

General overview

Corpus	Date	Language
hplt-v3-ssw_Latn	9/18/2025	Swati

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
2,789	94,956	58,597 (61.71 %)	2.1M	14,946,310	14.37 MB

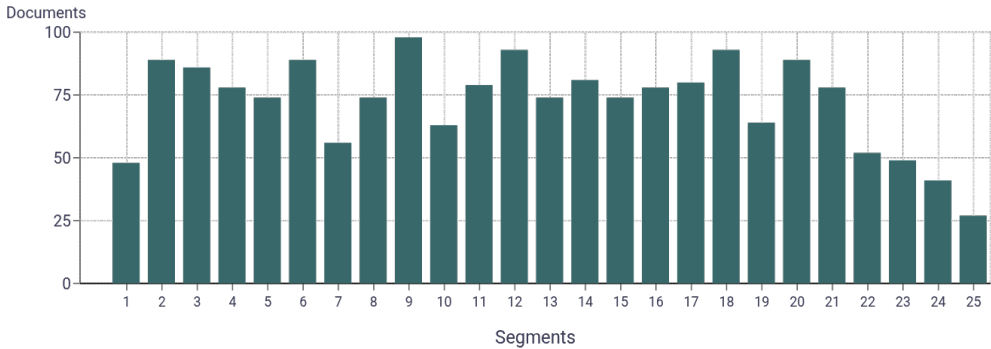
Top 10 domains

Domain	Docs	% of total
biblesa.co.za	887	31.80%
southafrica.co.za	795	28.50%
jw.org	381	13.66%
wikipedia.org	321	11.51%
nalibali.org	32	1.15%
myconstitution....	21	0.75%
www.gov.za	19	0.68%
shuters.co.za	14	0.50%
pansalb.org	12	0.43%
izithakazelo.blog	12	0.43%

Top 10 TLDs

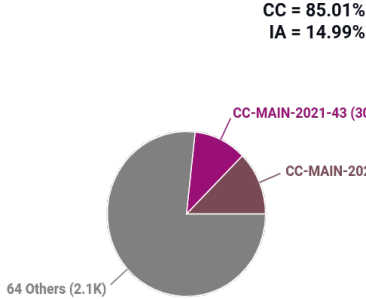
Domain	Docs	% of total
co.za	1.8K	64.54%
org	806	28.90%
com	56	2.01%
gov.za	37	1.33%
org.za	14	0.50%
mobi	14	0.50%
net	12	0.43%
blog	12	0.43%
co.uk	8	0.29%
fm	7	0.25%

Documents size (in segments) ⓘ



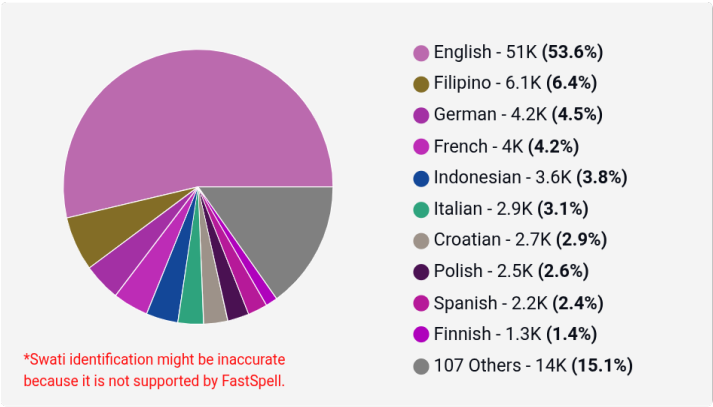
≤ 25 segments **64.79%** (1.8K documents)
> 25 segments **35.21%** (982 documents)

Document collections

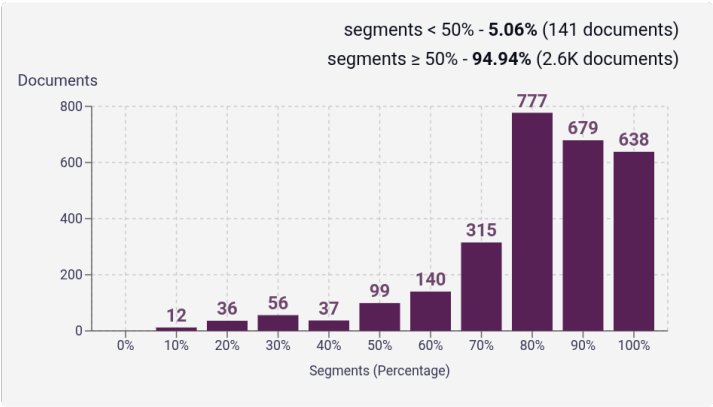


Language Distribution

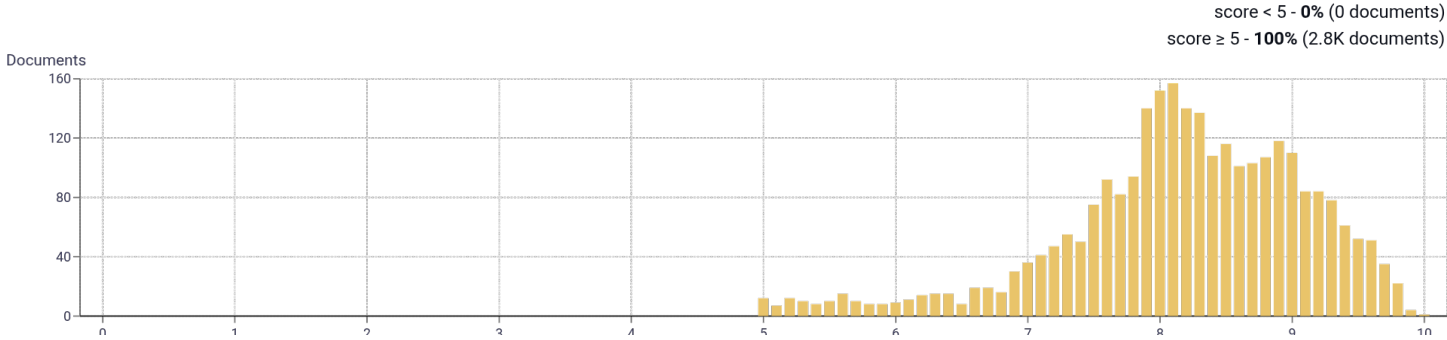
Number of segments in the Swati corpus



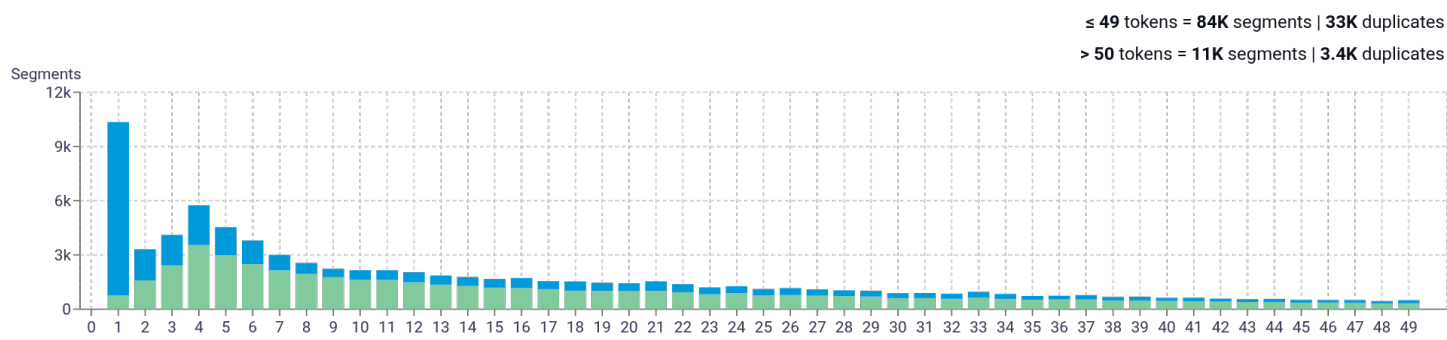
Percentage of segments in Swati inside documents



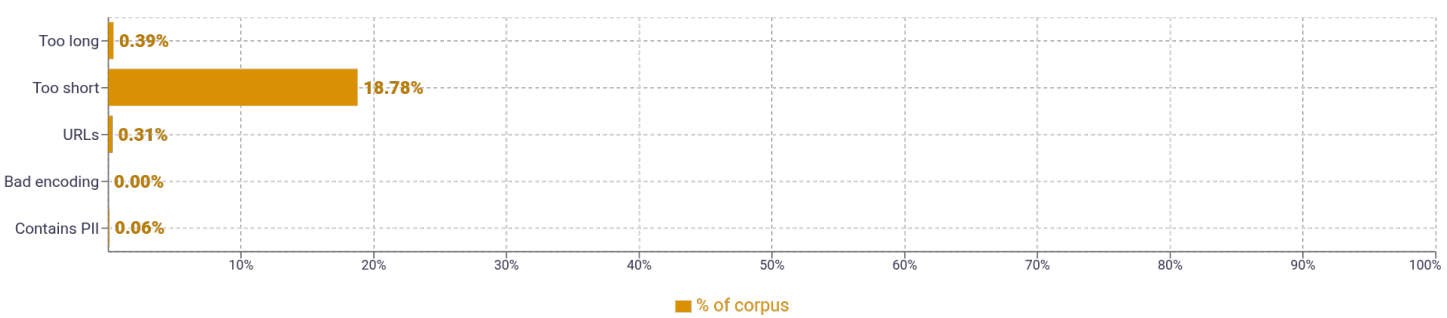
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	ke 6,469 nome 5,955 israyeli 5,629 ngobe 5,464 nkulunkulu 5,410	
2	bonkhe bantfu 850 simakadze nkulunkulu 768 nkulunkulu wenu 732 ningizimu afrika 709 eningizimu afrika 665	
3	simakadze nkulunkulu wenu 462 naku lokushiwo ngusimakadze 296 letinye tincwadzi tekucala 273 naku lokushiwo yinkhosi 258 lokushiwo yinkhosi simakadze 250	
4	naku lokushiwo yinkhosi simakadze 243 translated by phindile malotana 174 letinye tincwadzi tekucala titsi 169 wekucinisekisa emazinga kutemfundvo nekucecesha 114 liphunga lelinuka lusi lolumnandzi 90	
5	liphunga lelinuka lusi lolumnandzi kusimakadze 62 world translation of the holy 53 translation of the holy scriptures 53 new world translation of the 53 we are working towards a 52	

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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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>, , , 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				