HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
ace_Arab.jsonl.tsv	12/3/2024	Acehnese (ace)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
16	117	93 (79.49 %)	12K	49,626	84.34 KB

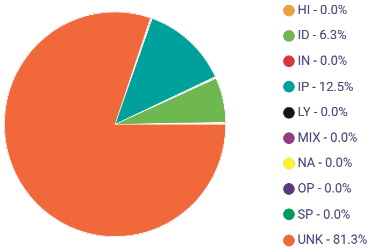
Top 10 domains

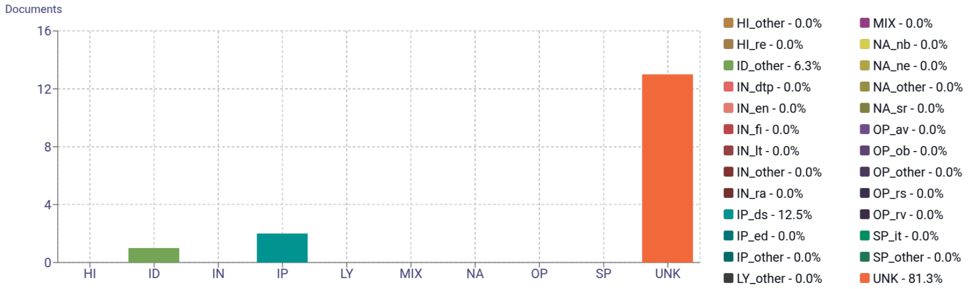
Domain	Docs	% of total
iran-eng.ir	4	25.00%
blogspot.com	2	12.50%
bayardservicewe...	1	6.25%
utusanmelayu.co...	1	6.25%
office-converte...	1	6.25%
salemalanzi.sa	1	6.25%
onlineyoutube.com	1	6.25%
all.biz	1	6.25%
blogspot.it	1	6.25%
blogspot.my	1	6.25%

Top 10 TLDs

Domain	Docs	% of total
com	6	37.50%
ir	4	25.00%
com.my	1	6.25%
sa	1	6.25%
biz	1	6.25%
it	1	6.25%
my	1	6.25%
net	1	6.25%

Register labels

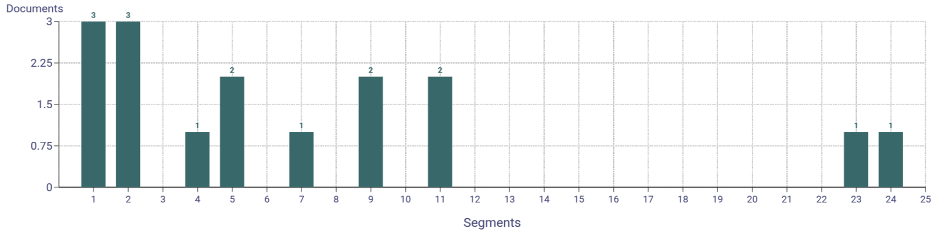




🤖 **MT**:56.3% | 9 Documents

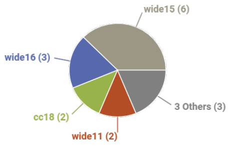
Documents size (in segments)

<= 25 segments **100%** (16 documents)
> 25 segments **0%** (0 documents)



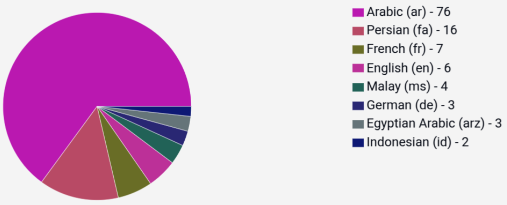
Documents by collection

CC = 18.75%
IA = 81.25%



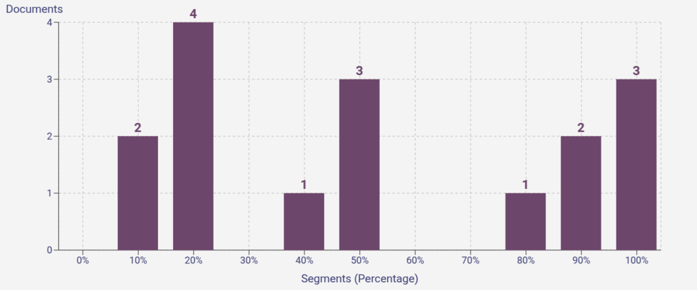
Language Distribution

Number of segments in the Acehnese (ace) corpus



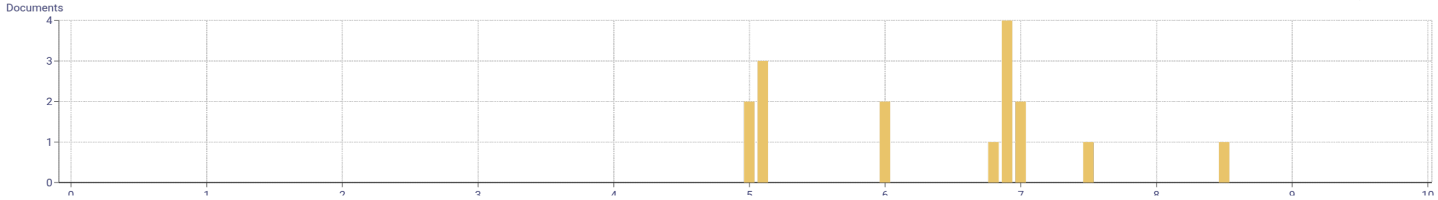
*Acehnese (ace) identification might be inaccurate because it is not supported by FastSpell

Percentage of segments in Acehnese (ace) inside documents



Distribution of documents by document score

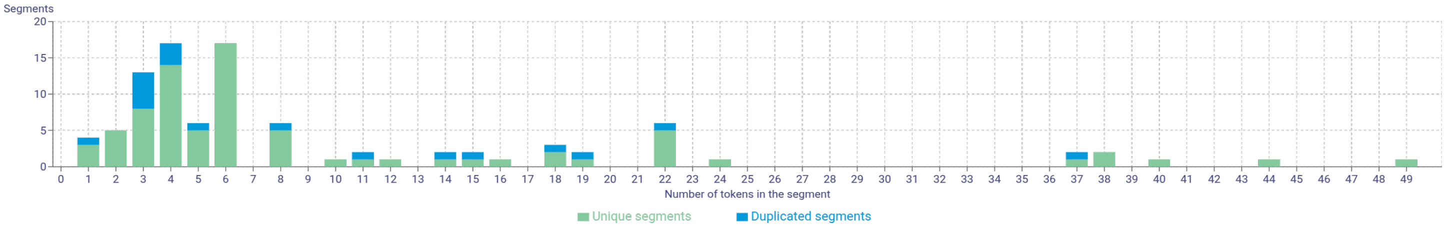
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (16 documents)



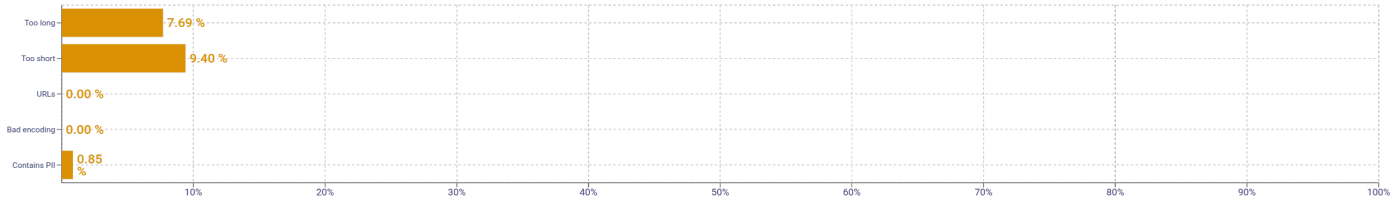
Segment length distribution by token

≤ **49** tokens = **78** segments | **18** duplicates

> **50** tokens = **21** segments | **6** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	1683 | کن · 1680 | مدون · 1530 | کجاس · 1530 | خونتون · 756 | اوک
2	1676 | مدون کن · 1673 | کن مدون · 1530 | خونتون کجاس · 755 | اوک اوک · ل ppm | 336
3	1670 | کن مدون کن · 1666 | مدون کن مدون · 754 | اوک اوک اوک · 56 | هیونداي جینیسیس کوبیه · 24 | چت ازیتازیتا چت
4	1665 | مدون کن مدون کن · 1660 | کن مدون کن مدون · 753 | اوک اوک اوک اوک · 13 | سعر هیونداي جینیسیس کوبیه · 8 | قورتل این اکن داقت
5	1659 | کن مدون کن مدون کن · 1655 | مدون کن مدون کن مدون · 752 | اوک اوک اوک اوک اوک · 6 | مدونکن مدون کن مدون کن · 6 | مدون کن مدونکن مدون کن

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