

General overview

Corpus	Analytics date	Language
tsn_Latn.jsonl.tsv	9/20/2024	Tswana (tn)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
6,050	132,173	77,616 (58.72 %)	6.1M	26.44 MB	27,545,230

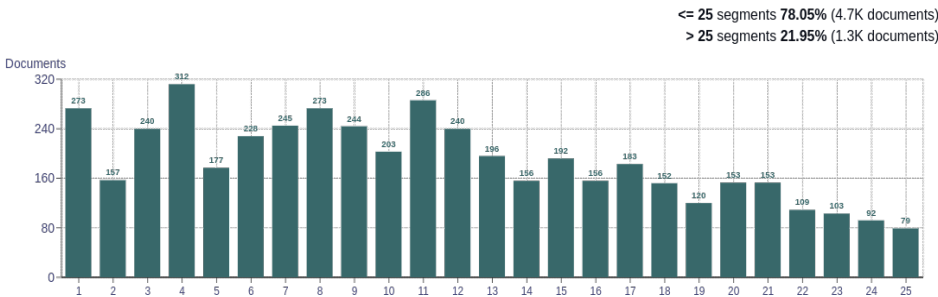
Top 10 domains

Domain	Docs	% of total
jw.org	2.4K	39.95
wikipedia.org	1.2K	19.49
biblesa.co.za	530	8.76
southafrica.co.za	247	4.08
gov.bw	145	2.40
sciencegraph.net	145	2.40
nwu.ac.za	137	2.26
dikgang24.news	76	1.26
oxforddictionaries.com	70	1.16
mmegi.bw	44	0.73

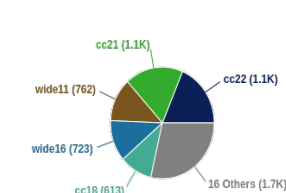
Top 10 TLDs

Domain	Docs	% of total
org	3.8K	62.86
co.za	976	16.13
com	390	6.45
net	190	3.14
bw	190	3.14
ac.za	151	2.50
news	76	1.26
support	38	0.63
org.za	37	0.61
gov.za	37	0.61

Documents size (in segments)

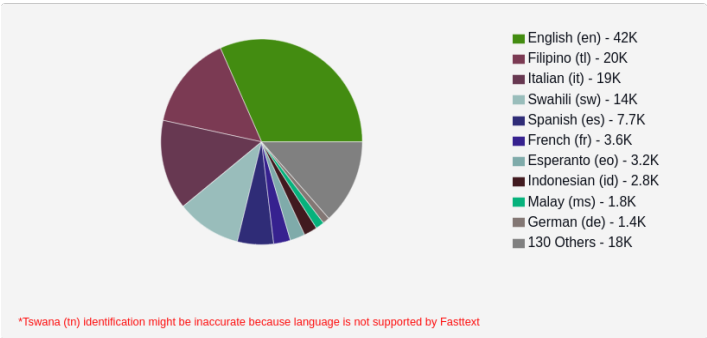


Documents by collection

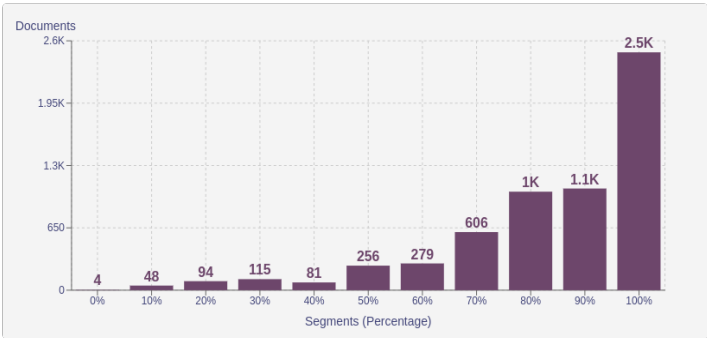


Language Distribution

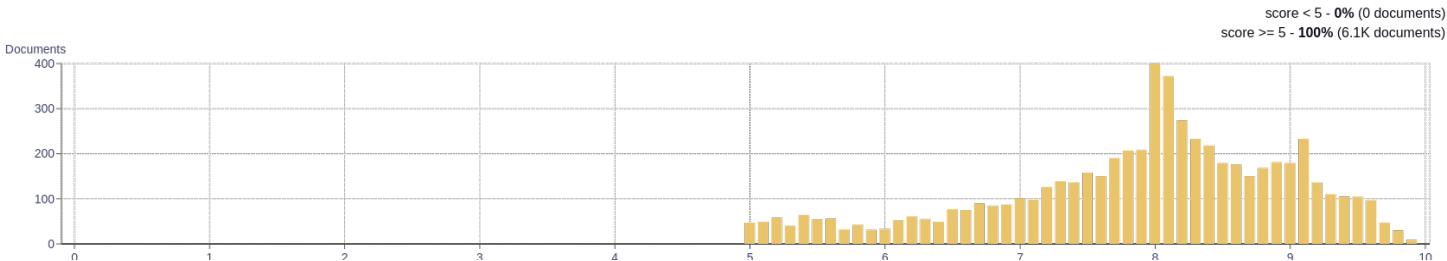
Number of segments



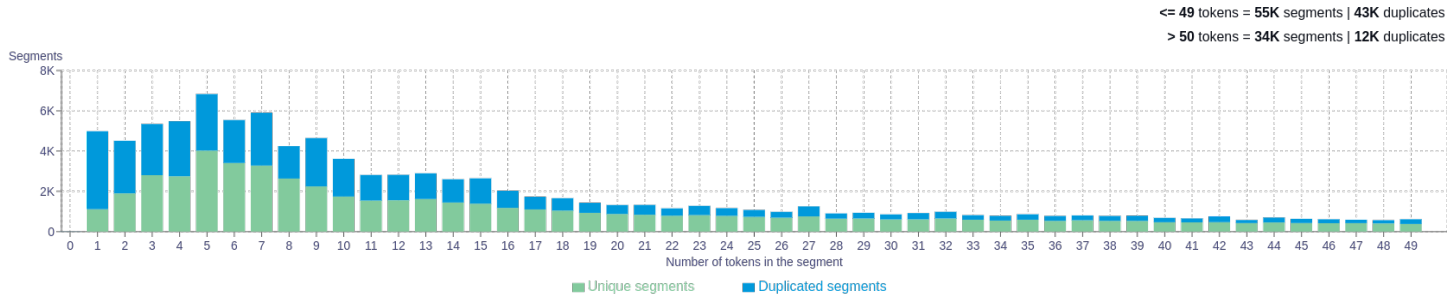
Percentage of segments in Tswana (tn) inside documents



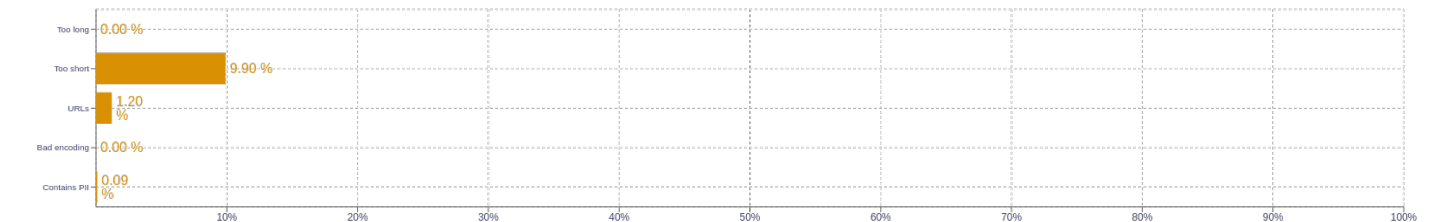
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>a 225859</div> <div>ya 123090</div> <div>wa 52375</div> <div>tse 46028</div> <div>mme 42568</div>
2	<div>a a 10824</div> <div>boing boing 7724</div> <div>jo bo 5198</div> <div>neng a 4639</div> <div>tse dingwe 4146</div>
3	<div>boing boing boing 7723</div> <div>a ne a 3938</div> <div>a neng a 3122</div> <div>a bo a 2156</div> <div>a ba a 1072</div>
4	<div>boing boing boing boing 7722</div> <div>jesu o ne a 1598</div> <div>basupi ba ga jehofa 1483</div> <div>jehofa o ne a 1083</div> <div>a se ka a 1034</div>
5	<div>boing boing boing boing boing 7721</div> <div>jaana o meri jaana film 505</div> <div>song jaana o meri jaana 504</div> <div>tsuki ga michibiku isekai douchuu 471</div> <div>thano!o ya lefatshe le lesa 391</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>