

General overview

Corpus	Analytics date	Language
bem_Latn.jsonl.tsv	10/4/2024	Bemba (bem)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
6,136	133,538	87,809 (65.76 %)	5.7M	31.03 MB	32,196,810

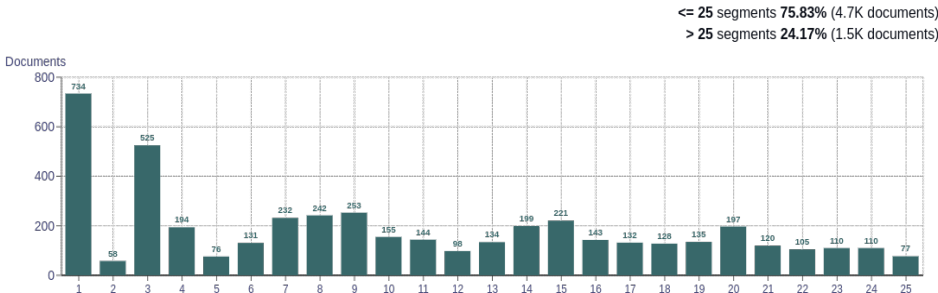
Top 10 domains

Domain	Docs	% of total
jw.org	4.4K	72.26
bible.is	893	14.55
worldslastchance.com	398	6.49
globalrecordings.net	28	0.46
bibles.org	24	0.39
kingsmanga.net	23	0.37
bible.com	19	0.31
watchtower.org	19	0.31
unicode.org	16	0.26
blogspot.com	12	0.20

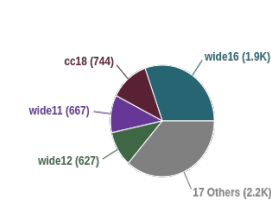
Top 10 TLDs

Domain	Docs	% of total
org	4.6K	74.56
is	893	14.55
com	527	8.59
net	69	1.12
cc	10	0.16
info	9	0.15
tv	7	0.11
org.za	6	0.10
co	5	0.08
in	4	0.07

Documents size (in segments)

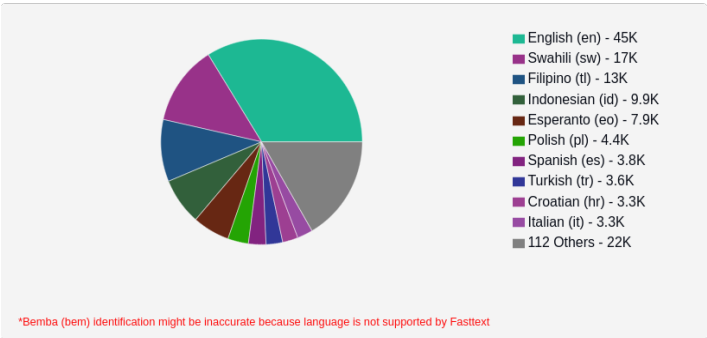


Documents by collection

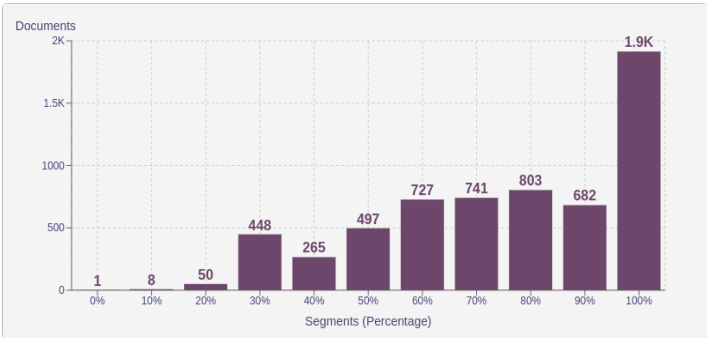


Language Distribution

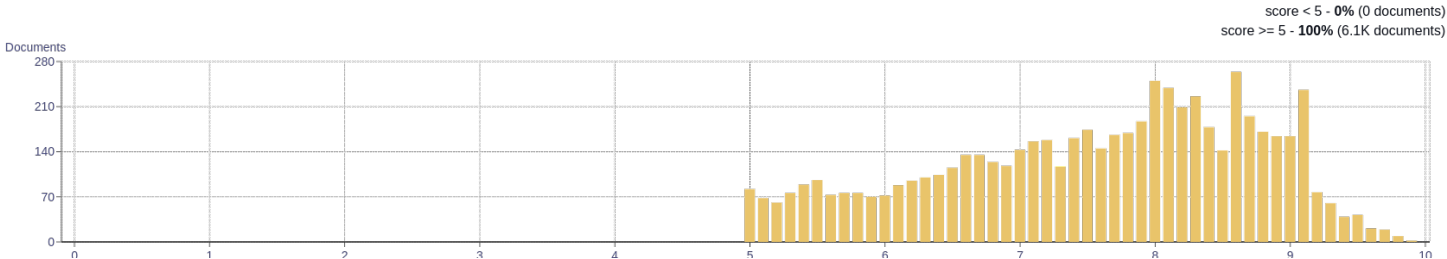
Number of segments



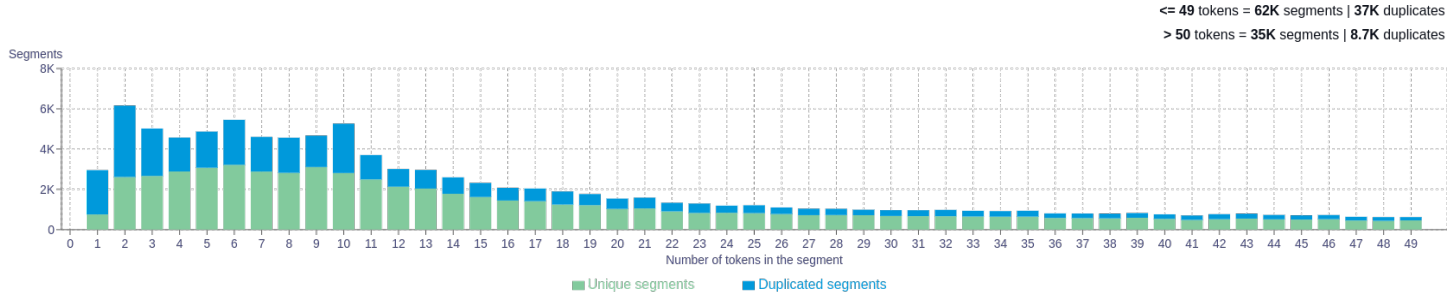
Percentage of segments in Bemba (bem) inside documents



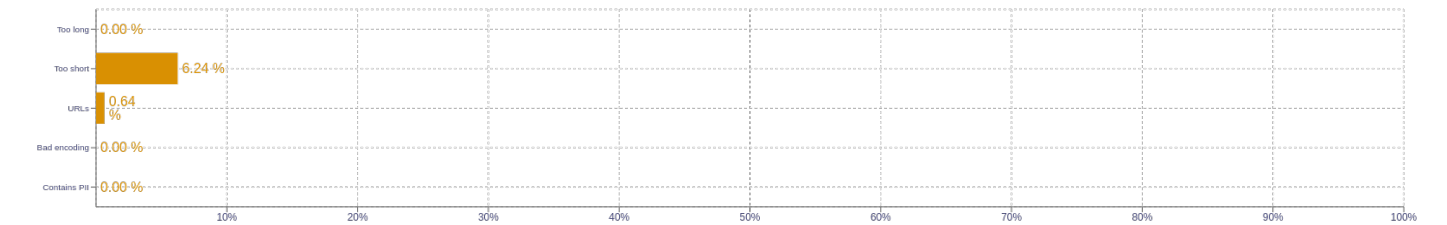
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>u 29843</div> <div>a 28044</div> <div>yesu 21515</div> <div>i 19488</div> <div>ng 17335</div>
2	<div>u ng 2040</div> <div>u yesu 1937</div> <div>mukashi banashi 1724</div> <div>furusato saisei 1717</div> <div>mambo a 1549</div>
3	<div>fumasabomba de fumasabomba 8710</div> <div>nihon no mukashi 1724</div> <div>mukashi banashi episode 1447</div> <div>ore wa shibushibu 671</div> <div>yuusha ni narenakatta 668</div>
4	<div>nihon no mukashi banashi 1724</div> <div>narenakatta ore wa shibushibu 671</div> <div>bakamonyi ba kwa yehoba 665</div> <div>yuusha ni narenakatta ore 648</div> <div>ore wa shibushibu shuushoku 604</div>
5	<div>fumasabomba de fumasabomba de fumasabomba 8708</div> <div>nihon no mukashi banashi episode 1447</div> <div>narenakatta ore wa shibushibu shuushoku 604</div> <div>ore wa shibushibu shuushoku o 358</div> <div>wine wine wine wine wine 234</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 "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>, , , 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>