

General overview

Corpus	Date	Language
hplt-v3-tsn_Latn	9/18/2025	Tswana

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
9,335	224,146	187,408 (83.61 %)	12M	53,338,978	51.28 MB

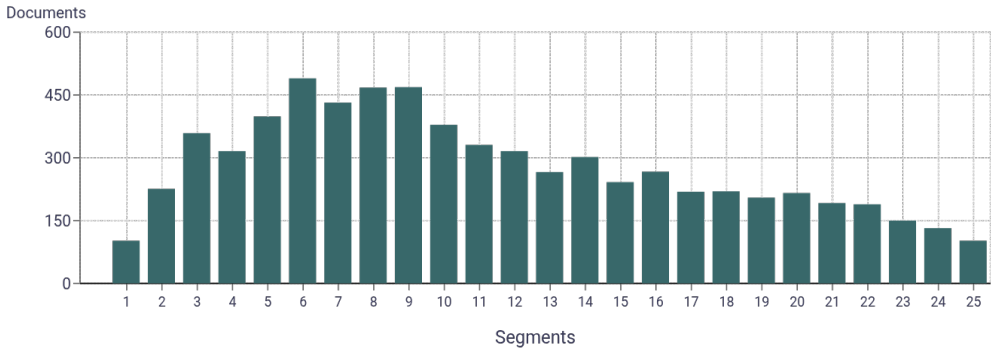
Top 10 domains

Domain	Docs	% of total
jw.org	4.1K	44.33%
southafrica.co.za	785	8.41%
biblesa.co.za	773	8.28%
mmegi.bw	565	6.05%
wikipedia.org	536	5.74%
dailynews.gov.bw	289	3.10%
dikgang24.news	271	2.90%
nwu.ac.za	176	1.89%
kutlwano.gov.bw	110	1.18%
www.gov.bw	93	1.00%

Top 10 TLDs

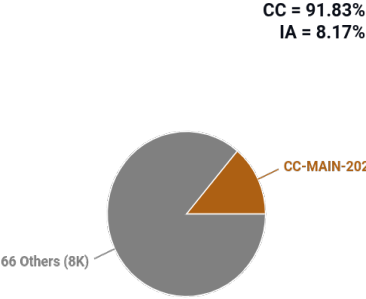
Domain	Docs	% of total
org	5K	53.80%
co.za	1.8K	19.81%
bw	566	6.06%
gov.bw	495	5.30%
com	465	4.98%
news	271	2.90%
ac.za	190	2.04%
gov.za	59	0.63%
net	55	0.59%
fm	44	0.47%

Documents size (in segments) ⓘ



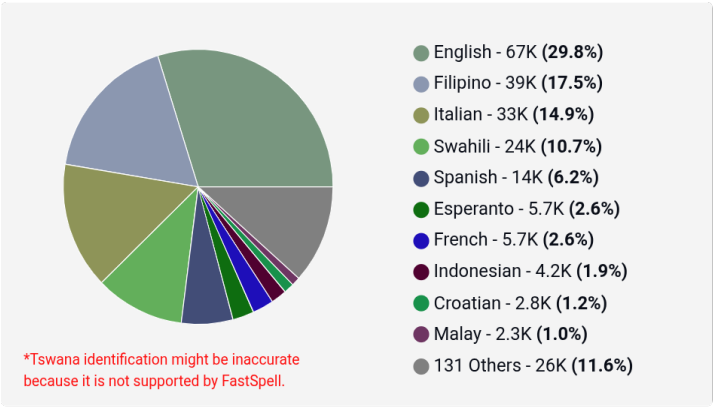
≤ 25 segments 74.87% (7K documents)  
> 25 segments 25.13% (2.3K documents)

Document collections

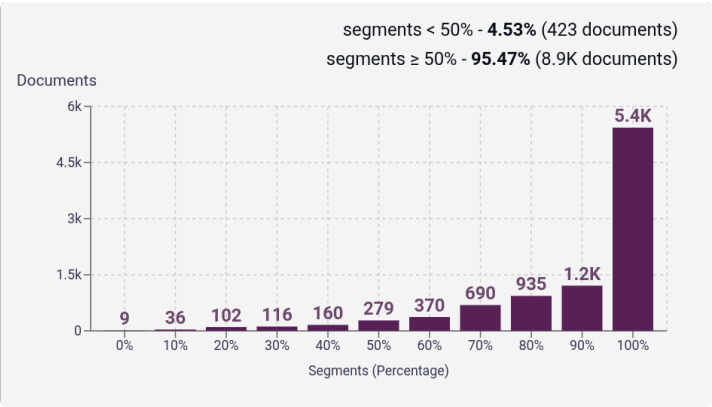


Language Distribution

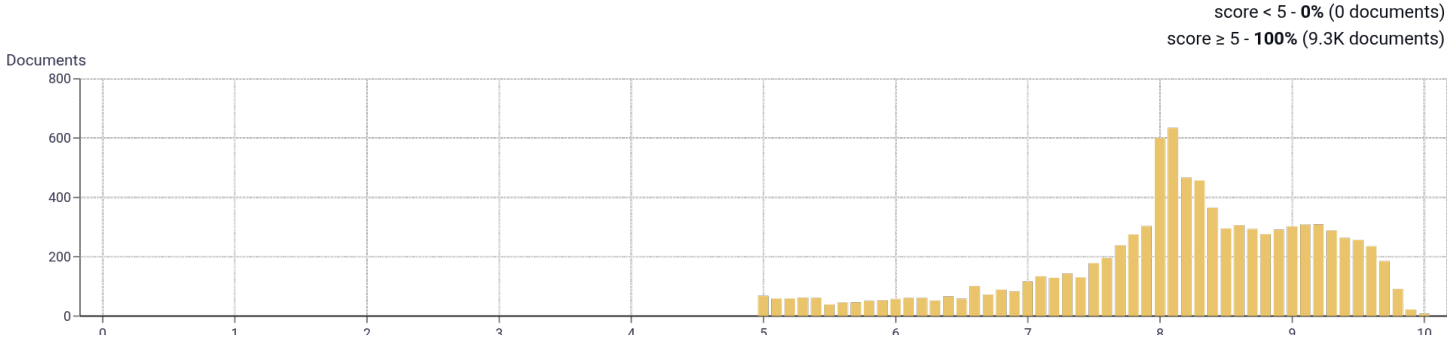
Number of segments in the Tswana corpus



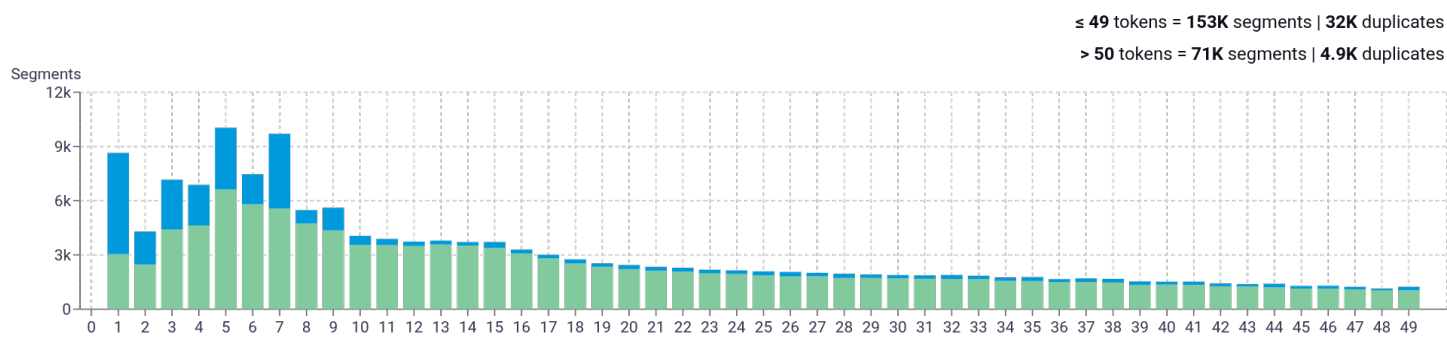
Percentage of segments in Tswana inside documents



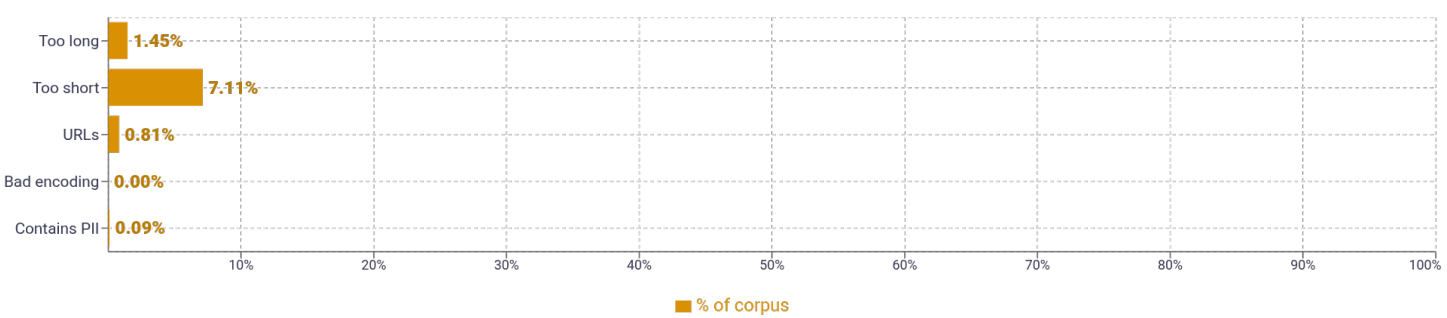
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>a   447,009</div> <div>ya   232,732</div> <div>wa   97,093</div> <div>tse   86,389</div> <div>mme   81,956</div>	
2	<div>a a   22,159</div> <div>jo bo   11,235</div> <div>neng a   11,192</div> <div>ya gagwe   8,815</div> <div>ntlha ya   8,627</div>	
3	<div>a ne a   9,620</div> <div>a neng a   7,370</div> <div>a bo a   3,972</div> <div>tse di molemo   2,596</div> <div>a ba a   2,088</div>	
4	<div>basupi ba ga jehofa   3,632</div> <div>jesu o ne a   3,411</div> <div>a se ka a   2,147</div> <div>jehofa o ne a   2,039</div> <div>batho ba le bantsi   2,028</div>	
5	<div>botshelo jo bo sa khutleng   669</div> <div>ya basupi ba ga jehofa   472</div> <div>moaposetoloi paulo o ne a   431</div> <div>thanolo ya lefatshe le lesa   393</div> <div>archived from the original on   336</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  $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$ , 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	ntp
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				