HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-sna_Latn	9/18/2025	Shona (sn)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
183,006	3,779,319	3,019,107 (79.88 %)	93M	624,924,975	597.31 MB

Top 10 domains

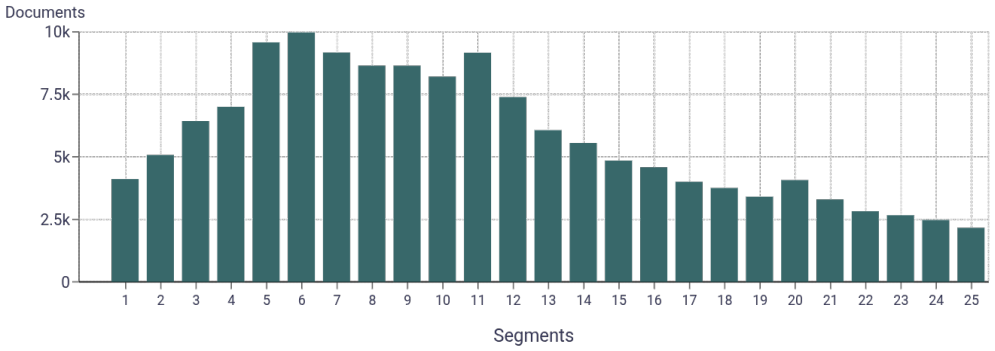
Domain	Docs	% of total
voashona.com	17K	9.48%
linuxadictos.com	6.8K	3.74%
jw.org	5.6K	3.03%
martech.zone	5.5K	3.01%
kwayedza.co.zw	5.2K	2.86%
eturbonews.com	4.6K	2.52%
actualidadiphon...	3.9K	2.16%
wikipedia.org	2.5K	1.36%
soydemac.com	2.4K	1.34%
actualidadliter...	1.5K	0.83%

Top 10 TLDs

Domain	Docs	% of total
com	141K	77.16%
org	11K	6.03%
co.zw	6.2K	3.36%
zone	5.5K	3.01%
net	4.4K	2.40%
es	1.3K	0.72%
online	1.3K	0.71%
fr	1.2K	0.67%
ru	1.2K	0.64%
news	862	0.47%

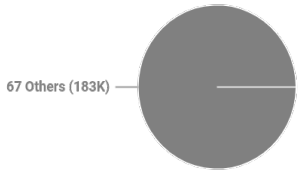
Documents size (in segments)

≤ 25 segments **78.21%** (143K documents)
> 25 segments **21.79%** (40K documents)



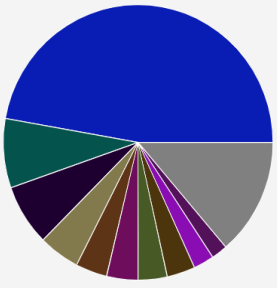
Document collections

CC = 97.66%
IA = 2.34%



Language Distribution

Number of segments in the Shona (sn) corpus

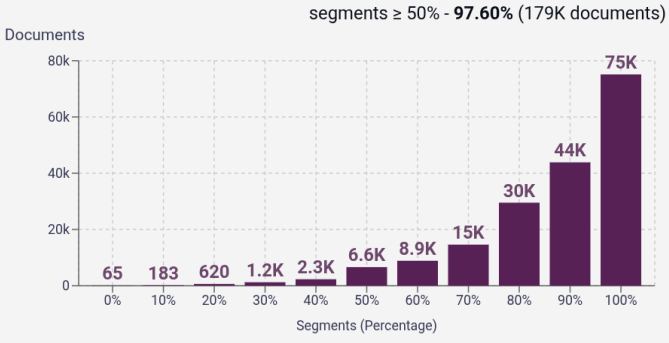


- English (en) - 1.8M **(47.2%)**
- Indonesian (id) - 307K **(8.1%)**
- Polish (pl) - 272K **(7.2%)**
- Croatian (hr) - 186K **(4.9%)**
- Finnish (fi) - 143K **(3.8%)**
- Swahili (sw) - 141K **(3.7%)**
- Esperanto (eo) - 132K **(3.5%)**
- Spanish (es) - 128K **(3.4%)**
- Italian (it) - 93K **(2.5%)**
- Dutch (nl) - 73K **(1.9%)**
- 156 Others - 519K **(13.7%)**

*Shona (sn) identification might be inaccurate because it is not supported by FastSpell

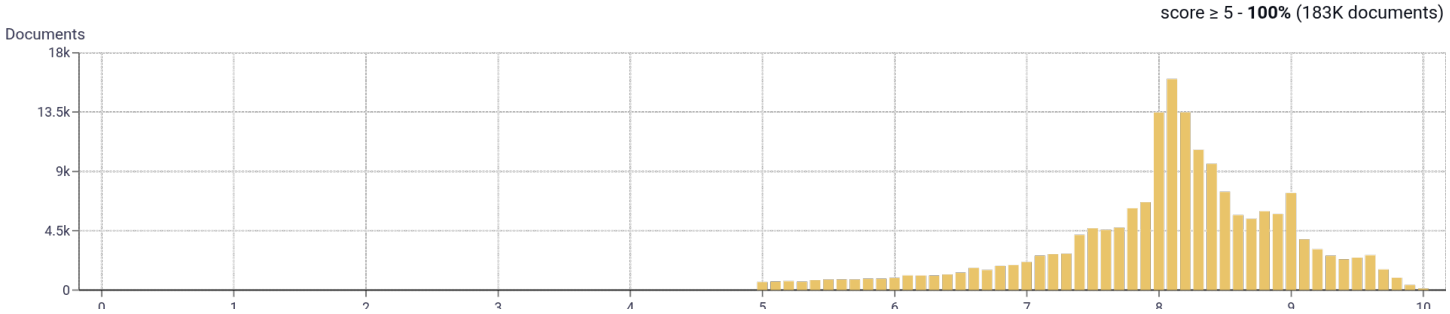
Percentage of segments in Shona (sn) inside documents

segments < 50% - **2.40%** (4.4K documents)
segments ≥ 50% - **97.60%** (179K documents)



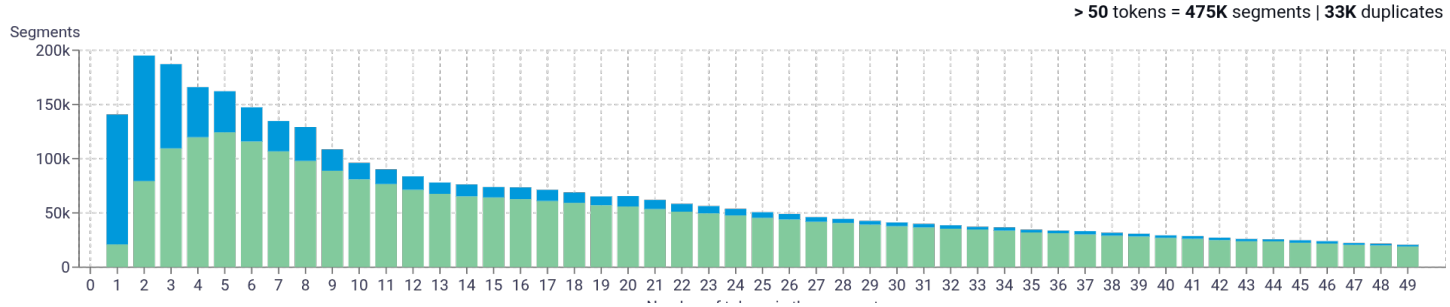
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (183K documents)

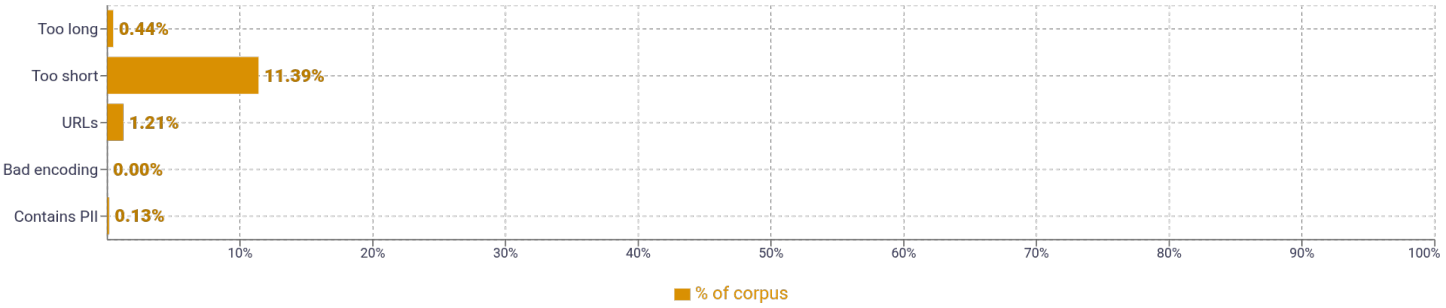


Segment length distribution by token

≤ 49 tokens = **3.3M** segments | **729K** duplicates
> 50 tokens = **475K** segments | **33K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	kana | 923,015 · iyo | 597,966 · kubva | 320,054 · asi | 310,302 · iri | 226,263
2	zviri nyore | 28,445 · makumi maviri | 23,767 · imwe chete | 16,409 · uchinge uchinge | 15,799 · iri nyore | 15,561
3	kana iwe uchida | 19,504 · uchinge uchinge uchinge | 15,155 · iva wekutanga kutaura | 13,543 · kana iwe uri | 9,554 · panguva imwe chete | 7,624
4	uchinge uchinge uchinge uchinge | 14,608 · plus untold biography facts | 2,208 · munguva pfupi iri kutevera | 1,520 · kana iwe uchida kuziva | 1,493 · kana muchida kupinda muchirongwa | 1,368
5	uchinge uchinge uchinge uchinge uchinge | 14,108 · kana muchida kupinda muchirongwa ichi | 1,366 · shanduro yenyika itsva yemagwaro matsvene | 1,138 · munokwanisawo kunzwa studio7 na6 am | 779 · childhood story plus untold biography | 726

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

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	**How-to or instructions**	**HI**	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	**Informational persuasion**	**IP**	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	**Opinion**	**OP**
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	**Informational description**	**IN**	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				