HPLT Analytics report

HPLT Analytics

General overview

Corpus	Date	Language
hplt-v3-ace_Arab	9/16/2025	Achinese

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
7	72	69 (95.83 %)	3.2K	13,796	23.64 KB

Top 10 domains

Domain	Docs	% of total
wikimedia.org	2	28.57%
utusanmelayu.co...	2	28.57%
issuu.com	1	14.29%
blogspot.com	1	14.29%
bayardservicewe...	1	14.29%

Top 10 TLDs

Domain	Docs	% of total
com	3	42.86%
org	2	28.57%
com.my	2	28.57%

Documents size (in segments) ⓘ
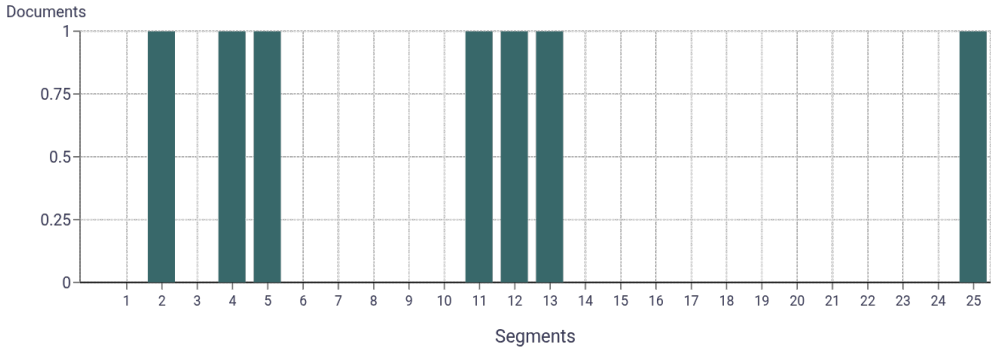
≤ **25** segments **100%** (7 documents)
> **25** segments **0%** (0 documents)

Document collections

CC = 100.00%
IA = 0.00%

CC-MAIN-2024-51 (1)
CC-MAIN-2017-04
C-MAIN-2024-22 (1)
CC-MAIN-201
CC-MAIN-2018-43 (1)
CC-MAIN-2018-39 (1)
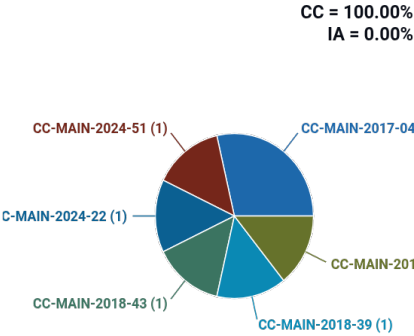
Language Distribution

Number of segments in the Achinese corpus

- Arabic - 34 **(47.2%)**
- Persian - 13 **(18.1%)**
- Mazanderani - 6 **(8.3%)**
- English - 5 **(6.9%)**
- French - 3 **(4.2%)**
- Urdu - 2 **(2.8%)**
- Uyghur - 2 **(2.8%)**
- Malay - 2 **(2.8%)**
- South Azerbaijani - 2 **(2.8%)**
- Sindhi - 1 **(1.4%)**
- 2 Others - 2 **(2.8%)**

*Achinese identification might be inaccurate because it is not supported by FastSpell.
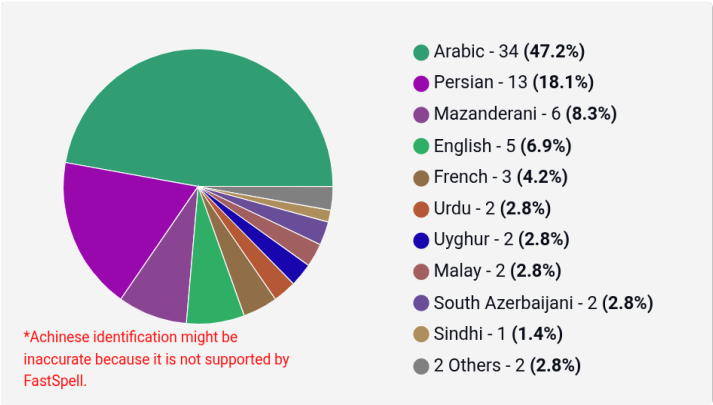
Percentage of segments in Achinese inside documents

segments < 50% - **42.86%** (3 documents)
segments ≥ 50% - **57.14%** (4 documents)

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (7 documents)
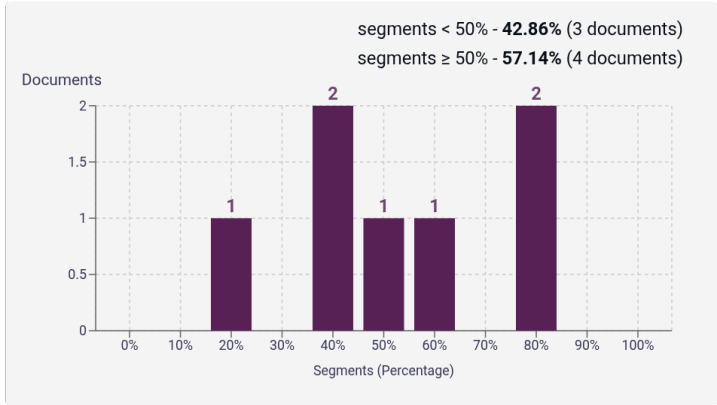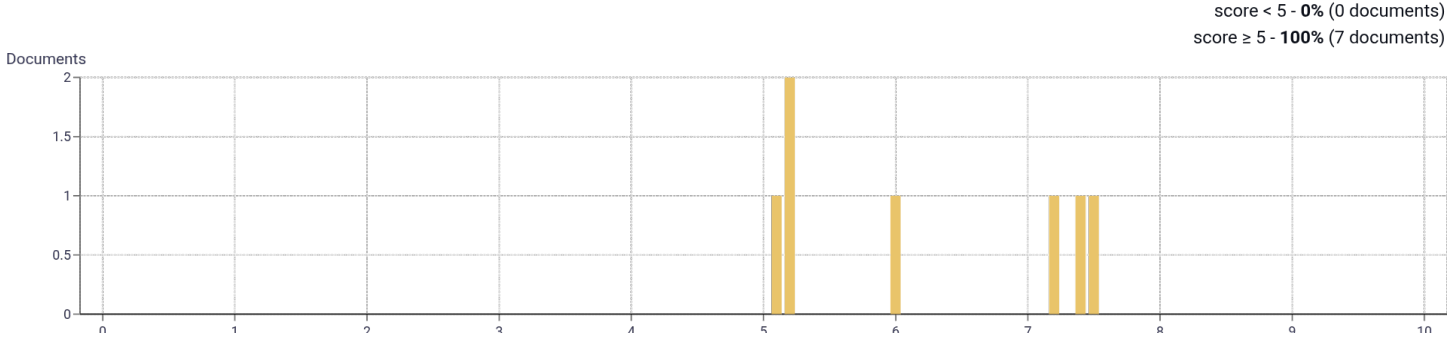
Segment length distribution by token

≤ 49 tokens = **52** segments | **2** duplicates

> 50 tokens = **20** segments | **1** duplicates

Segments (y-axis: 0, 3, 6, 9, 12)

Number of tokens in the segment (x-axis: 0–49)

Segment noise distribution

Category	% of corpus
Too long	0.00%
Too short	15.28%
URLs	5.56%
Bad encoding	0.00%
Contains PII	5.56%

■ % of corpus
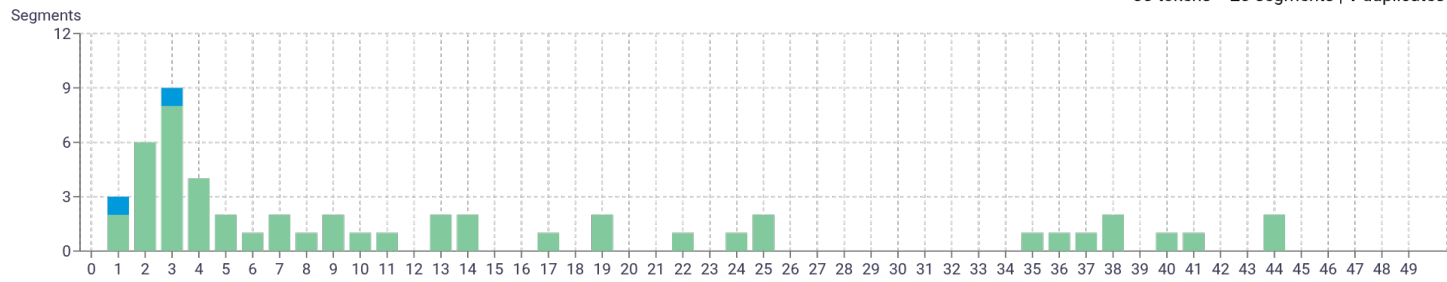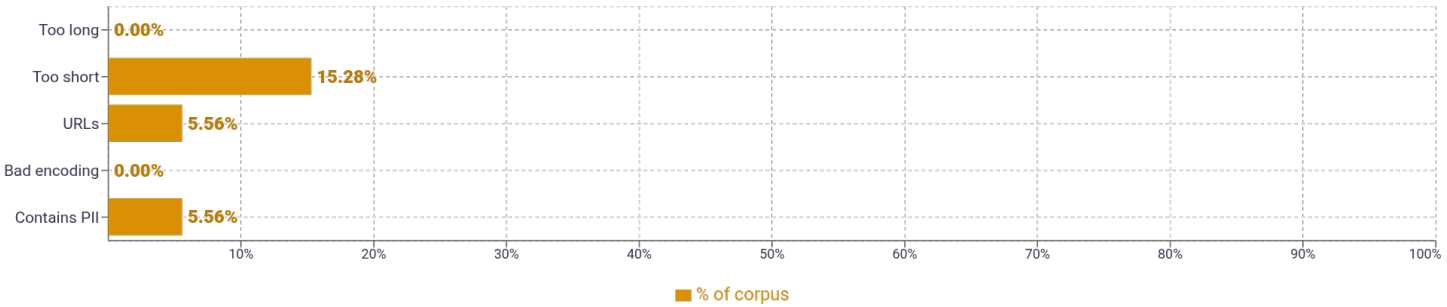
Frequent n-grams

SIZE	N-GRAMS
1	ا | 218 و | 54 ن | 53 ل | 39 ب | 35
2	ا ا | 30 اوڠن كومريݢ | 9 ل ا | 7 كومريݢ اولو | 6 ڤورتل اين | 6
3	اين اكن داڤت | 4 دي كبوڤاتين اوڠن | 4 كبوڤاتين اوڠن كومريݢ | 4 ڤورتل اين اكن | 4 اوڠن كومريݢ اولو | 6
4	ڤڠڬونا'ن توليسن جاوي سڤنوهن | 2 كوجودن ڤورتل اين اكن | 2 كبوڤاتين اوڠن كومريݢ اولو | 3 دي كبوڤاتين اوڠن كومريݢ | 4 ڤورتل اين اكن داڤت | 4
5	دي كبوڤاتين اوڠن كومريݢ اولو | 3 كوجودن ڤورتل اين اكن داڤت | 2 ڤڠڬونا داڤت ممنڤعتكن اف يݢ | 2 ڤڠڬونا باݢي ڤريڤكت عمور يݢ | 2 ڤڠتاهوان ڤڠڬونا باݢي ڤريڤكت عمور | 2

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				