

General overview

Corpus	Analytics date	Language
umb_Latn.jsonl.tsv	11/27/2024	Umbundu (umb)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
2,471	59,912	41,987 (70.08 %)	2.9M	14.8 MB	15,348,339

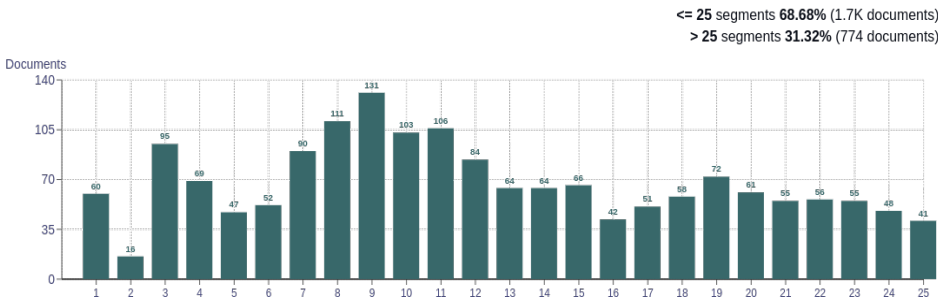
Top 10 domains

Domain	Docs	% of total
jw.org	1.9K	78.19
neweralive.na	167	6.76
bible.is	108	4.37
kundana.com.na	95	3.84
blogspot.com	28	1.13
globalrecordings.net	26	1.05
pngscriptures.org	19	0.77
watchtower.org	12	0.49
unicode.org	11	0.45
blogspot.pt	7	0.28

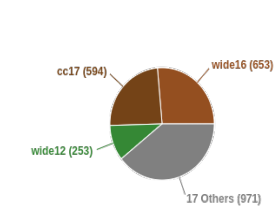
Top 10 TLDs

Domain	Docs	% of total
org	2K	80.49
na	167	6.76
is	108	4.37
com.na	98	3.97
com	57	2.31
net	29	1.17
pt	7	0.28
bible	4	0.16
co	3	0.12
cc	3	0.12

Documents size (in segments)

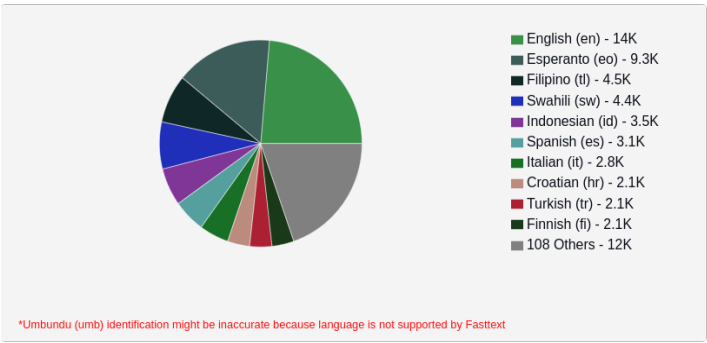


Documents by collection

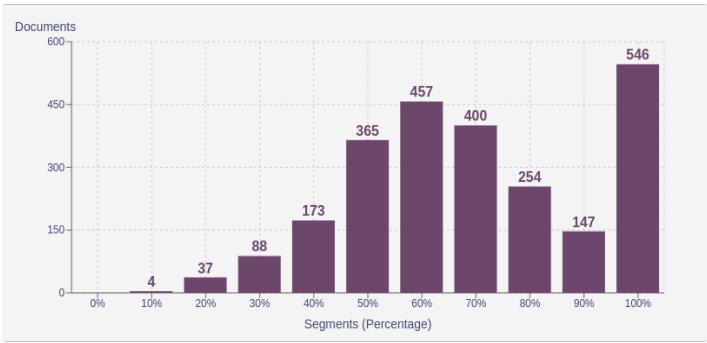


Language Distribution

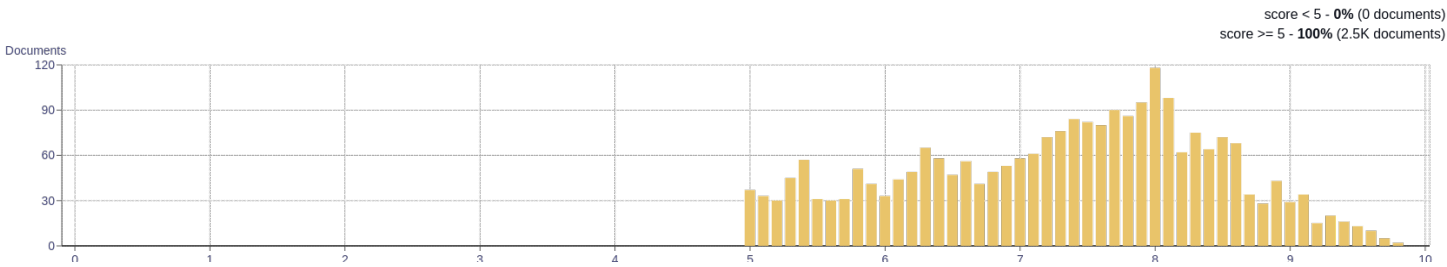
Number of segments



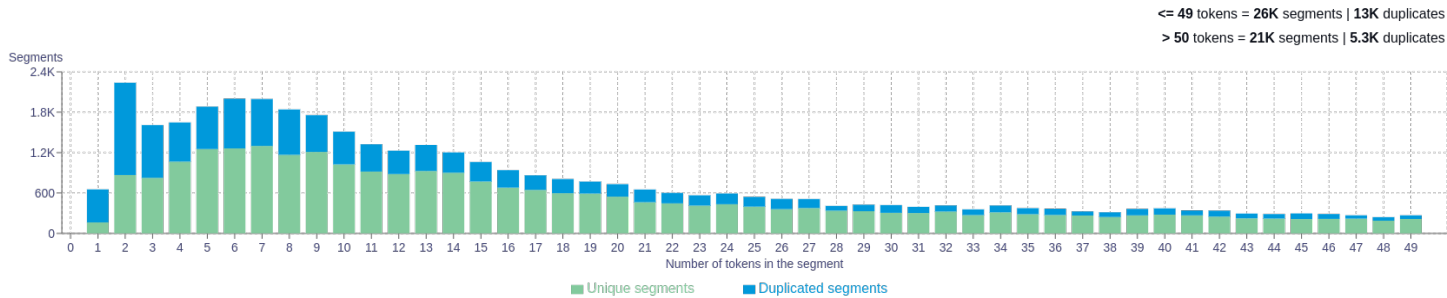
Percentage of segments in Umbundu (umb) inside documents



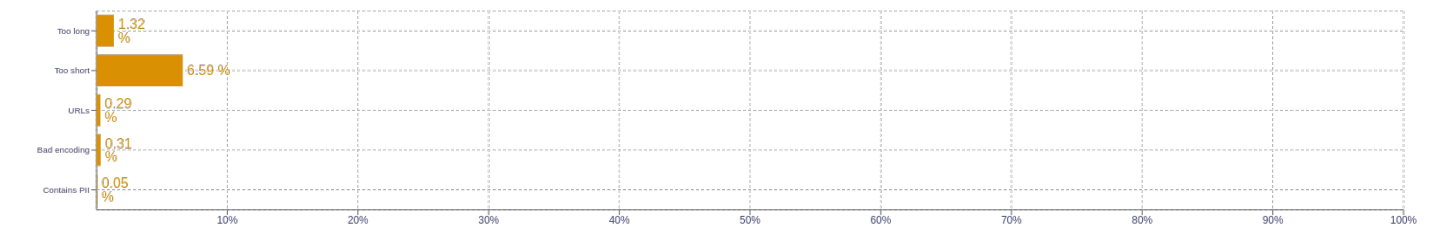
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>li   58413</div> <div>va   38896</div> <div>ka   27550</div> <div>okwa   25024</div> <div>oku   21506</div>
2	<div>okwa li   12853</div> <div>ova li   4999</div> <div>li va   4042</div> <div>nosho yo   3530</div> <div>ka kala   3162</div>
3	<div>ova li va   2625</div> <div>shi na sha   2374</div> <div>jesus okwa li   1452</div> <div>ovo va li   1078</div> <div>oo a li   1038</div>
4	<div>lie lie lie lie   496</div> <div>okwa li a lombwela   429</div> <div>wa omae dake ka   329</div> <div>omae dake ka yo   329</div> <div>wo suki nano wa   328</div>
5	<div>lie lie lie lie lie   493</div> <div>wa omae dake ka yo   329</div> <div>wo suki nano wa omae   328</div> <div>suki nano wa omae dake   328</div> <div>nano wa omae dake ka   328</div>

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 of 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>