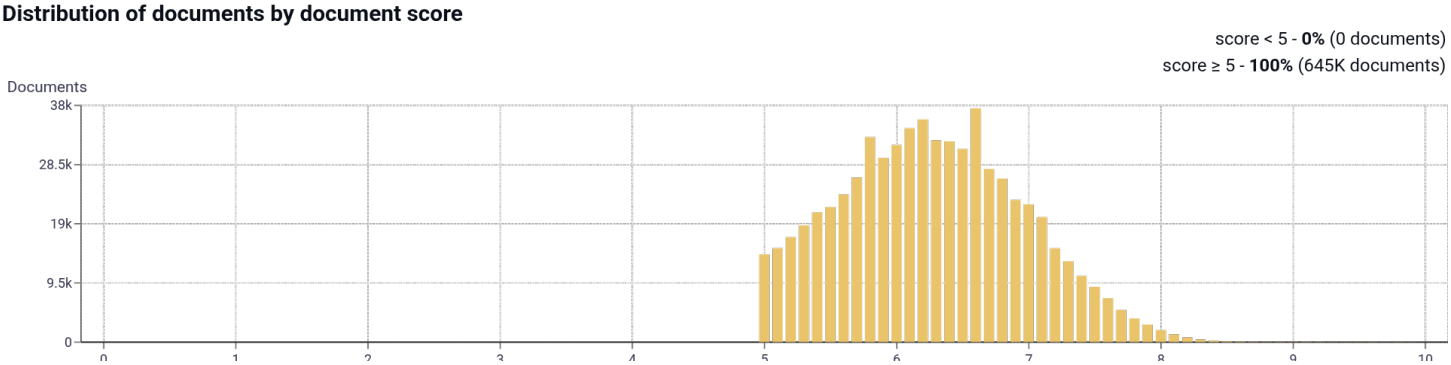
Number of segments in the Uyghur (ug) corpus



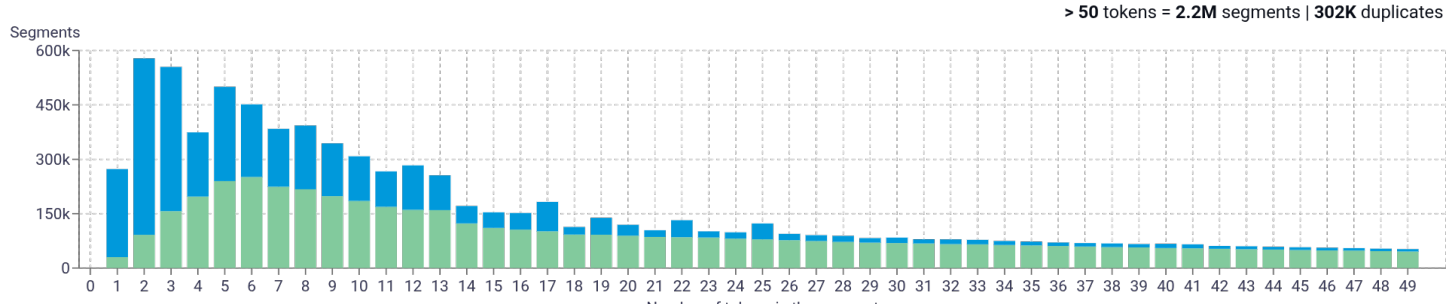
Percentage of segments in Uyghur (ug) inside documents



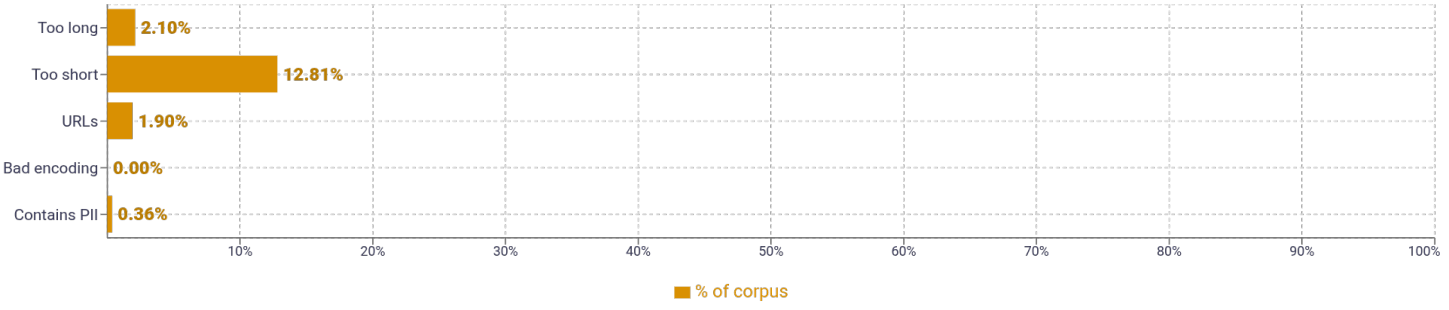
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	بىلەن | 1,709,206 · مەن | 1,055,228 · دەپ | 905,541 · م | 859,067 · بولۇپ | 805,073
2	خەلق تورى | 76,190 · مۇنداق دېدى | 73,407 · شۇنىڭ بىلەن | 65,726 · كەلۈ قايناري | 58,744 · ئاپتونوم رايونلۇق | 57,117
3	ۋقىعى حالىق تورابنا | 54,731 · حالىق تورابنا ءتان | 54,731 · خەلق تورى ئۇيغۇرچە | 30,879 · all rights reserved | 30,703 · جۇڭحۇا حالىق رەسپۇبليكاسىنىڭ | 29,000
4	ۋقىعى حالىق تورابنا ءتان | 54,731 · مەنشىك ۋقىعى حالىق تورابنا | 26,728 · بەتتىڭ مەنشىك ۋقىعى حالىق | 26,728 · بۇل بەتتىڭ مەنشىك ۋقىعى | 26,728 · ەگەر باسقارۋ جاعىنا پىكىربىڭىز | 26,727
5	مەنشىك ۋقىعى حالىق تورابنا ءتان | 26,728 · بەتتىڭ مەنشىك ۋقىعى حالىق تورابنا | 26,728 · بۇل بەتتىڭ مەنشىك ۋقىعى حالىق | 26,728 · ەگەر باسقارۋ جاعىنا پىكىربىڭىز بولسا | 26,727 · گازەتتى مەكەمەسىنىڭ توراپ ورتالىعىنا اكىس | 26,727

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				