HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-nus_Latn	9/18/2025	Nuer

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
139	3,278	3,214 (98.05 %)	441K	1,417,341	1.63 MB

Top 10 domains

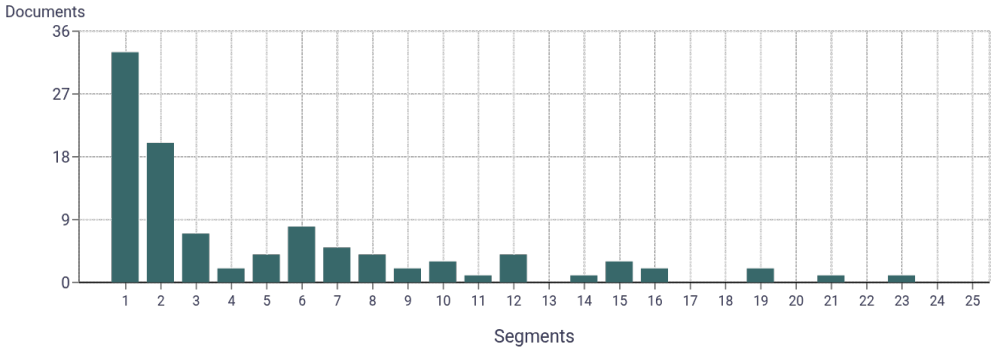
Domain	Docs	% of total
bible.is	33	23.74%
nanetya-foundat...	24	17.27%
jw.org	18	12.95%
ebible.org	18	12.95%
indiana.edu	11	7.91%
consumer.vic.go...	6	4.32%
jobs.vic.gov.au	4	2.88%
vietnamimmigrat...	2	1.44%
triquidechicahu...	2	1.44%
lchs.com.au	2	1.44%

Top 10 TLDs

Domain	Docs	% of total
org	65	46.76%
is	33	23.74%
vic.gov.au	15	10.79%
edu	11	7.91%
com	9	6.47%
com.au	3	2.16%
net	1	0.72%
io	1	0.72%
biz	1	0.72%

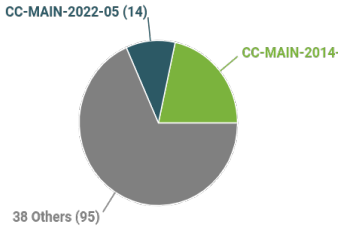
Documents size (in segments)

≤ 25 segments **74.1%** (103 documents)
> 25 segments **25.9%** (36 documents)



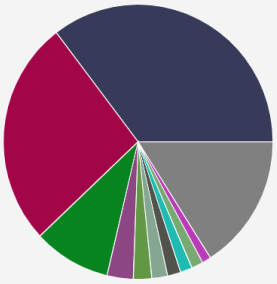
Document collections

CC = 84.17%
IA = 15.83%



Language Distribution

Number of segments in the Nuer corpus

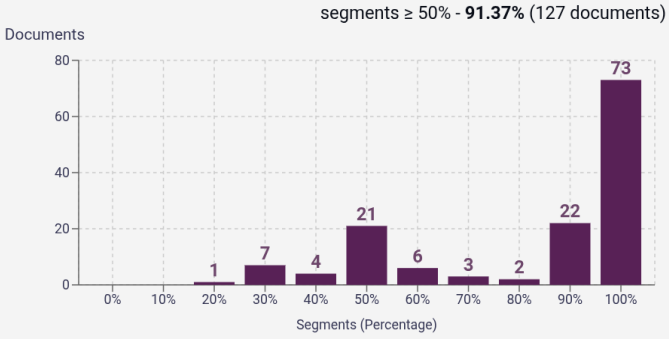


- Finnish - 1.2K **(35.4%)**
- Slovenian - 873 **(26.6%)**
- English - 304 **(9.3%)**
- Spanish - 103 **(3.1%)**
- Estonian - 72 **(2.2%)**
- Dutch - 62 **(1.9%)**
- Tatar - 52 **(1.6%)**
- Albanian - 47 **(1.4%)**
- Tosk Albanian - 46 **(1.4%)**
- Czech - 35 **(1.1%)**
- 77 Others - 524 **(16.0%)**

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

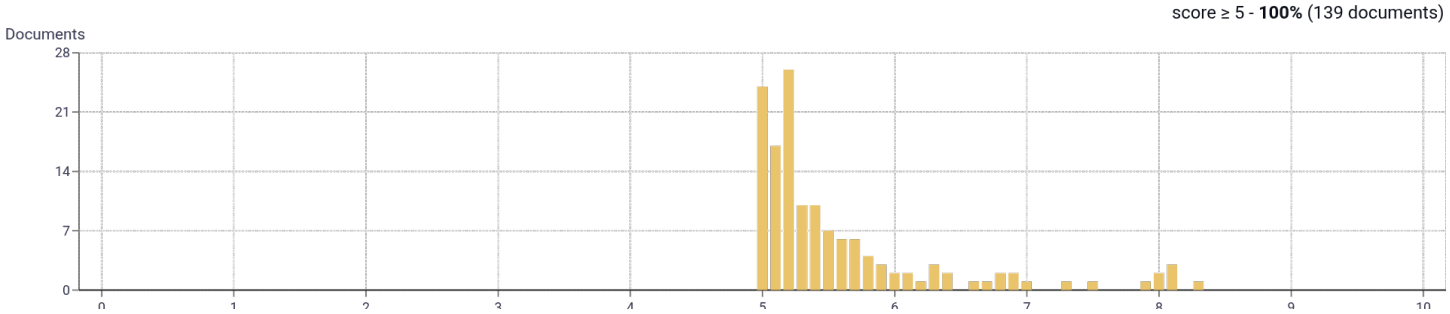
Percentage of segments in Nuer inside documents

segments < 50% - **8.63%** (12 documents)
segments ≥ 50% - **91.37%** (127 documents)



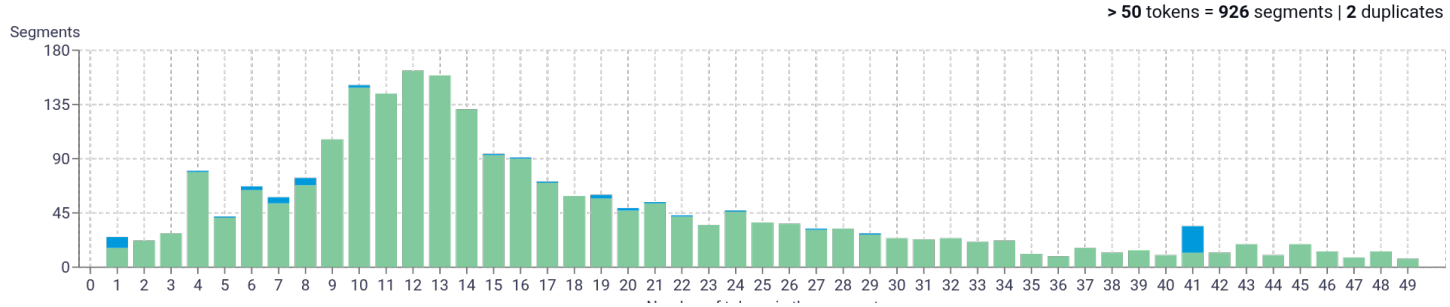
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (139 documents)

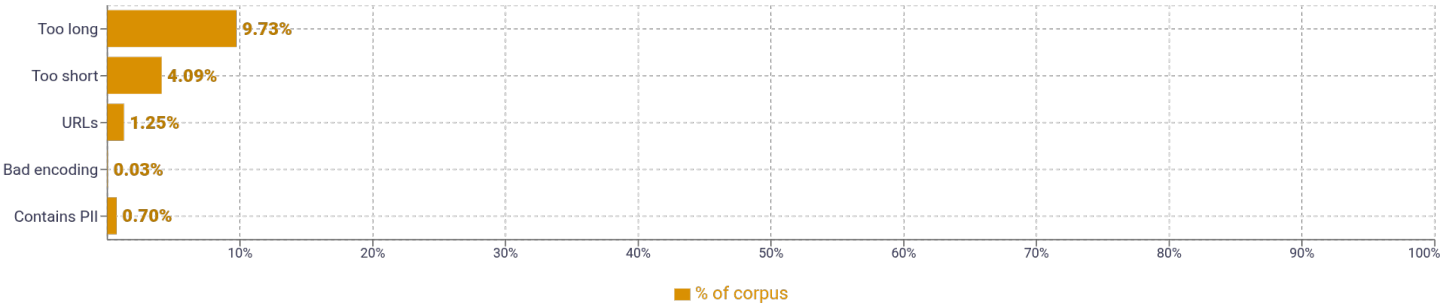


Segment length distribution by token

≤ 49 tokens = **2.4K** segments | **62** duplicates
> 50 tokens = **926** segments | **2** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	kɛ | 15,771 · kä | 14,317 · ɣöö | 5,736 · ɛ | 4,924 · jɛ | 4,688
2	kɛ ɣöö | 2,836 · kɛ kui | 1,732 · ɛn ɣöö | 1,460 · kä cuɛ | 1,122 · kä cu | 1,058
3	kɛ ɣöö bɛ | 242 · kä duundɛ ɣöö | 242 · kä cuɛ jɛ | 205 · kä cuɛ kɛ | 195 · kä cu yecu | 191
4	gɔaa in rɛl rɔ | 92 · kä cuɛ kɛ jiök | 85 · cu yecu jɛ jiök | 60 · cu yecu kɛ jiök | 56 · kä cu yecu kɛ | 52
5	yiëë in gɔaa in rɛl | 75 · kä ni ɛn guäth ɛmɔ | 34 · kä cu yecu kɛ jiök | 27 · written by michael thilyang gatkek | 24 · with special thanks to fr | 24

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				