HPLT Analytics report

General overview

Corpus	Date	Language
hplt-v3-taq_Latn	9/18/2025	Tamasheq

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
827	48,115	43,515 (90.44 %)	2.3M	8,927,633	9.65 MB

Top 10 domains

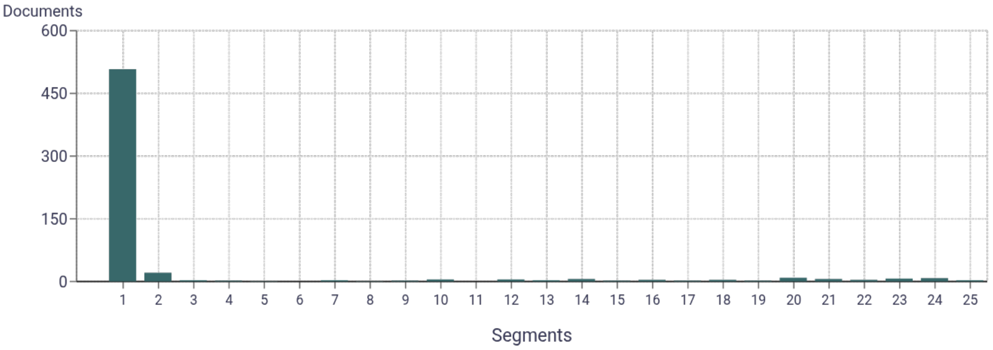
Domain	Docs	% of total
bible.is	519	62.76%
biblehub.com	141	17.05%
newchristianbib...	79	9.55%
ebible.org	39	4.72%
baebol.org	6	0.73%
stepbible.org	5	0.60%
gospelgo.com	4	0.48%
bible.com	4	0.48%
stalk.info	3	0.36%
omniglot.com	3	0.36%

Top 10 TLDs

Domain	Docs	% of total
is	519	62.76%
com	162	19.59%
org	135	16.32%
net	4	0.48%
info	3	0.36%
gq	1	0.12%
es	1	0.12%
edu.vn	1	0.12%
ca	1	0.12%

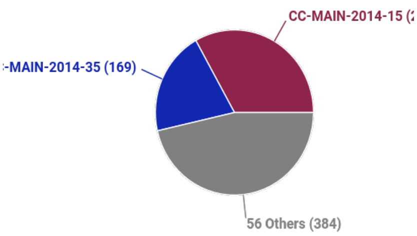
Documents size (in segments)

≤ 25 segments **73.88%** (611 documents)
> 25 segments **26.12%** (216 documents)



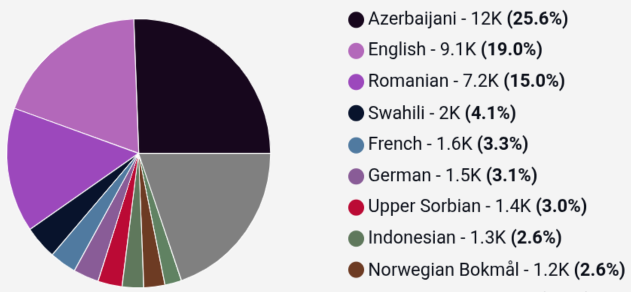
Document collections

CC = 98.43%
IA = 1.57%



Language Distribution

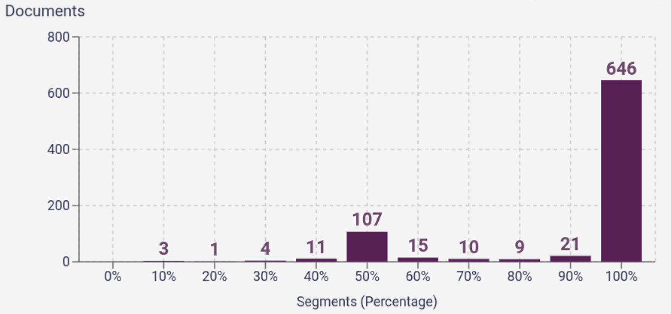
Number of segments in the Tamasheq corpus



*Tamasheq identification might be inaccurate because it is not supported by FastSpell.

Percentage of segments in Tamasheq inside documents

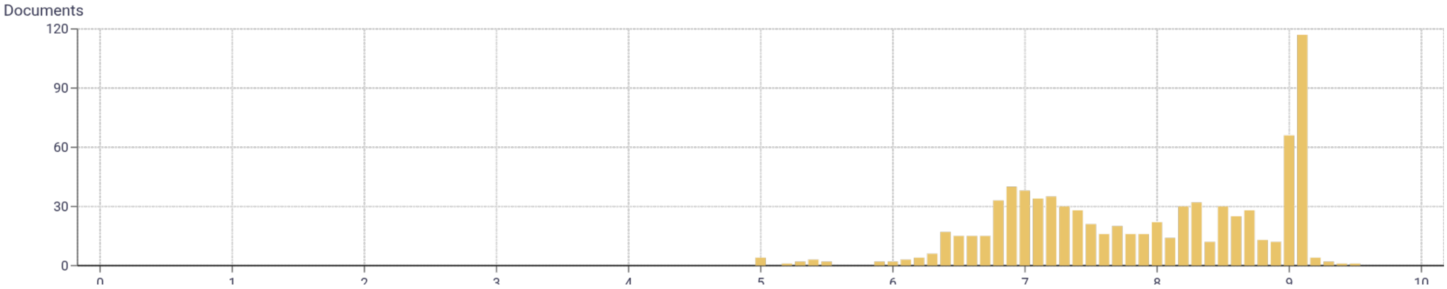
segments < 50% - **2.30%** (19 documents)
segments ≥ 50% - **97.70%** (808 documents)



Distribution of documents by document score

score < 5 - **0%** (0 documents)

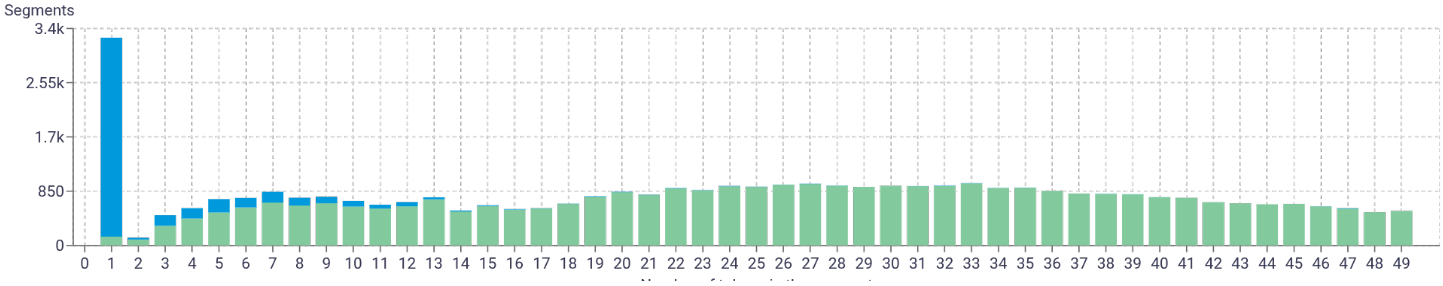
score ≥ 5 - **100%** (827 documents)



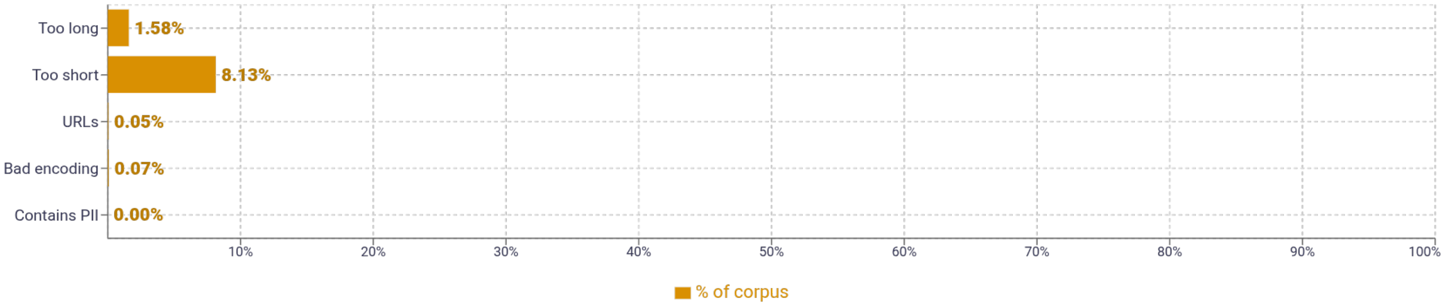
Segment length distribution by token

≤ **49** tokens = **40K** segments | **4.6K** duplicates

> **50** tokens = **8.6K** segments | **8** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	n | 38,547 s | 28,992 dăy | 26,710 net | 18,871 ahay | 18,049
2	ata awan | 3,617 do ahay | 3,181 ata nà | 2,685 ɗo sə | 2,573 anà do | 2,422
3	mer su way | 1,719 anà do ahay | 1,026 wulen su doh | 657 dăy a făl | 654 à man ata | 627
4	ga mer su way | 432 do a yesu ahay | 402 à wulen su doh | 378 gəɗan dungo anà way | 368 sə gəɗan dungo anà | 366
5	sə gəɗan dungo anà way | 364 doh sə mazlab a mbərom | 291 do si mer su way | 279 sa ga mer su way | 239 gəɗan dungo anà way ahay | 239

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				