HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
azb_Arab.jsonl.tsv	9/27/2024	South Azerbaijani (azb)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
66,112	2,389,200	1,055,552 (44.18 %)	49M	257,866,139	446.12 MB

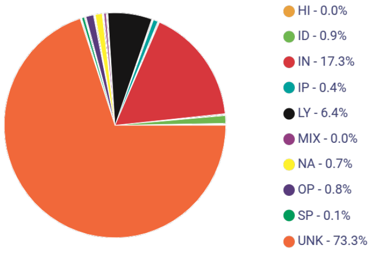
Top 10 domains

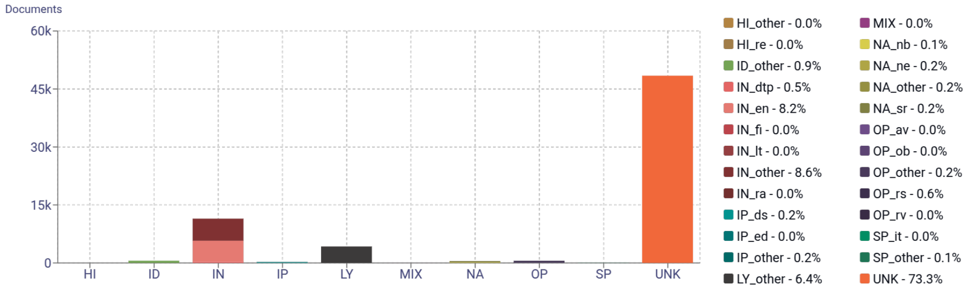
Domain	Docs	% of total
blogfa.com	14K	21.68%
wikipedia.org	11K	16.86%
axar.az	4.7K	7.09%
trt.net.tr	4.5K	6.80%
arzublog.com	2.9K	4.45%
ishiq.net	2.9K	4.35%
baybak.com	2.1K	3.12%
bilimsesi.com	1.5K	2.23%
blogsky.com	1.4K	2.12%
mihanblog.com	1.2K	1.88%

Top 10 TLDs

Domain	Docs	% of total
com	33K	50.20%
org	13K	19.92%
ir	5K	7.60%
az	4.7K	7.09%
net.tr	4.5K	6.80%
net	3.3K	4.96%
info	380	0.57%
biz	271	0.41%
se	253	0.38%
ca	209	0.32%

Register labels

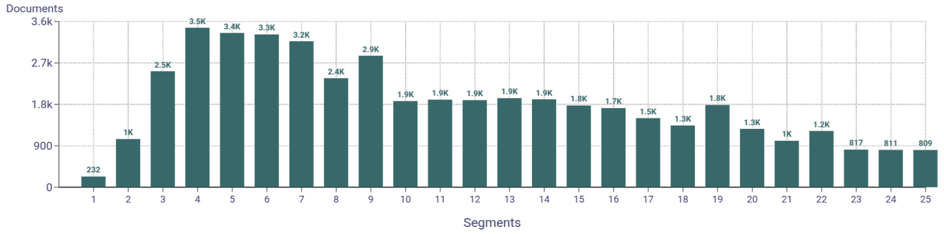




🤖 **MT**:43.3% | 29K Documents

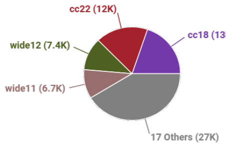
Documents size (in segments)

<= 25 segments **69.41%** (46K documents)
> 25 segments **30.59%** (20K documents)



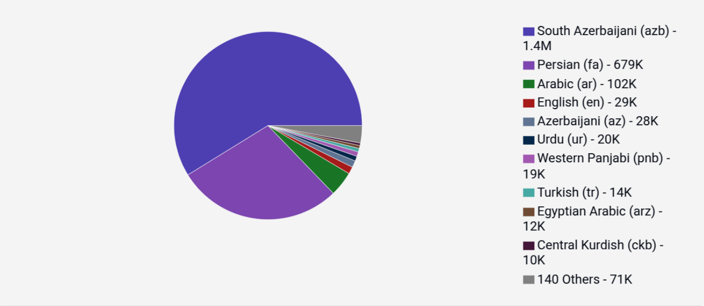
Documents by collection

CC = 48.77%
IA = 51.23%

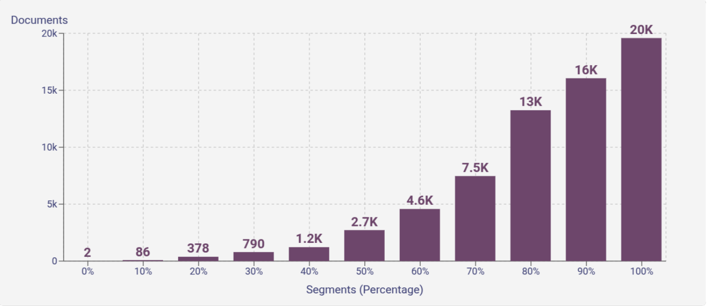


Language Distribution

Number of segments in the South Azerbaijani (azb) corpus

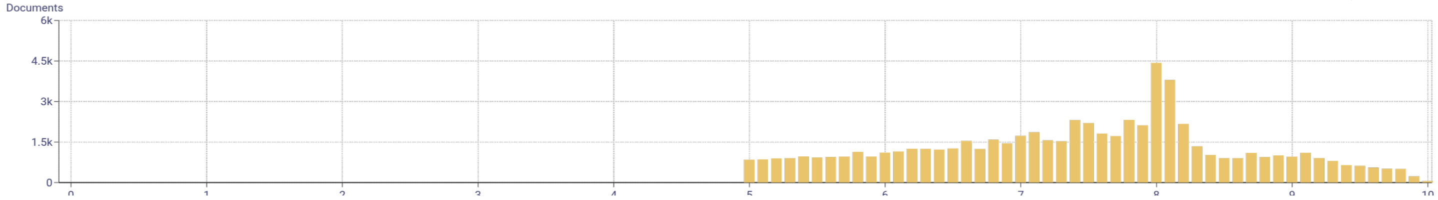


Percentage of segments in South Azerbaijani (azb) inside documents



Distribution of documents by document score

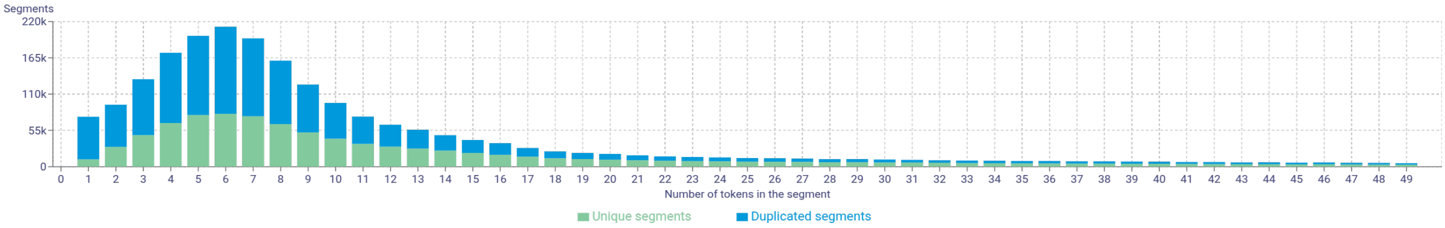
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (66K documents)



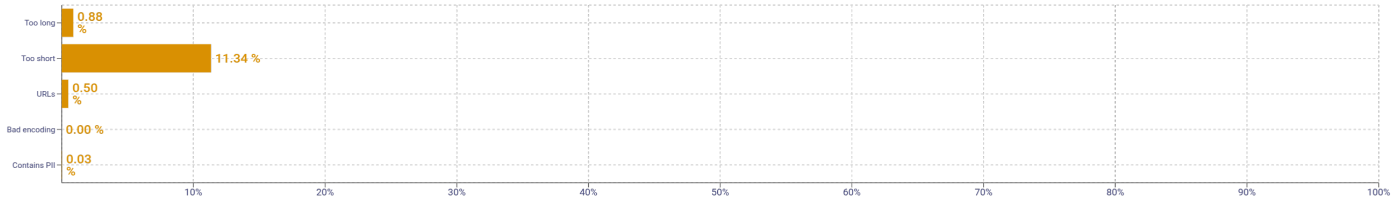
Segment length distribution by token

≤ **49** tokens = **917K** segments | **1.2M** duplicates

> **50** tokens = **241K** segments | **102K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	172533 | ایله 156791 | دا 109314 | کیمی 98960 | چوخ 92750 | دیر
2	14647 | جی ایلده read more | 9442 6535 | بونا گؤره 6317 | داها چوخ 5500 | داها آرتیق
3	3082 | گؤره داها آرتیق 3079 | داها آرتیق بیلگیلر 3078 | آرتیق بیلگیلر تاپابیلرسینیز 2740 | ایشلدنلری طرفیندن یارانمیش 2718 | ویکیپدیاسینین ایشلدنلری طرفیندن
4	3078 | گؤره داها آرتیق بیلگیلر 3078 | داها آرتیق بیلگیلر تاپابیلرسینیز 2718 | ویکیپدیاسینین ایشلدنلری طرفیندن یارانمیش 2099 | اینگیلیسجه ویکیپدیاسینین ایشلدنلری طرفیندن 1537 | ایران ممالیکی محروسه سینده
5	3078 | گؤره داها آرتیق بیلگیلر تاپابیلرسینیز 2099 | اینگیلیسجه ویکیپدیاسینین ایشلدنلری طرفیندن یارانمیش 561 | ییشیک موبایل وبی ایله د 496 | ی تیمی ترکیبینده چیخیش ائدیب 414 | ایلام ایلام ایلام ایلام ایلام

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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av