HPLT Analytics report

General overview

Corpus	Date	Language
hplt-v3-dyu_Latn	9/17/2025	Dyula

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,747	45,172	43,814 (96.99 %)	2M	7,323,745	7.59 MB

Top 10 domains

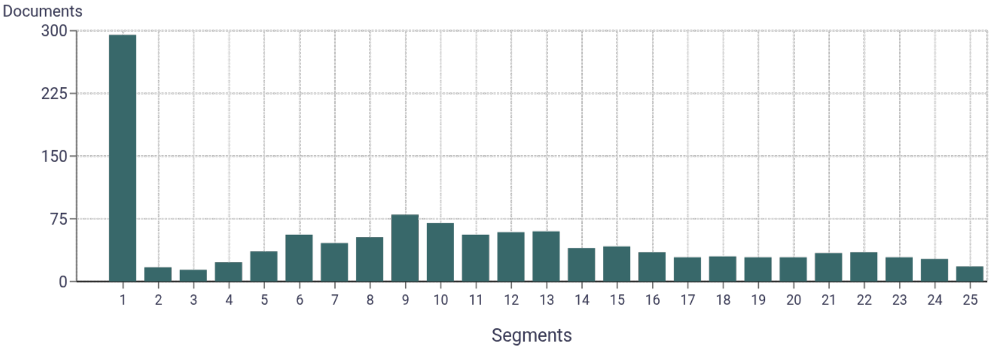
Domain	Docs	% of total
bibles.org	907	51.92%
jw.org	419	23.98%
bible.is	304	17.40%
ebible.org	60	3.43%
gotquestions.org	8	0.46%
premiumtextread...	6	0.34%
kayira.info	5	0.29%
twr360.org	4	0.23%
rfi.fr	4	0.23%
coastsystems.net	3	0.17%

Top 10 TLDs

Domain	Docs	% of total
org	1.4K	80.19%
is	304	17.40%
com	25	1.43%
net	5	0.29%
info	5	0.29%
fr	4	0.23%
ru	1	0.06%
io	1	0.06%
cx	1	0.06%

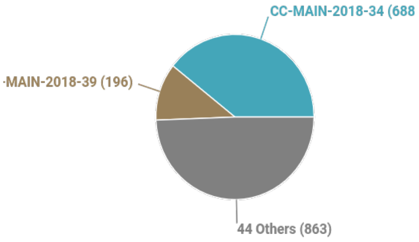
Documents size (in segments)

≤ 25 segments **71.09%** (1.2K documents)
> 25 segments **28.91%** (505 documents)



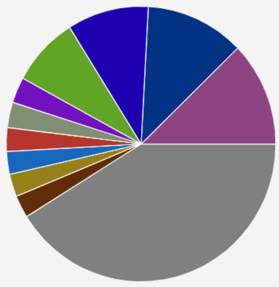
Document collections

CC = 89.12%
IA = 10.88%



Language Distribution

Number of segments in the Dyula corpus

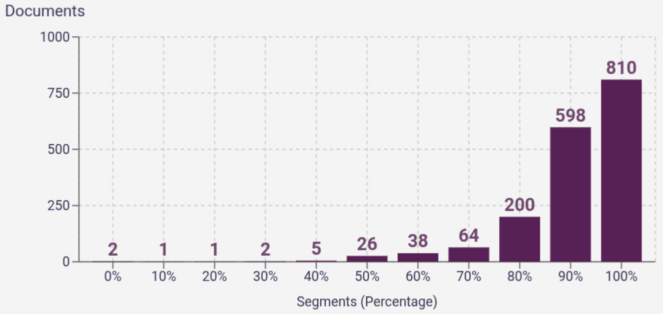


- ● Filipino - 5.6K (**12.3%**)
- ● Welsh - 5.4K (**11.9%**)
- ● English - 4.4K (**9.7%**)
- ● Sundanese - 3.7K (**8.2%**)
- ● Spanish - 1.4K (**3.0%**)
- ● Esperanto - 1.3K (**3.0%**)
- ● Indonesian - 1.3K (**2.8%**)
- ● Finnish - 1.2K (**2.7%**)
- ● Turkish - 1.2K (**2.7%**)
- ● Lojban - 1.2K (**2.6%**)
- ● 147 Others - 19K (**41.2%**)

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

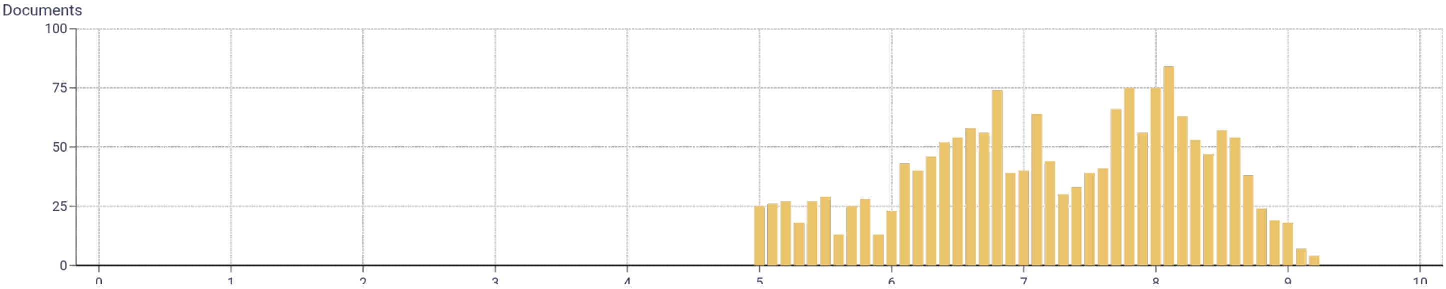
Percentage of segments in Dyula inside documents

segments < 50% - **0.63%** (11 documents)
segments ≥ 50% - **99.37%** (1.7K documents)



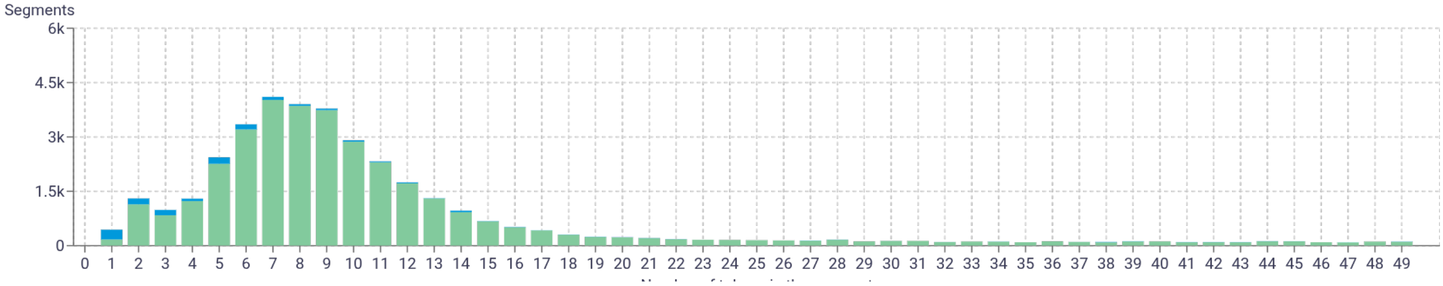
Distribution of documents by document score

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

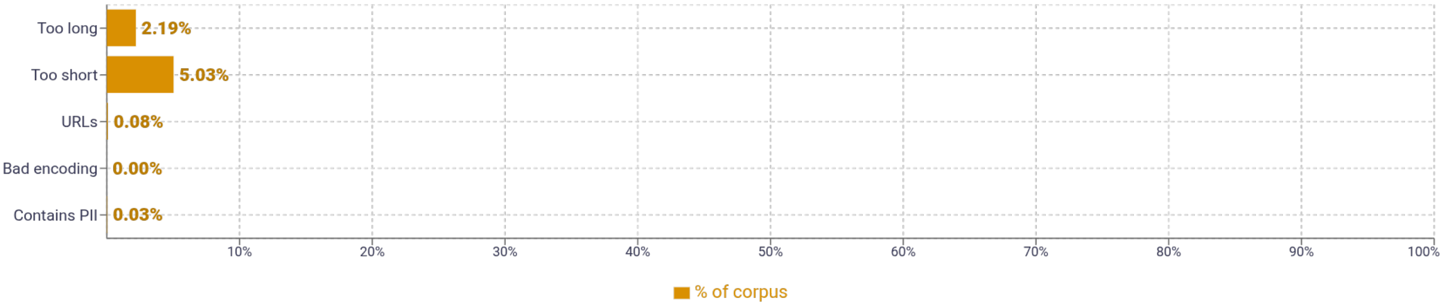


Segment length distribution by token

≤ **49** tokens = **37K** segments | **1.3K** duplicates
> **50** tokens = **8.3K** segments | **30** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	o | 38,858 u | 36,272 ko | 30,175 be | 26,006 k | 23,114
2	zan zan | 5,161 tun be | 4,273 tuma min | 2,417 cogo min | 2,111 tun b | 1,802
3	zan zan zan | 5,032 be se k | 872 u ye ko | 790 ala ka kuma | 535 kelen kelen bɛɛ | 469
4	zan zan zan zan | 4,903 masaba aw ka ala | 256 masaba le ko ten | 228 u kelen kelen bɛɛ | 134 zan zan zan zanzan | 128
5	zan zan zan zan zan | 4,774 zan zan zan zan zanzan | 128 zanzan zan zan zan zan | 127 zan zanzan zan zan zan | 127 zan zan zanzan zan zan | 127

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				