

General overview

Corpus	Analytics date	Language
sna_Latn.jsonl.tsv	11/28/2024	Shona (sn)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
61,076	1,201,679	864,345 (71.93 %)	29M	183.12 MB	191,477,676

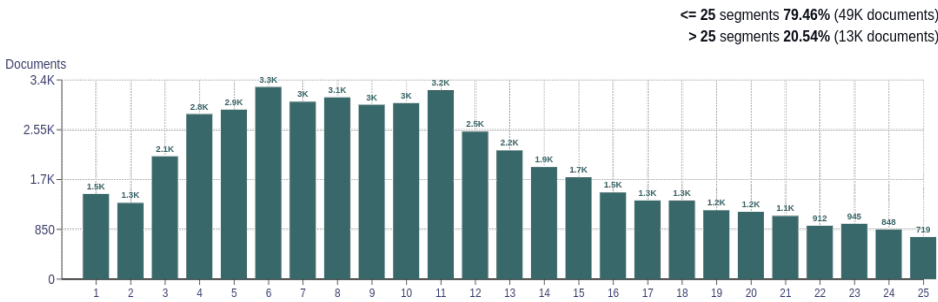
Top 10 domains

Domain	Docs	% of total
voashona.com	6.7K	11.04
wikipedia.org	5.5K	8.95
jw.org	5K	8.23
linuxadictos.com	2.3K	3.84
eturbonews.com	1.6K	2.66
kwayedza.co.zw	1.6K	2.57
martech.zone	1K	1.64
actualidadiphone.com	857	1.40
zimkatorike.com	726	1.19
masasieharare.com	642	1.05

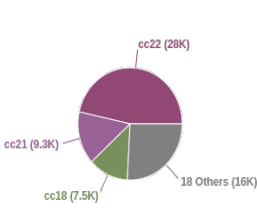
Top 10 TLDs

Domain	Docs	% of total
com	41K	66.65
org	12K	20.40
co.zw	2.3K	3.74
zone	1K	1.64
net	998	1.63
africa	374	0.61
fr	214	0.35
co.za	212	0.35
ru	194	0.32
org.uk	170	0.28

Documents size (in segments)

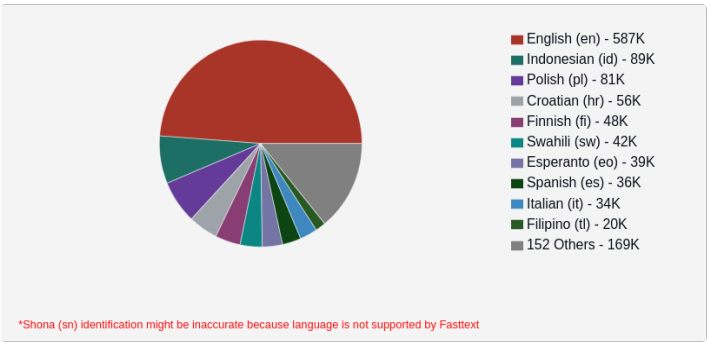


Documents by collection

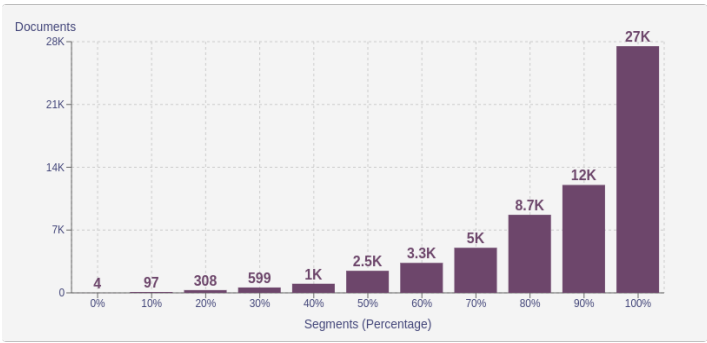


Language Distribution

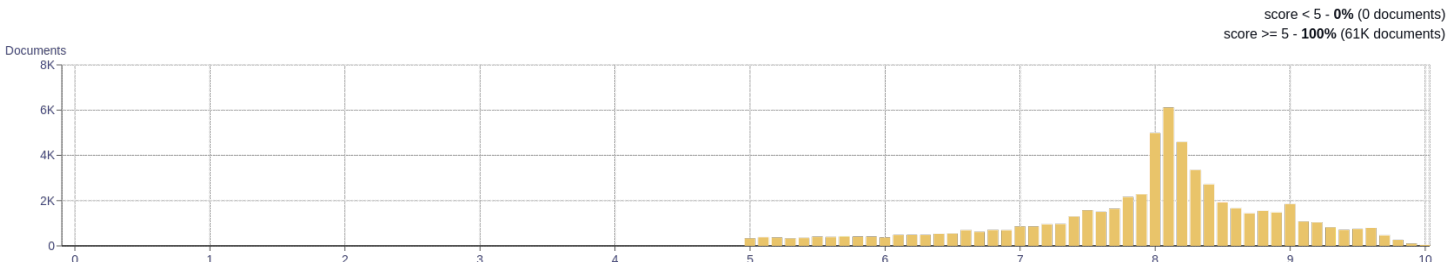
Number of segments



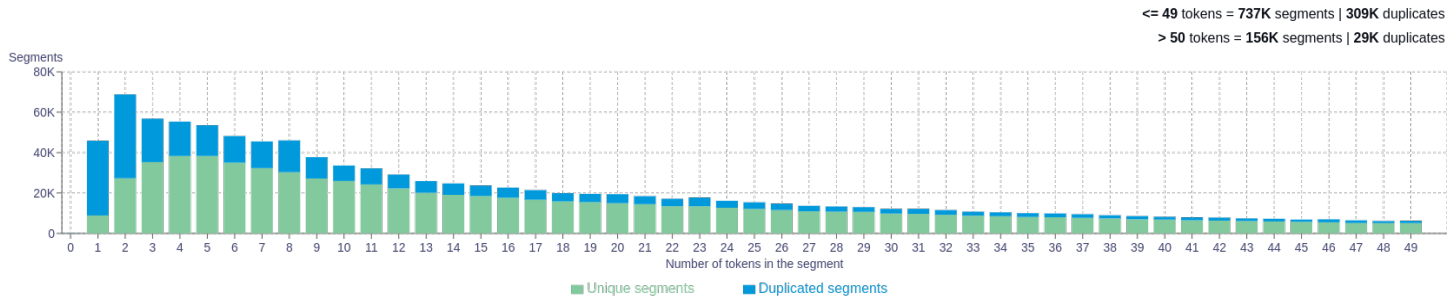
Percentage of segments in Shona (sn) inside documents



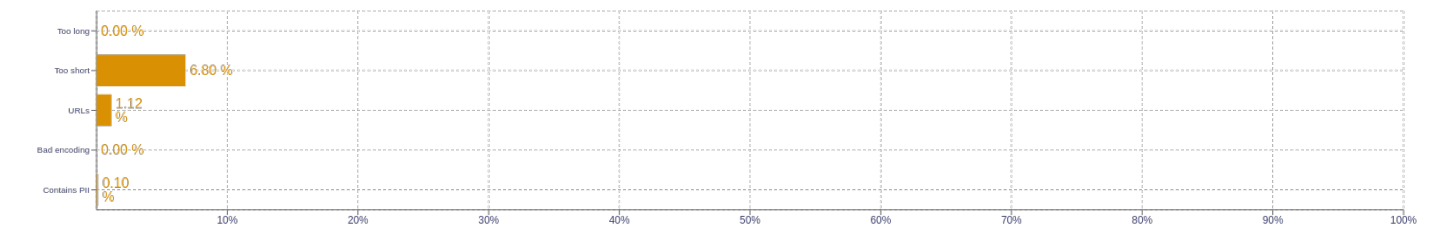
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>kana   282828</div> <div>iyo   138946</div> <div>asi   102289</div> <div>kubva   86647</div> <div>iri   73004</div>
2	<div>chirp chirp   15672</div> <div>zviri nyore   6956</div> <div>edit source   6669</div> <div>makumi maviri   6263</div> <div>imwe chete   4739</div>
3	<div>chirp chirp chirp   15667</div> <div>kana iwe uchida   4547</div> <div>uchinge uchinge uchinge   3636</div> <div>kana iwe uri   2168</div> <div>panguva imwe chete   1991</div>
4	<div>chirp chirp chirp chirp   15664</div> <div>uchinge uchinge uchinge uchinge   3488</div> <div>kuverenga nyaya ino mumutauro   1379</div> <div>here kuverenga nyaya ino   1379</div> <div>yenyika itsva yemagwaro matsvene   1355</div>
5	<div>chirp chirp chirp chirp chirp   15661</div> <div>uchinge uchinge uchinge uchinge uchinge   3347</div> <div>ungada here kuverenga nyaya ino   1379</div> <div>kuverenga nyaya ino mumutauro we   1379</div> <div>here kuverenga nyaya ino mumutauro   1379</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>