HPLT Analytics report

General overview

Corpus	Analytics date	Language
nus_Latn.jsonl.tsv	12/6/2024	Nuer (nus)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
272	8,514	2,720 (31.95 %)	583K	2.13 MB	1,873,800

Top 10 domains

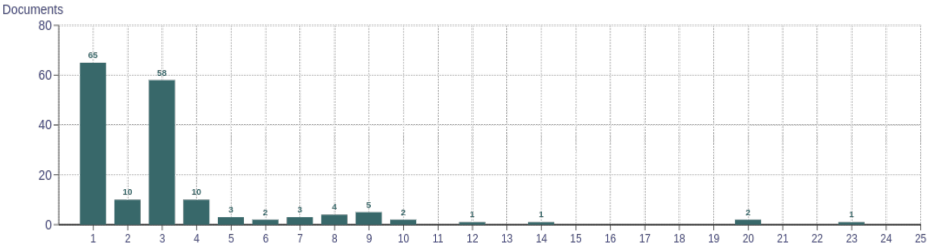
Domain	Docs	% of total
bible.is	115	42.28
indiana.edu	65	23.90
consumer.vic.gov.au	27	9.93
jw.org	25	9.19
nanetya-foundation.org	9	3.31
ironchariots.org	7	2.57
triquidechicahuaxtla.org	5	1.84
multicultural.vic.gov.au	3	1.10
4laws.com	3	1.10
nueronline.com	2	0.74

Top 10 TLDs

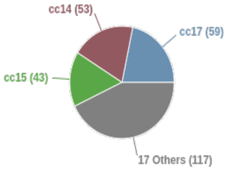
Domain	Docs	% of total
is	115	42.28
edu	65	23.90
org	47	17.28
vic.gov.au	33	12.13
com	9	3.31
com.au	3	1.10

Documents size (in segments)

<= 25 segments **61.4%** (167 documents)
> 25 segments **38.6%** (105 documents)

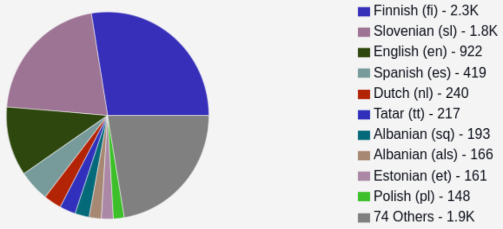


Documents by collection



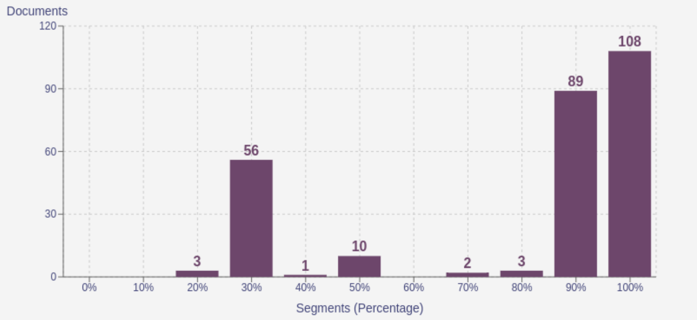
Language Distribution

Number of segments



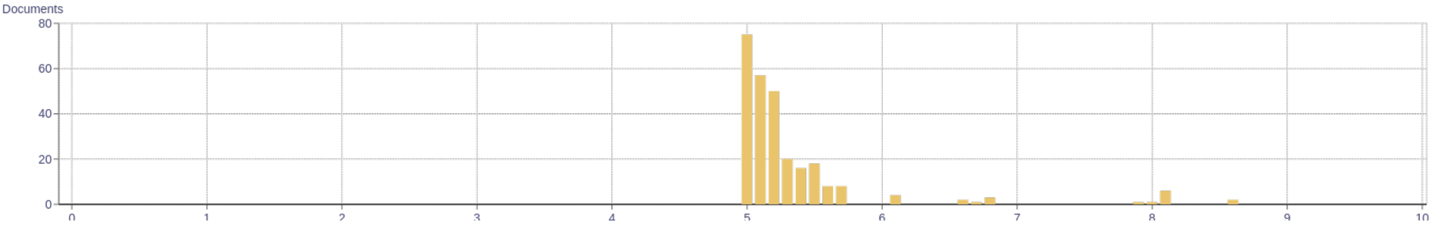
*Nuer (nus) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Nuer (nus) inside documents



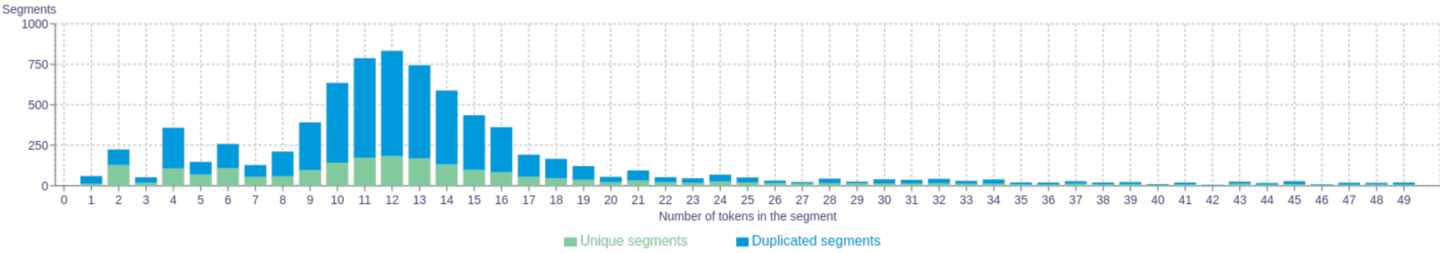
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (272 documents)

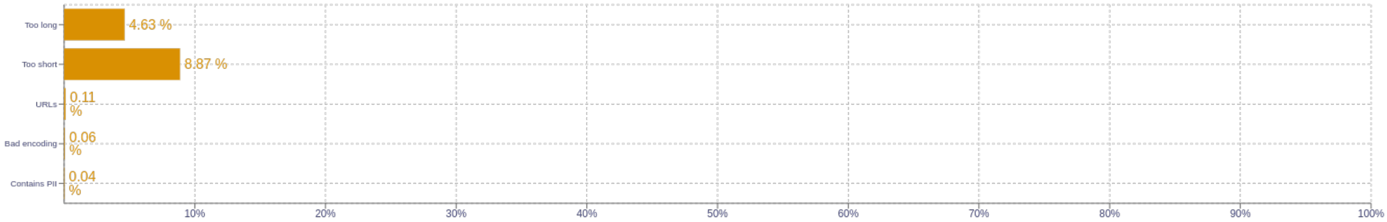


Segment length distribution by token

<= 49 tokens = **2.2K** segments | **5.5K** duplicates
> 50 tokens = **857** segments | **301** duplicates



Segment noise distribution



Frequent n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>kɛ | 18498
kä | 16475
ɣöö | 5735
ji | 5551
ɛ | 5545</td></tr>
<tr><td>2</td><td>kɛ ɣöö | 3025
kɛ kui | 1751
kä cu | 1364
ɛn ɣöö | 1360
kä cuɛ | 1327</td></tr>
<tr><td>3</td><td>kä duundɛ ɣöö | 257
kɛ ɣöö bɛ | 256
kä cuɛ jɛ | 234
kä cuɛ kɛ | 206
irä jiye kuna | 204</td></tr>
<tr><td>4</td><td>gɔaa in rɛl rɔ | 92
kä cuɛ kɛ jiök | 83
kɛ run ti kur | 71
cu yecu jɛ jiök | 62
cu yecu kɛ jiök | 56</td></tr>
<tr><td>5</td><td>yiëë in gɔaa in rɛl | 76
lɛ te kɛ run ti | 40
kä ni ɛn guäth ɛmɔ | 34
tin ci kuoth kɛ lɛy | 33
jesús ktei bata shäk wa | 33</td></tr>
</table>

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