HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
ace_Latn.jsonl.tsv	10/2/2024	Acehnese (ace)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
12,930	206,187	107,160 (51.97 %)	9.7M	50,639,739	49.23 MB

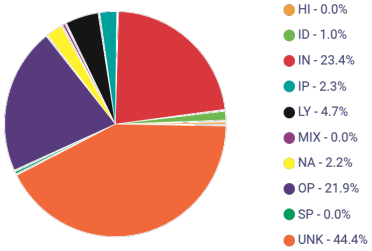
Top 10 domains

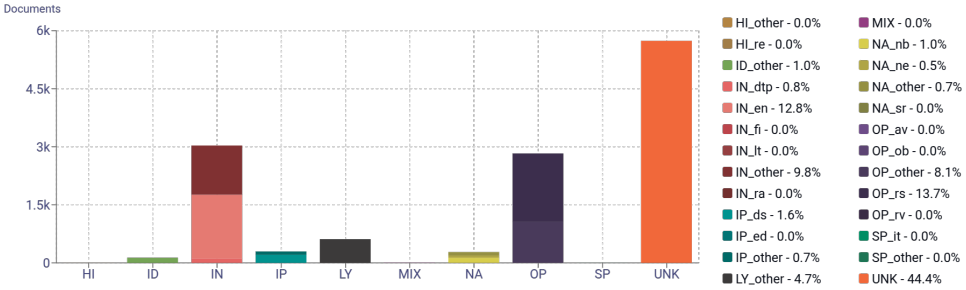
Domain	Docs	% of total
bible.is	4.3K	33.40%
wikipedia.org	2.8K	21.99%
wordproject.org	1.2K	9.15%
petalokasi.org	362	2.80%
blogspot.com	338	2.61%
wordpress.com	209	1.62%
azlyricdb.com	180	1.39%
kodeposindo.xyz	166	1.28%
fanskpop.com	148	1.14%
jerseysu.com	85	0.66%

Top 10 TLDs

Domain	Docs	% of total
org	4.6K	35.52%
is	4.3K	33.40%
com	2.6K	19.81%
xyz	172	1.33%
net	166	1.28%
com.br	93	0.72%
ru	87	0.67%
mobi	71	0.55%
co.id	65	0.50%
tv	50	0.39%

Register labels

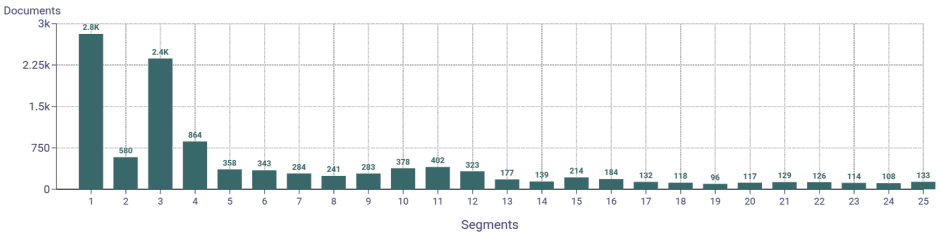




MT:5.7% | 731 Documents

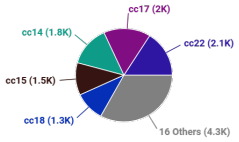
Documents size (in segments)

<= 25 segments **85.25%** (11K documents)
> 25 segments **14.75%** (1.9K documents)



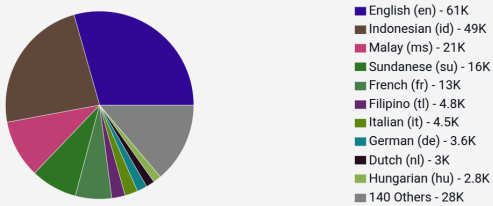
Documents by collection

CC = 73.23%
IA = 26.77%



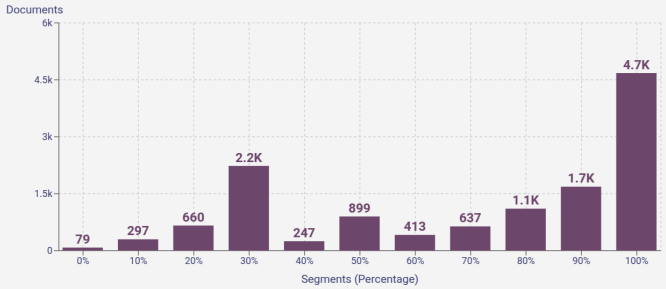
Language Distribution

Number of segments in the Acehnese (ace) corpus



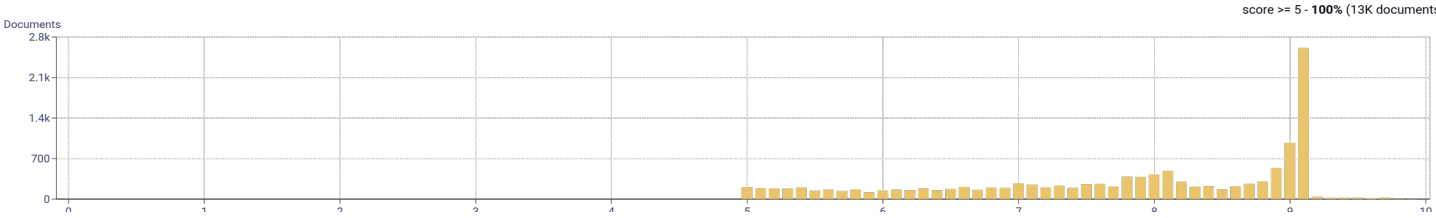
*Acehnese (ace) identification might be inaccurate because it is not supported by FastSpell

Percentage of segments in Acehnese (ace) inside documents



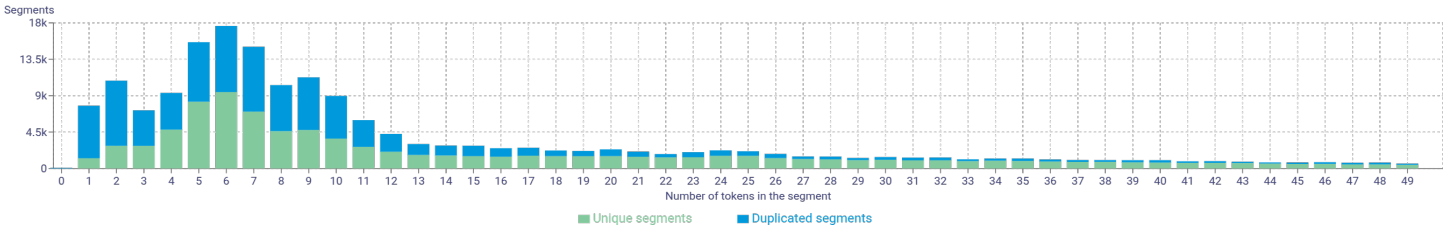
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (13K documents)

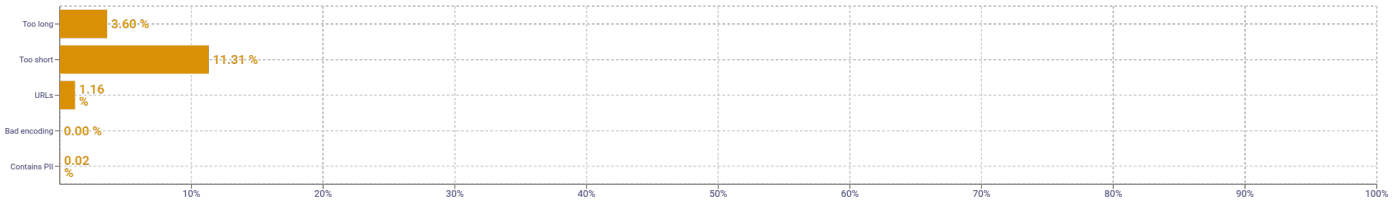


Segment length distribution by token

≤ 49 tokens = **94K** segments | **87K** duplicates
> 50 tokens = **25K** segments | **12K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	blah | 290258 · bak | 100221 · ureuëng | 74691 · gata | 72273 · ulôn | 63060
2	blah blah | 258752 · teu allah | 10208 · zis gas | 9104 · meunan cit | 8720 · blahblah blah | 8433
3	blah blah blah | 236064 · blah blahblah blah | 7693 · blahblah blah blah | 6509 · blah blah blahblah | 6088 · taiq zis gas | 5663
4	blah blah blah blah | 217503 · blahblah blah blah blah | 6054 · blah blah blahblah blah | 5964 · blah blahblah blah blah | 5821 · blah blah blah blahblah | 5639
5	blah blah blah blah blah | 204731 · blah blah blahblah blah blah | 5730 · blah blah blah blahblah blah | 5533 · blah blahblah blah blah blah | 5371 · blahblah blah blah blah blah | 4115

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