HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
taq_Latn.jsonl.tsv	11/28/2024	Tamasheq (taq)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,747	13,884	7,842 (56.48 %)	2.3M	9.41 MB	8,833,192

Top 10 domains

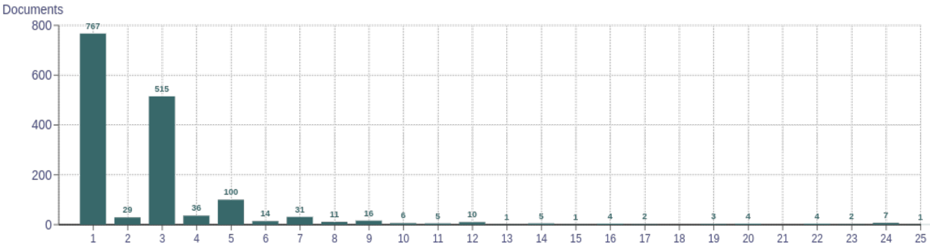
Domain	Docs	% of total
bible.is	1.2K	69.83
worddetector.com	79	4.52
newchristianbiblestudy.org	60	3.43
ebible.org	53	3.03
biblehub.com	18	1.03
vanuatubibles.org	16	0.92
tuspalabras.com	14	0.80
blogspot.com	12	0.69
case.edu	9	0.52
omniglot.com	9	0.52

Top 10 TLDs

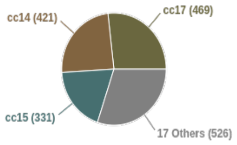
Domain	Docs	% of total
is	1.2K	69.83
com	250	14.31
org	162	9.27
net	33	1.89
edu	9	0.52
de	8	0.46
co	8	0.46
es	7	0.40
com.pl	4	0.23
me	4	0.23

Documents size (in segments)

<= 25 segments **90.1%** (1.6K documents)
> 25 segments **9.9%** (173 documents)

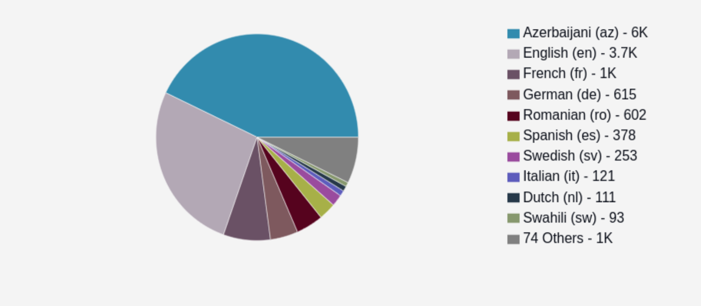


Documents by collection



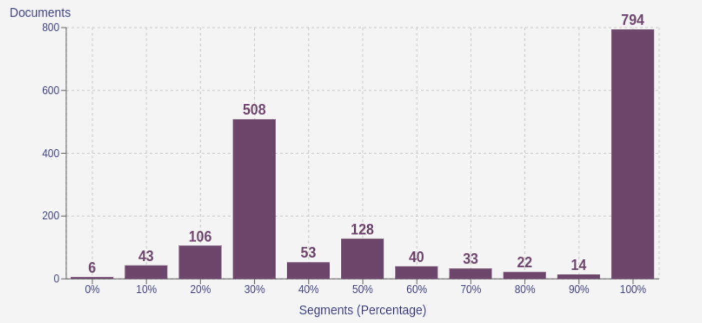
Language Distribution

Number of segments



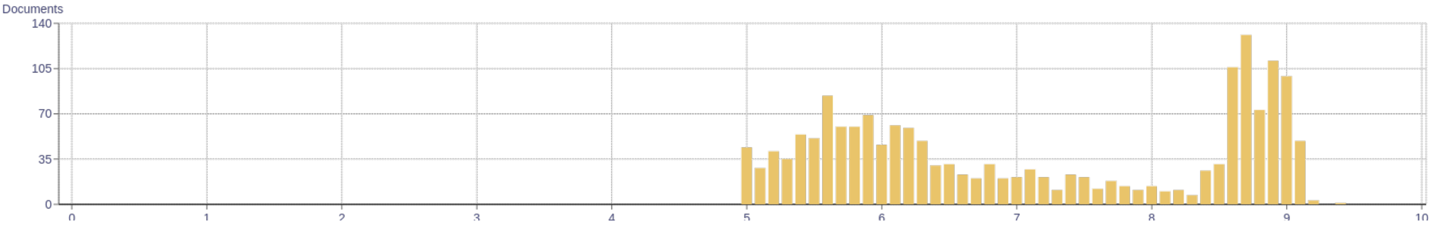
*Tamasheq (taq) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Tamasheq (taq) inside documents



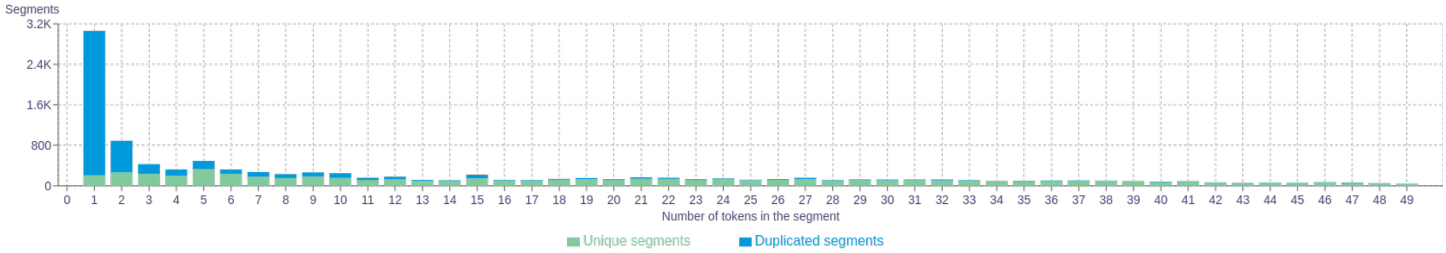
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (1.7K documents)

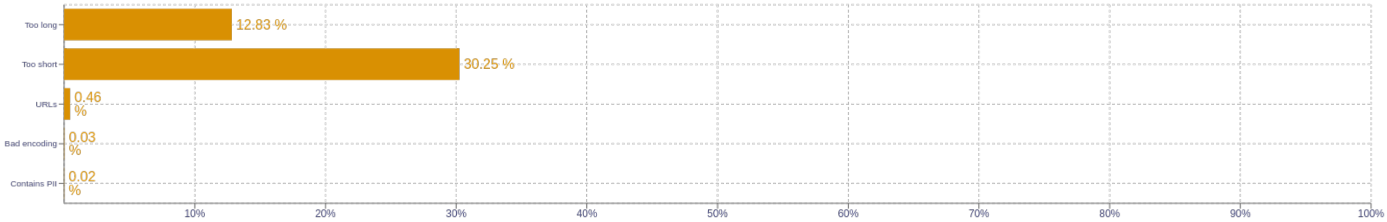


Segment length distribution by token

<= 49 tokens = **5.9K** segments | **5K** duplicates
> 50 tokens = **3K** segments | **1.1K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	n | 29099 s | 23680 dăɣ | 22105 ahay | 21633 sə | 20455
2	win win | 5792 ata awan | 4379 ɗo ahay | 3716 ata nà | 3207 ɗo sə | 3019
3	win win win | 5787 mer su way | 2029 anà ɗo ahay | 1261 asd asd asd | 1200 sdjflk asdfkas df | 990
4	win win win win | 5782 laskdj flka sdjflk asdfkas | 990 flka sdjflk asdfkas df | 990 d faslkdfj asdlkfj as | 990 asdf aasd f asdklfj | 990
5	win win win win win | 5777 laskdj flka sdjflk asdfkas df | 990 asdf aasd f asdklfj as | 990 flsadfasdf asdf aasd f asdklfj | 986 aksd flsadfasdf asdf aasd f | 986

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