HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-zul_Latn	9/18/2025	Zulu

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
336,440	8,015,560	5,321,756 (66.39 %)	168M	1,113,609,767	1.04 GB

Top 10 domains

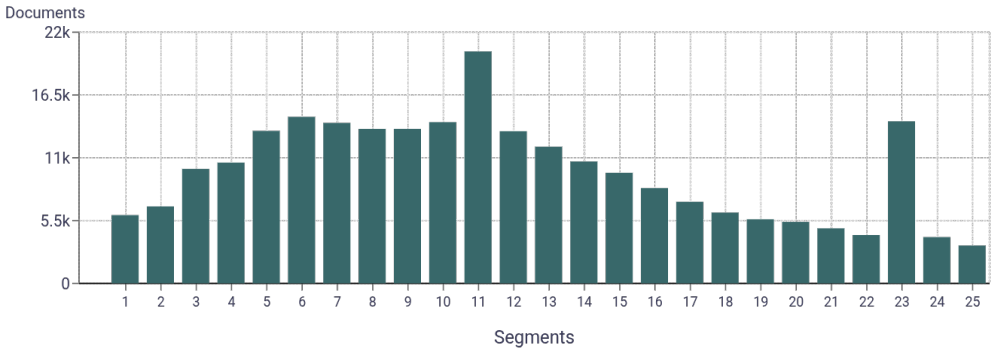
Domain	Docs	% of total
airbnb.com	55K	16.29%
voandebele.com	18K	5.25%
news24.com	14K	4.21%
scrolla.africa	10K	3.04%
unansea.com	9K	2.68%
bayedenews.com	7.3K	2.18%
jw.org	6.4K	1.91%
isolezwe.co.za	6.2K	1.85%
martech.zone	5K	1.50%
impempe.com	4.9K	1.46%

Top 10 TLDs

Domain	Docs	% of total
com	235K	69.73%
co.za	34K	9.98%
org	21K	6.12%
africa	10K	3.05%
net	5.9K	1.74%
zone	5K	1.50%
co.zw	3.1K	0.91%
info	2.2K	0.66%
xyz	1.9K	0.56%
ru	1.3K	0.39%

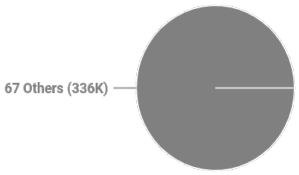
Documents size (in segments)

≤ 25 segments **73.1%** (246K documents)
> 25 segments **26.9%** (91K documents)



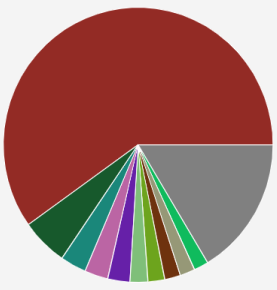
Document collections

CC = 94.85%
IA = 5.15%



Language Distribution

Number of segments in the Zulu corpus

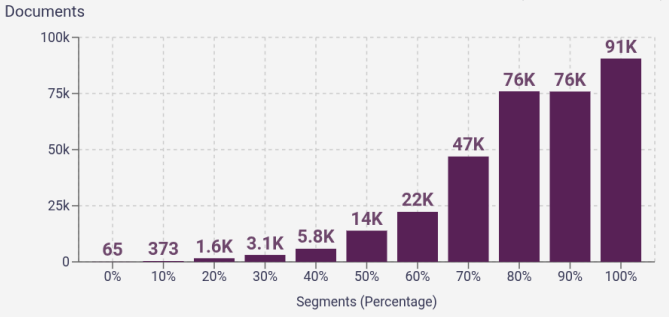


*Zulu identification might be inaccurate because it is not supported by FastSpell.

- English - 4.8M **(59.9%)**
- Filipino - 448K **(5.6%)**
- Croatian - 249K **(3.1%)**
- Italian - 226K **(2.8%)**
- Indonesian - 214K **(2.7%)**
- Swahili - 170K **(2.1%)**
- German - 162K **(2.0%)**
- Turkish - 152K **(1.9%)**
- Polish - 145K **(1.8%)**
- Spanish - 139K **(1.7%)**
- 160 Others - 1.3M **(16.4%)**

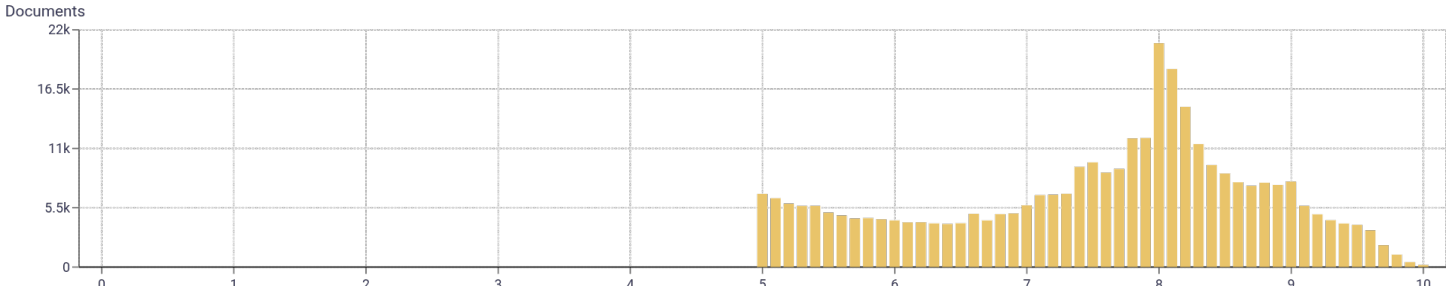
Percentage of segments in Zulu inside documents

segments < 50% - **3.27%** (11K documents)
segments ≥ 50% - **96.73%** (325K documents)



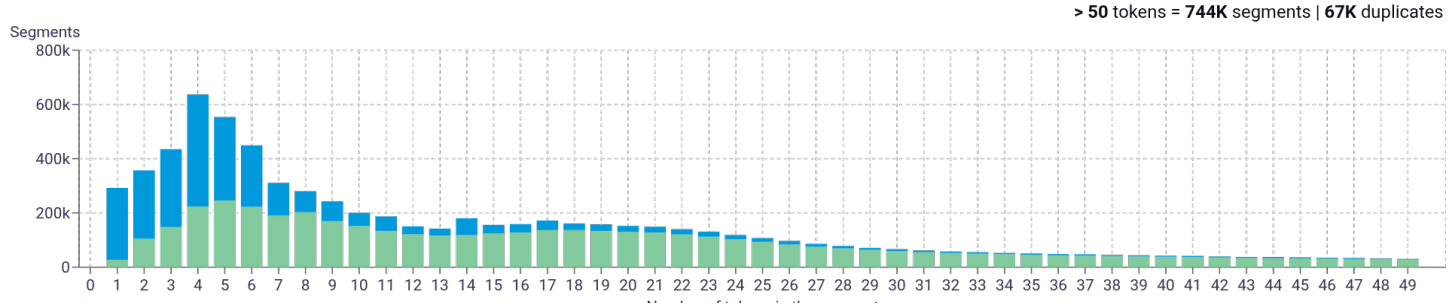
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (336K documents)

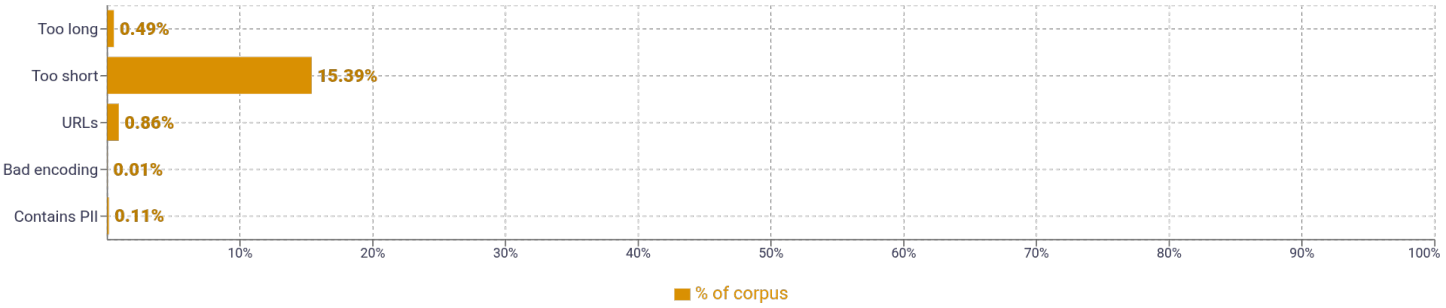


Segment length distribution by token

≤ 49 tokens = **7.3M** segments | **2.6M** duplicates
> 50 tokens = **744K** segments | **67K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	i | 1,659,953 · e | 1,044,287 · okungu | 523,781 · ukuphawula | 521,149 · isilinganiso | 515,187
2	ukuphawula okungu | 506,543 · isilinganiso esingu | 491,609 · ikhaya e | 193,809 · ifulethi e | 100,895 · iyunithi yokuqasha | 83,414
3	iyunithi yokuqasha e | 81,908 · izindawo eziqashwayo zeholide | 54,403 · thola futhi ubhukhe | 37,049 · zeholide ezifanela sonke | 37,023 · ubungako bendawo obukwenele | 37,023
4	zeholide ezifanela sonke isitayela | 37,023 · thola ubungako bendawo obukwenele | 37,023 · izindawo eziqashwayo zeholide ezifanela | 37,023 · eziqashwayo zeholide ezifanela sonke | 37,023 · thola futhi ubhukhe izindawo | 26,650
5	izindawo eziqashwayo zeholide ezifanela sonke | 37,023 · eziqashwayo zeholide ezifanela sonke isitayela | 37,023 · thola futhi ubhukhe izindawo zokuhlala | 17,769 · ubhukhe izindawo zokuhlala eziyingqayivele ku | 17,397 · zindawo zokuhlala zinconywa kakhulu ngokwendawo | 17,394

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				