HPLT Analytics report

HPLT Analytics

General overview

Corpus	Analytics date	Language
taq_Latn.jsonl.tsv	11/5/2024	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

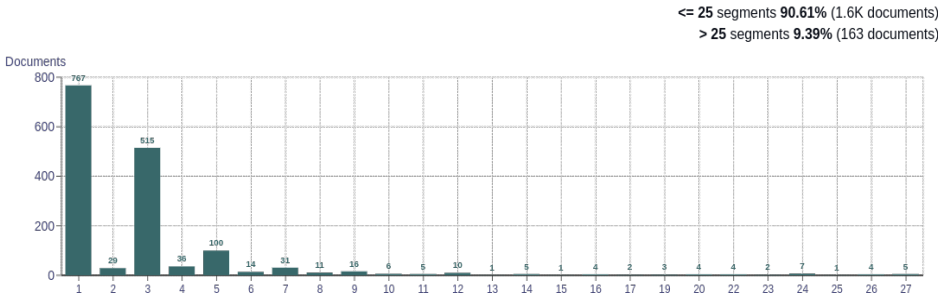### Top 10 domains

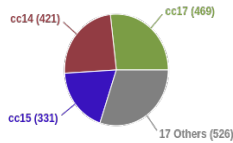| Domain	Docs	% of total
bible.is	1.2K	69.83
worddetector.com	79	4.52
newchristianbiblestudy.org	60	3.43
ebible.org	53	3.03
biblehub.com	18	1.03
vanuatubibles.org	16	0.92
tuspalabras.com	14	0.80
blogspot.com	12	0.69
case.edu	9	0.52
omniglot.com	9	0.52

Top 10 TLDs

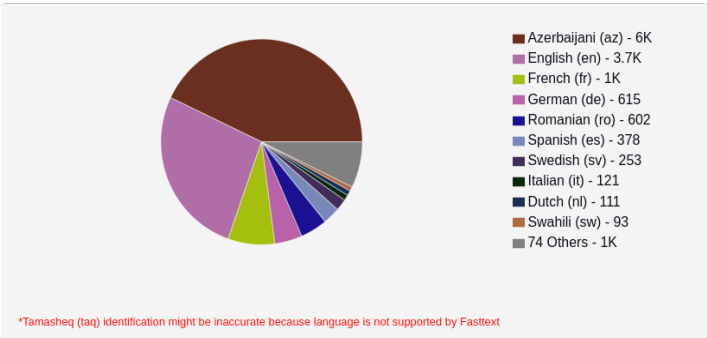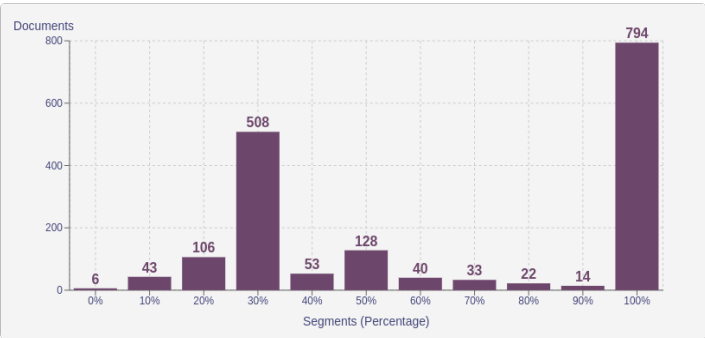| Domain	Docs	% of total
is	1.2K	69.83
com	250	14.31
org	162	9.27
net	33	1.89
edu	9	0.52
de	8	0.46
co	8	0.46
es	7	0.40
com.pl	4	0.23
me	4	0.23

Documents size (in segments)

<= 25 segments **90.61%** (1.6K documents)
> 25 segments **9.39%** (163 documents)

Documents by collection

cc14 (421) · cc17 (469) · cc15 (331) · 17 Others (526)

Language Distribution

Number of segments

- Azerbaijani (az) - 6K
- English (en) - 3.7K
- French (fr) - 1K
- German (de) - 615
- Romanian (ro) - 602
- Spanish (es) - 378
- Swedish (sv) - 253
- Italian (it) - 121
- Dutch (nl) - 111
- Swahili (sw) - 93
- 74 Others - 1K

*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

Unique segments · Duplicated segments

Segment noise distribution

Too long	0.00 %
Too short	30.14 %
URLs	0.30 %
Bad encoding	0.03 %
Contains PII	0.02 %
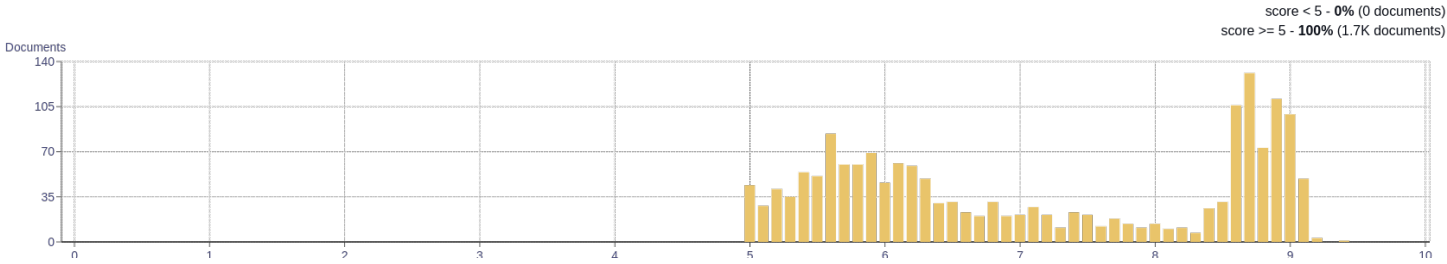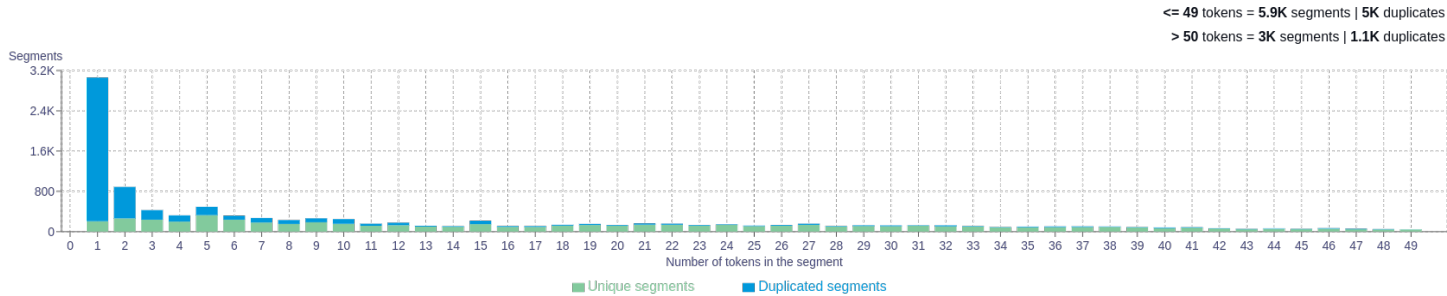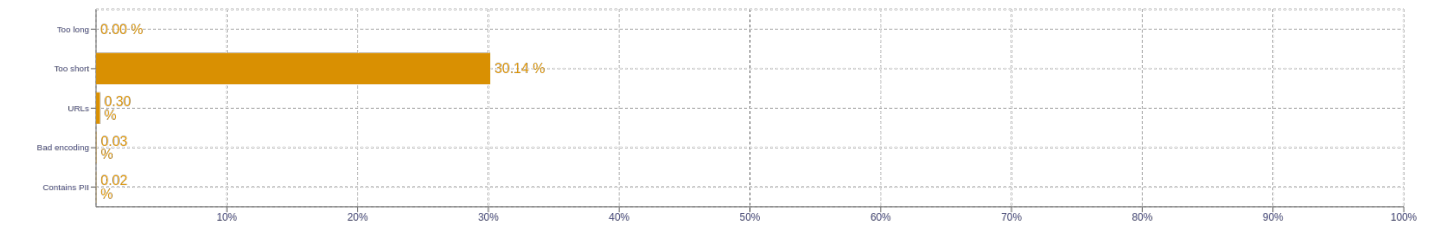
About HPLT Analytics

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