HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-azj_Latn	9/17/2025	Azerbaijani (az)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
11,068,894	244,002,387	105,314,568 (43.16 %)	6.7B	41,023,605,815	44.01 GB

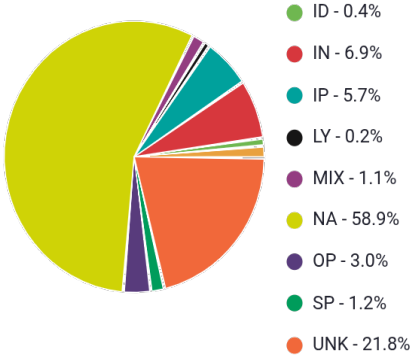
Top 10 domains

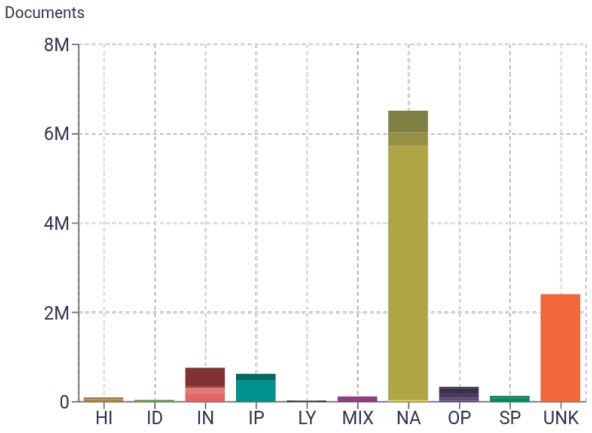
Domain	Docs	% of total
trend.az	219K	1.98%
publika.az	159K	1.44%
azadliq.org	148K	1.33%
azertag.az	141K	1.28%
report.az	122K	1.10%
stadium.az	116K	1.05%
sputnik.az	110K	1.00%
wikipedia.org	107K	0.97%
netlify.app	98K	0.89%
amerikaninsesi.org	92K	0.83%

Top 10 TLDs

Domain	Docs	% of total
az	6M	54.00%
com	2.2M	20.23%
org	687K	6.20%
info	333K	3.01%
net	248K	2.24%
gov.az	211K	1.91%
tv	117K	1.06%
app	100K	0.90%
edu.az	97K	0.87%
biz	58K	0.53%

Register labels



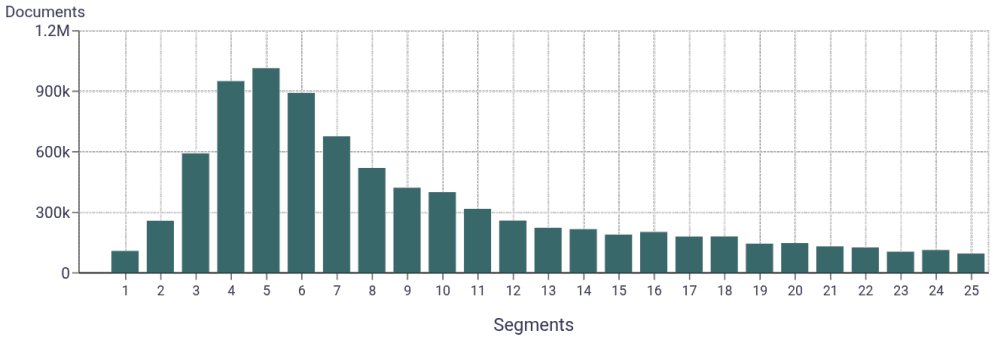


HI_other - 0.5% | MIX - 1.1%
HI_re - 0.4% | NA_nb - 0.4%
ID_other - 0.4% | NA_ne - 51.4%
IN_dtp - 1.5% | NA_other - 2.7%
IN_en - 1.3% | NA_sr - 4.4%
IN_fi - 0.0% | OP_av - 0.4%
IN_lt - 0.4% | OP_ob - 0.4%
IN_other - 3.6% | OP_other - 0.9%
IN_ra - 0.0% | OP_rs - 1.1%
IP_ds - 4.3% | OP_rv - 0.2%
IP_ed - 0.0% | SP_it - 0.8%
IP_other - 1.4% | SP_other - 0.4%
LY_other - 0.2% | UNK - 21.8%

MT:18.0% | 2M Documents

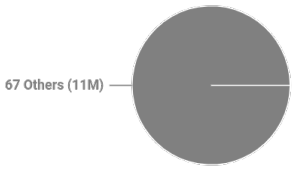
Documents size (in segments)

≤ 25 segments **76.51%** (8.5M documents)
> 25 segments **23.49%** (2.6M documents)



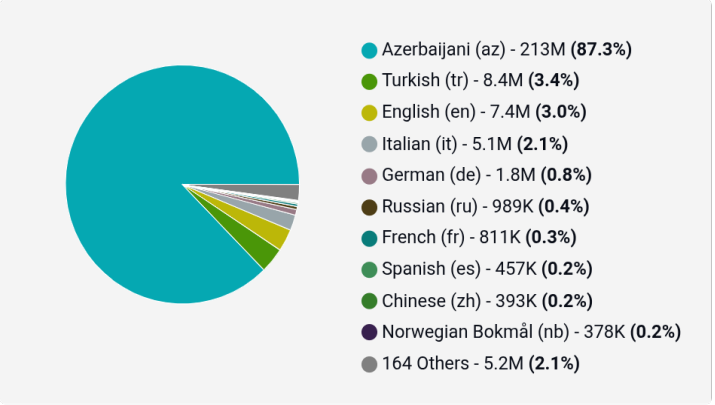
Document collections

CC = 87.62%
IA = 12.38%

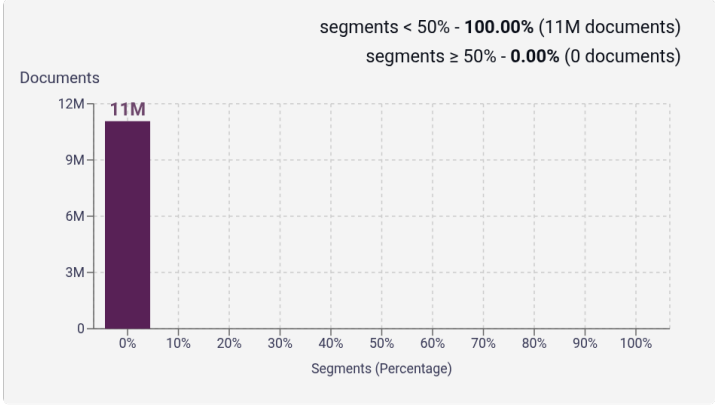


Language Distribution

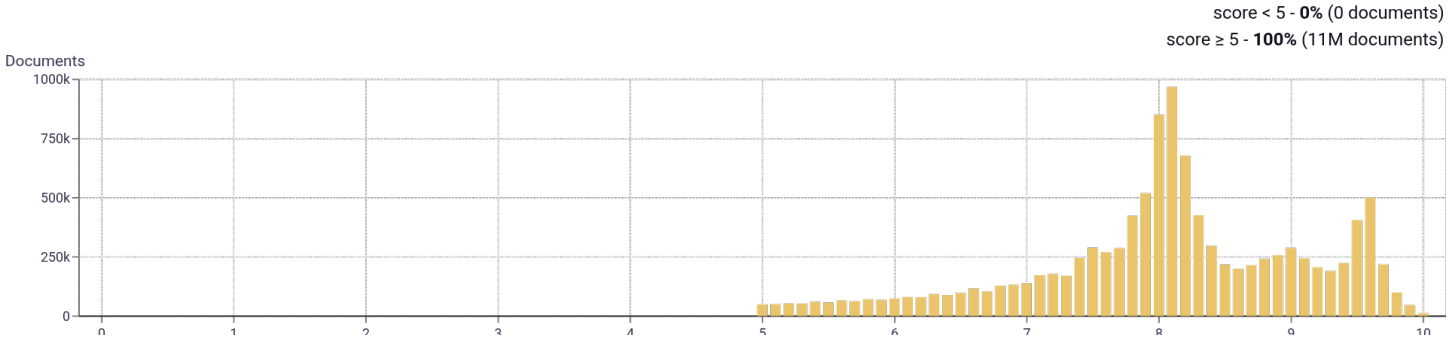
Number of segments in the Azerbaijani (az) corpus



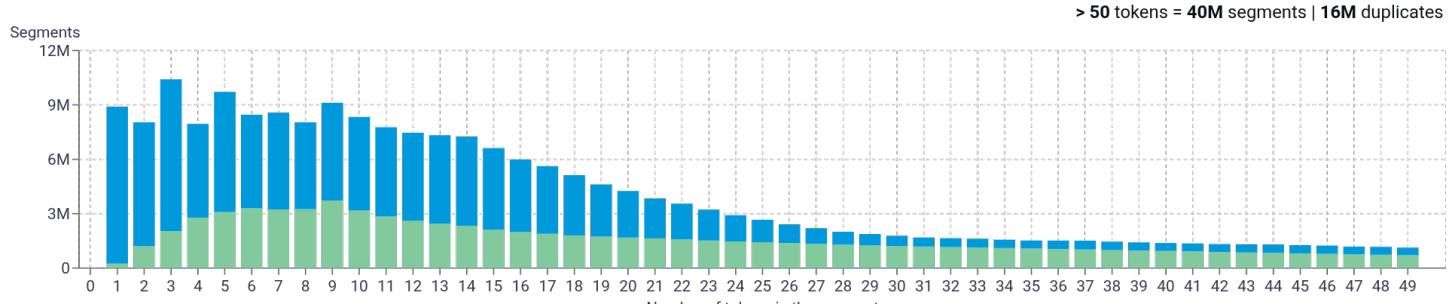
Percentage of segments in Azerbaijani (az) inside documents



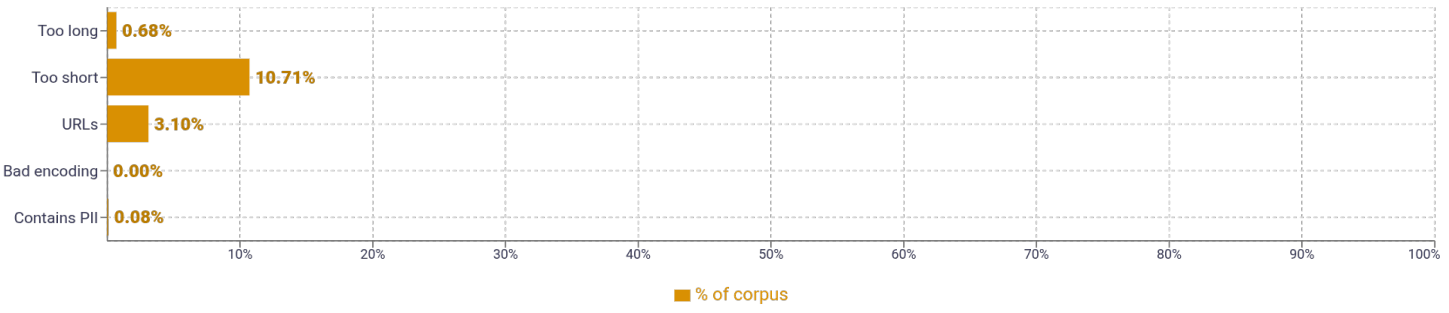
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	mostbet | 20,325,732 · mərc | 20,040,089 · azərbaycan | 17,244,761 · bilərsiniz | 16,577,783 · pin | 15,279,013
2	pin up | 7,741,086 · edə bilərsiniz | 5,178,530 · up casino | 4,812,949 · xəbər verir | 3,583,786 · imkan verir | 2,336,943
3	pin up casino | 2,570,658 · əldə edə bilərsiniz | 960,449 · istinadən xəbər verir | 832,706 · mərc edə bilərsiniz | 646,933 · etməyə imkan verir | 630,817
4	dəstək xidməti ilə əlaqə | 435,167 · pin up casino online | 264,280 · up on line casino | 199,428 · azərbaycan respublikasının prezidenti ilham | 179,932 · android və ya ios | 161,892
5	dəstək xidməti ilə əlaqə saxlaya | 149,391 · kazino azerbaycan ən yaxşı bukmeyker | 123,829 · azerbaycan ən yaxşı bukmeyker rəsmi | 122,791 · əmək və əhalinin sosial müdafiəsi | 121,405 · azərbaycan respublikasının prezidenti ilham əliyev | 115,830

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				