

General overview

Corpus	Date	Language
hplt-v3-azb_Arab	9/16/2025	Azerbaijani (az)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
94,756	2,584,448	1,756,065 (67.95 %)	55M	293,610,352	509.44 MB

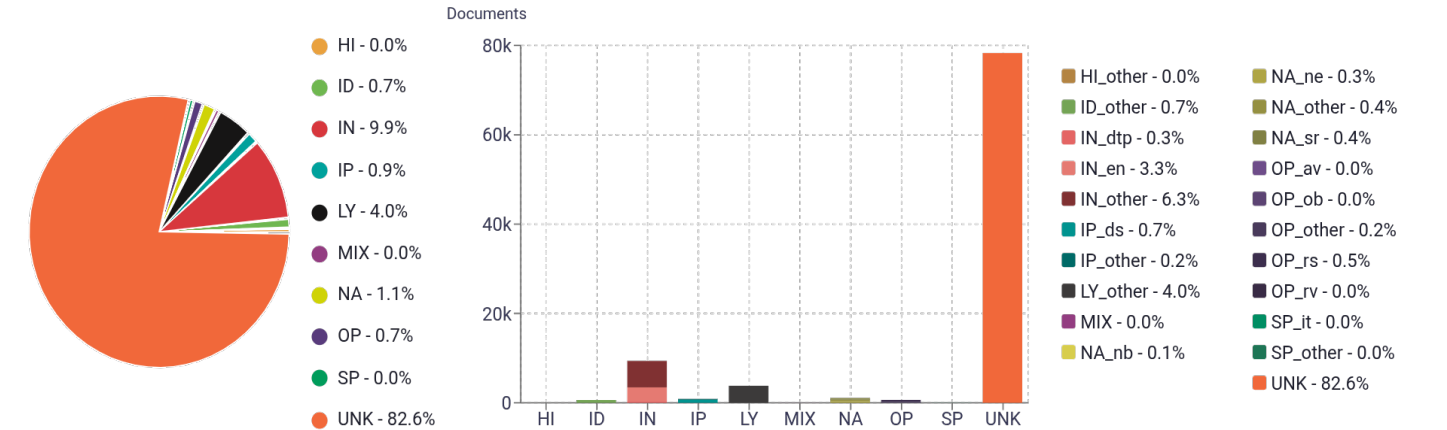
Top 10 domains

Domain	Docs	% of total
trt.net.tr	19K	20.13%
blogfa.com	11K	12.04%
axar.az	11K	11.43%
wikipedia.org	6.5K	6.90%
arzublog.com	4.9K	5.14%
baybak.com	3.4K	3.61%
ishiq.net	2.9K	3.01%
bbc.com	1.9K	2.02%
swn.af	1.8K	1.91%
inform.kz	1.3K	1.32%

Top 10 TLDs

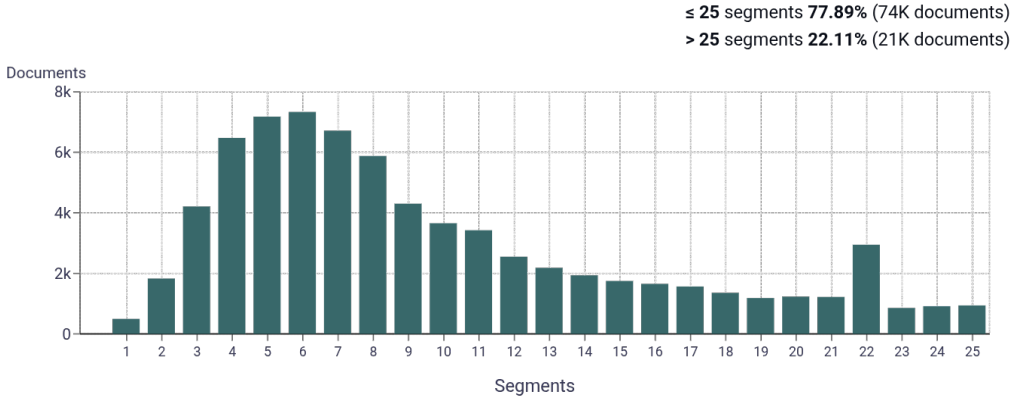
Domain	Docs	% of total
com	38K	40.23%
net.tr	19K	20.13%
az	11K	11.44%
org	8.6K	9.05%
ir	7.5K	7.87%
net	3.4K	3.57%
af	2.6K	2.78%
kz	1.3K	1.32%
se	716	0.76%
gov.af	551	0.58%

Register labels

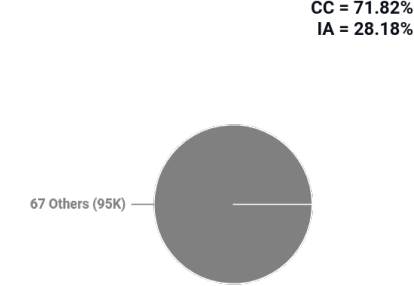


MT:58.4% | 55K Documents

Documents size (in segments) ⓘ

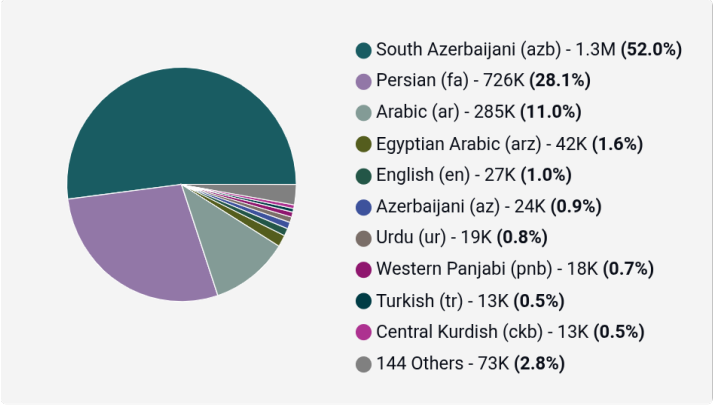


Document collections

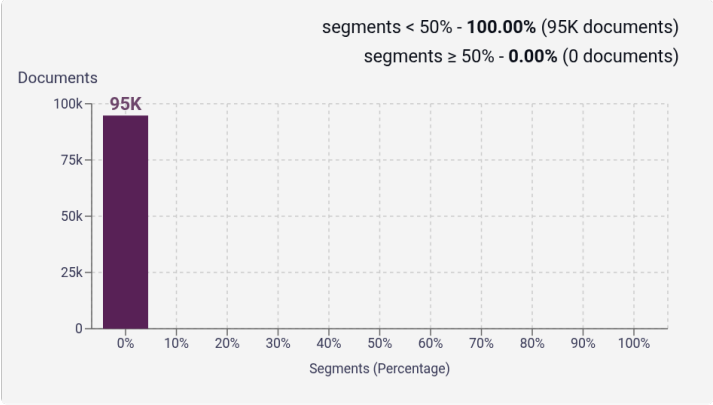


Language Distribution

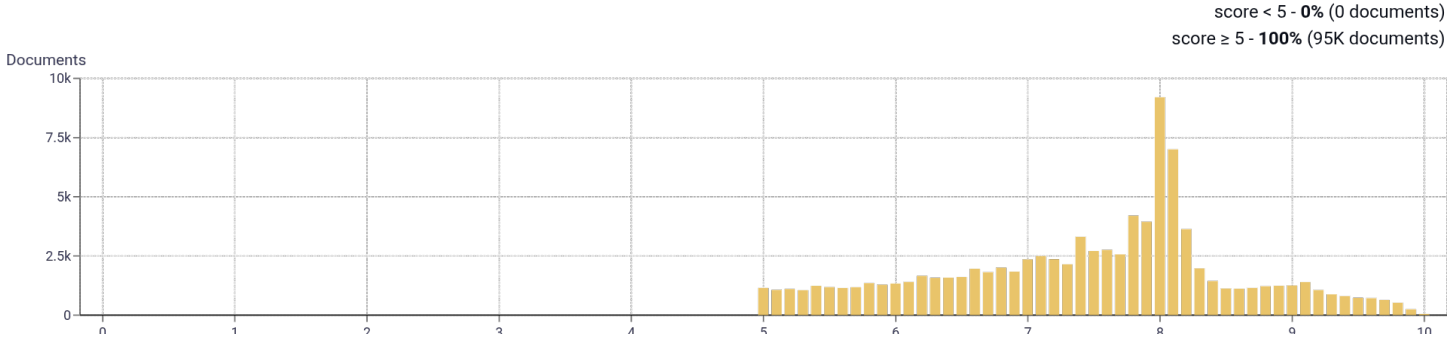
Number of segments in the Azerbaijani (az) corpus



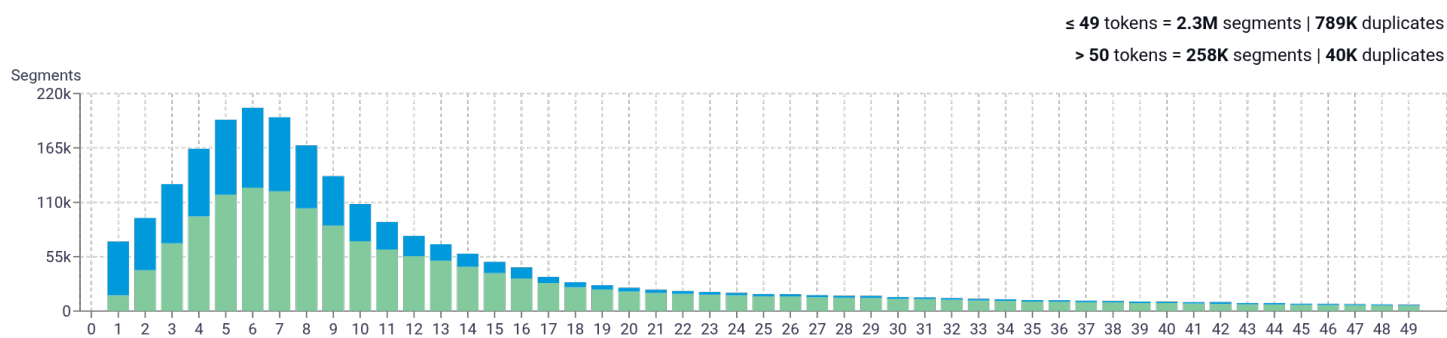
Percentage of segments in Azerbaijani (az) inside documents



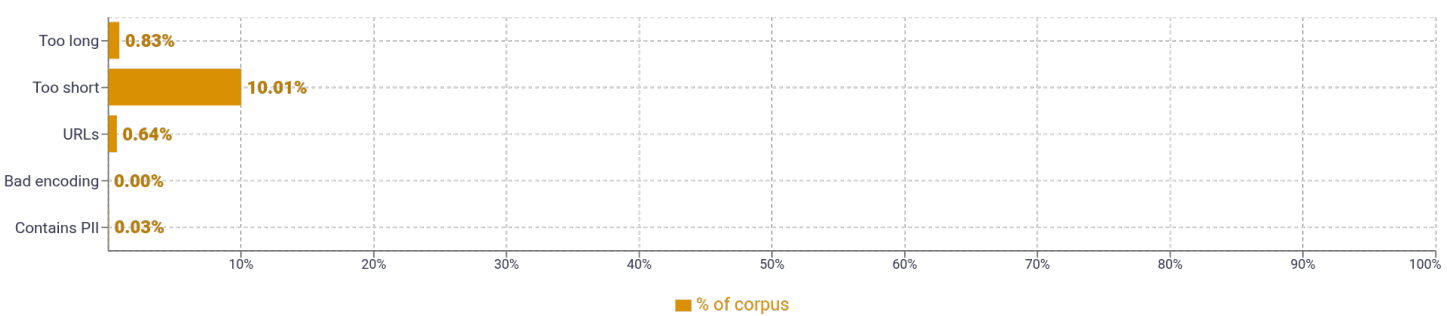
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	1,069,312 و 510,436 بیر 498,789 بو 316,664 ده 270,700 او	
2	19,354 و بو 17,487 و یا 12,904 بوگون 11,810 حی ایله 10,982 هنج بیر	
3	7,665 خبر وئیر کی 4,477 بیر میللتین سسی 4,269 آذ خبر وئیر 3,425 نوشته شده در 3,007 یونا گۆره ده	
4	4,266 آذ خبر وئیر کی 1,620 ایران ممالیکی محروسه سینده 1,562 گۆنده ر بۆلوم لر 1,124 ویکیپداسنین ایشلدنلری طرفیندن یارانمیش	
5	1,055 بای یک ساینیدان پرینت اولوب 818 ایران ممالیکی محروسه سینده کی 764 فروشگاهیان دان آلیش وئریش اندین 763 یاختی اولماقی اوجون آنیل باتیل 763 باتیل فروشگاهیان دان آلیش وئریش	

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	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				