

General overview

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

Volumes

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

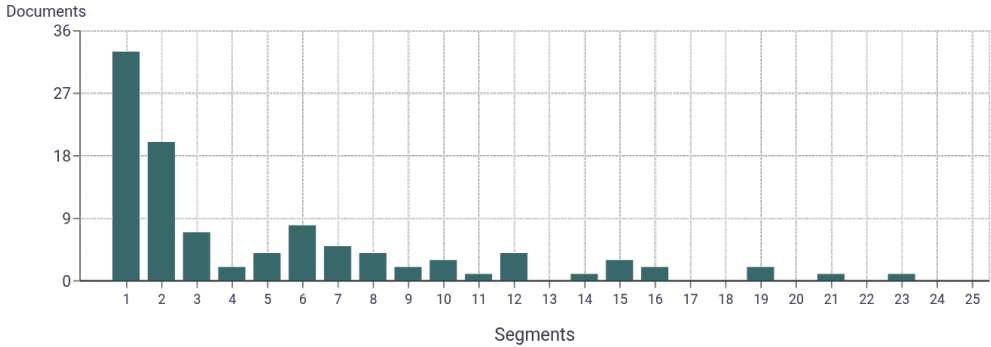
Top 10 domains

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

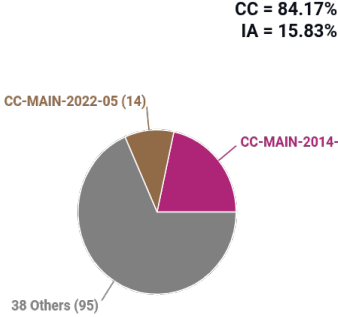
Domain	Docs	% of total
org	65	46.76%
is	33	23.74%
vic.gov.au	15	10.79%
edu	11	7.91%
com	9	6.47%
com.au	3	2.16%
net	1	0.72%
io	1	0.72%
biz	1	0.72%

Documents size (in segments) ⓘ



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

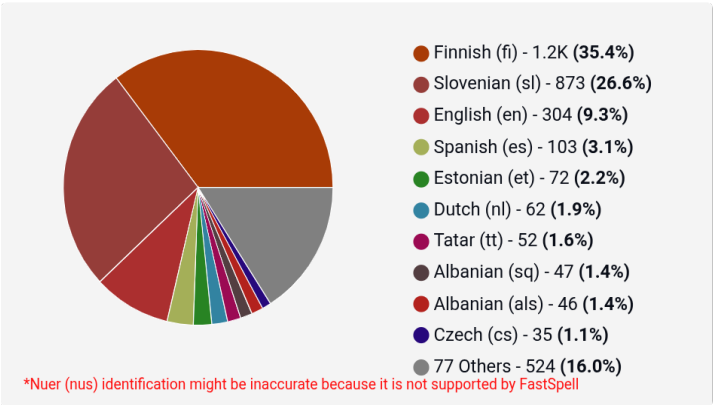
Document collections



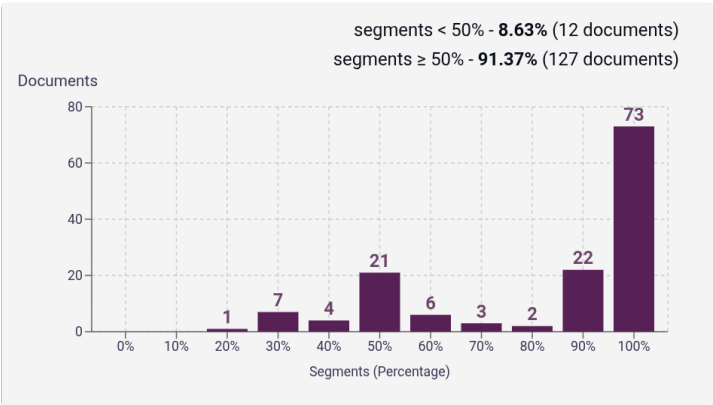
CC = 84.17%
IA = 15.83%

Language Distribution

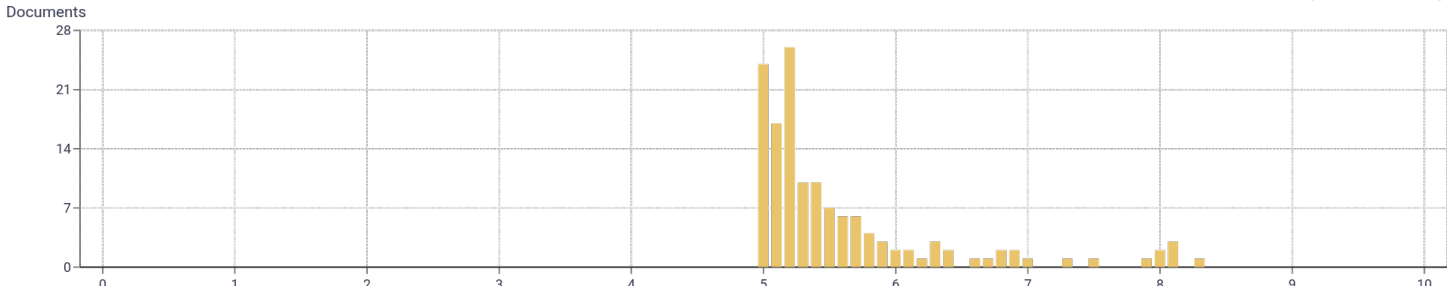
Number of segments in the Nuer (nus) corpus



Percentage of segments in Nuer (nus) inside documents

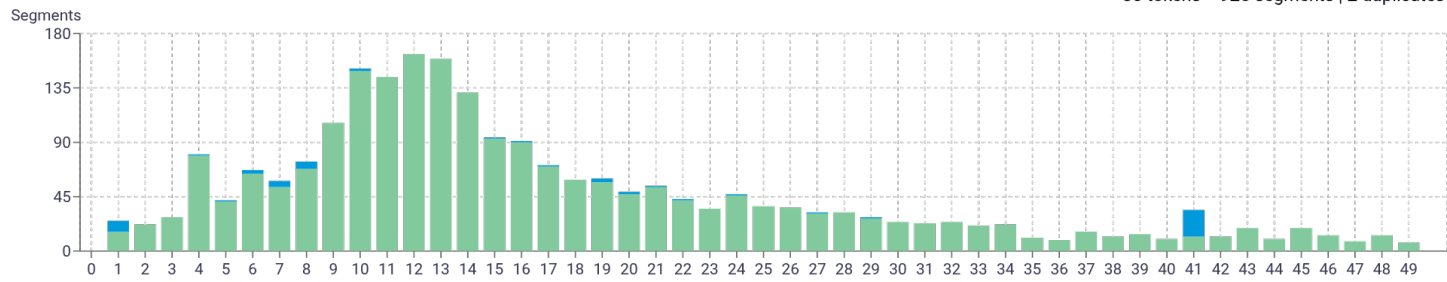


Distribution of documents by document score

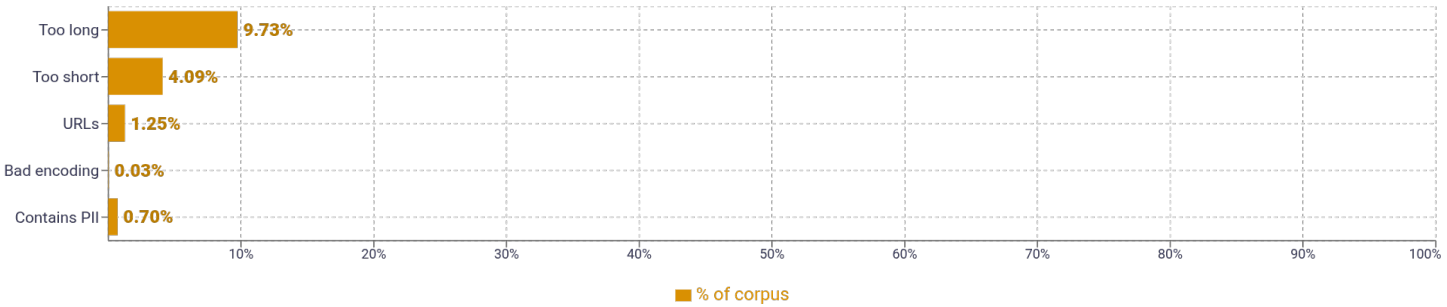


Segment length distribution by token

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



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>kε 15,771</div> <div>kā 14,317</div> <div>γöö 5,736</div> <div>ε 4,924</div> <div>jε 4,688</div>	
2	<div>kε γöö 2,836</div> <div>kε kui 1,732</div> <div>εn γöö 1,460</div> <div>kā cυε 1,122</div> <div>kā cu 1,058</div>	
3	<div>kε γöö bε 242</div> <div>kā duunde γöö 242</div> <div>kā cυε jε 205</div> <div>kā cυε kε 195</div> <div>kā cu yecu 191</div>	
4	<div>ɔpaa in rεl rɔ 92</div> <div>kā cυε kε jïök 85</div> <div>cu yecu jε jïök 60</div> <div>cu yecu kε jïök 56</div> <div>kā cu yecu kε 52</div>	
5	<div>yiëë in ɔpaa in rεl 75</div> <div>kā ni en guäth εmɔ 34</div> <div>kā cu yecu kε jïök 27</div> <div>written by michael thilyang gatkek 24</div> <div>with special thanks to fr 24</div>	

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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>, , , 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	ntp
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				