HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-azb_Arab	9/23/2025	South Azerbaijani

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

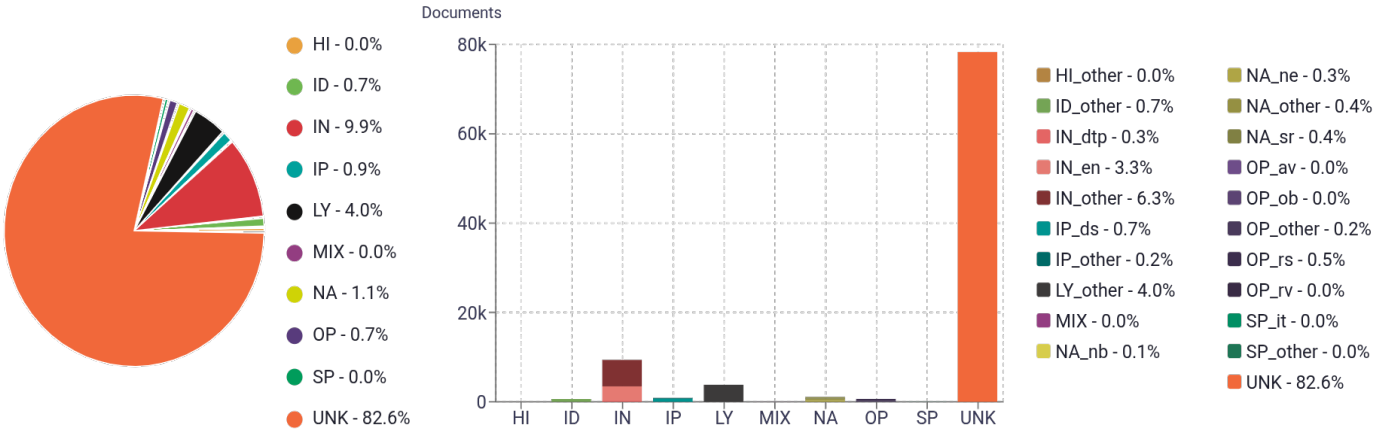
Top 10 domains

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%

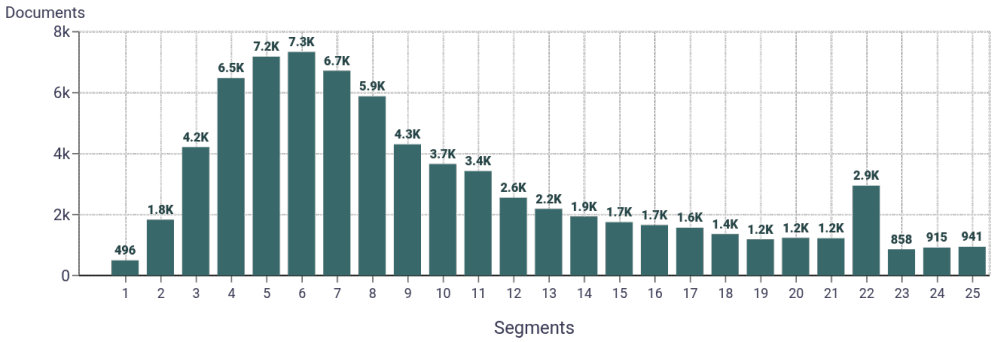
Register labels



MT:58.4% | 55K Documents

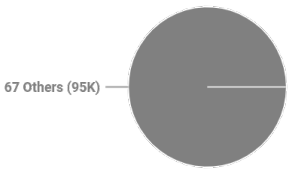
Documents size (in segments)

≤ 25 segments **77.89%** (74K documents)
> 25 segments **22.11%** (21K documents)



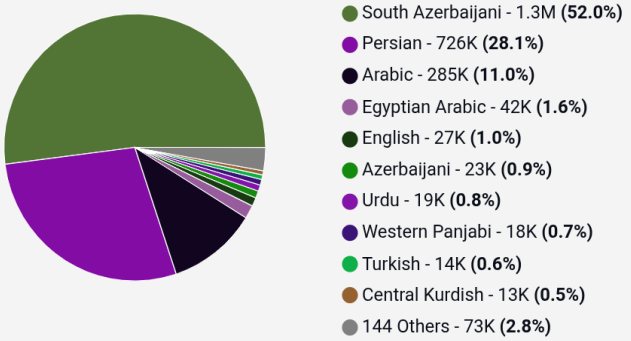
Document collections

CC = 71.82%
IA = 28.18%

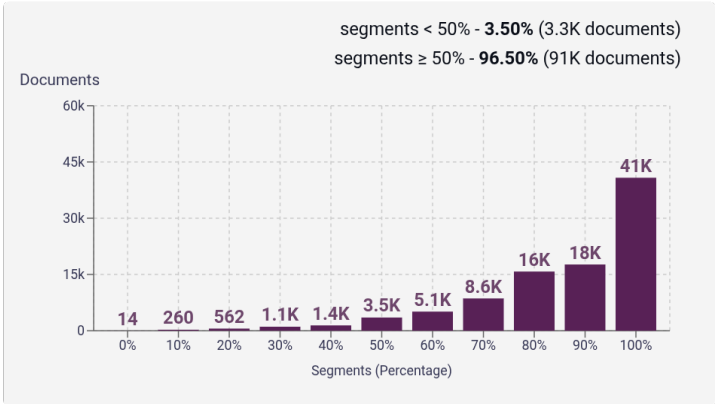


Language Distribution

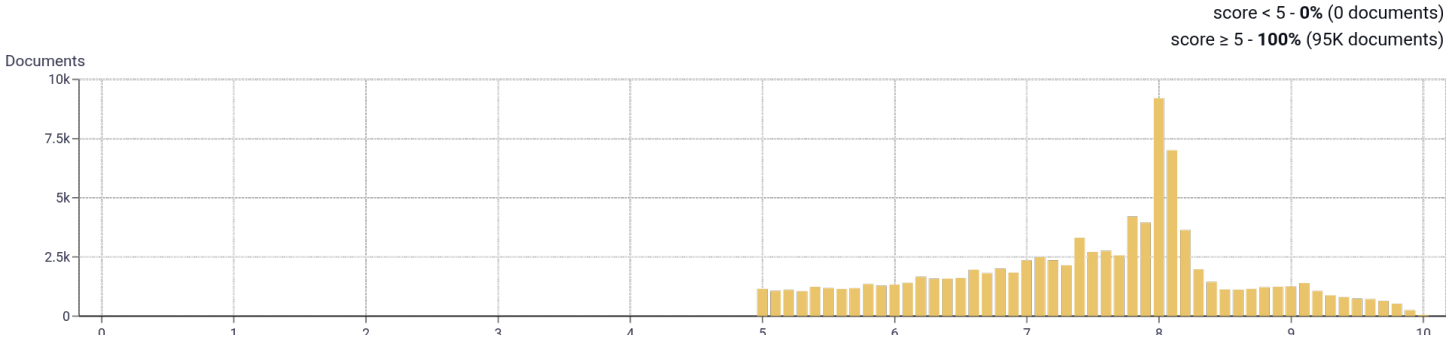
Number of segments in the South Azerbaijani corpus



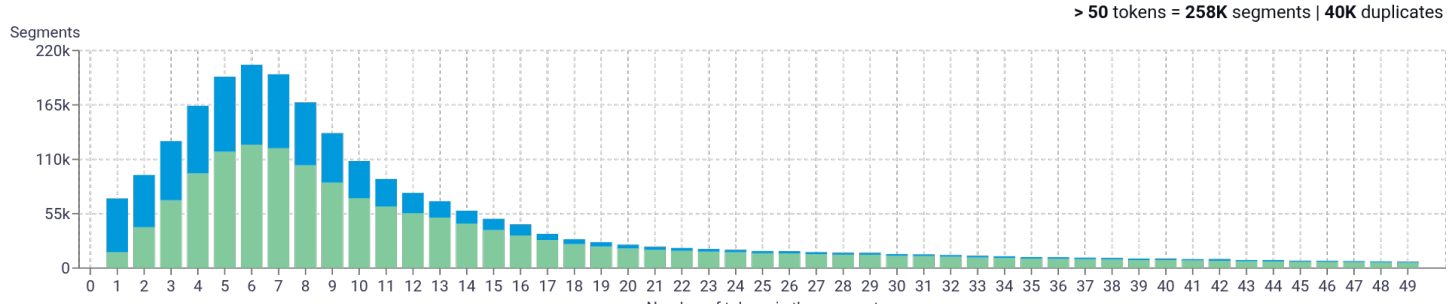
Percentage of segments in South Azerbaijani inside documents



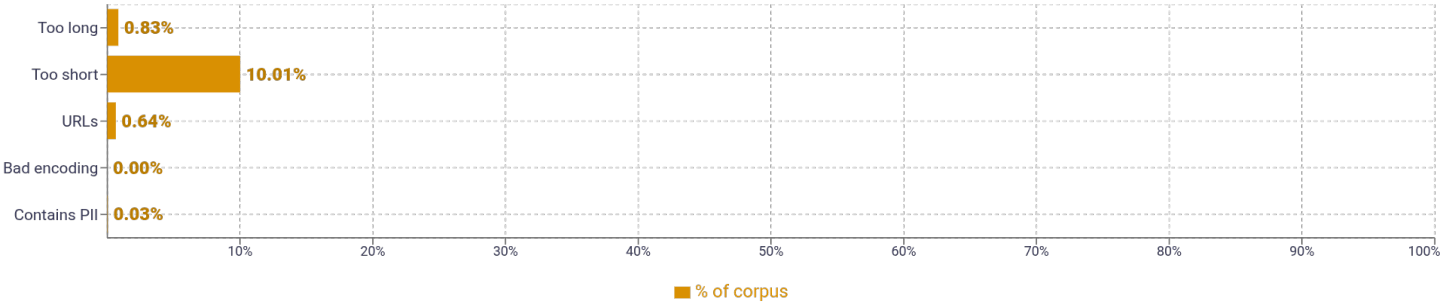
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	دا | 176,174 · ایله | 161,467 · بیر | 136,665 · نینگ | 129,179 · چوخ | 106,118
2	جی ایلده | 11,810 · read more | 9,435 · ایله باغلی | 8,795 · خبر وئریر | 8,133 · داها چوخ | 7,363
3	آذ خبر وئریر | 4,269 · ياي بک آذربایجان | 2,645 · امریکا قوشمه ایالتلری | 2,632 · ایران ممالیکی محروسه | 2,414 · خبر وئریر az 2,048
4	ایران ممالیکی محروسه سینده | 1,620 · گؤنده ر بؤلوم لر | 1,562 · ویکیپدیاسینین ایشلدنلری طرفیندن یارانمیش | 1,124 · بک سایتیندان پرینت اولوب | 1,055 · بای بک سایتیندان پرینت | 1,055
5	بای بک سایتیندان پرینت اولوب | 1,055 · فروشگاهین دان آلیش وئریش ائدین | 764 · یاخشی اولماقی اوچون آتیل باتیل | 763 · باتیل فروشگاهین دان آلیش وئریش | 763 · اولماقی اوچون آتیل باتیل فروشگاهین | 763

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				