

General overview

Corpus	Date	Language
hplt-v3-tso_Latn	9/18/2025	Tsonga

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
11,680	292,332	236,361 (80.85 %)	12M	57,759,670	56.08 MB

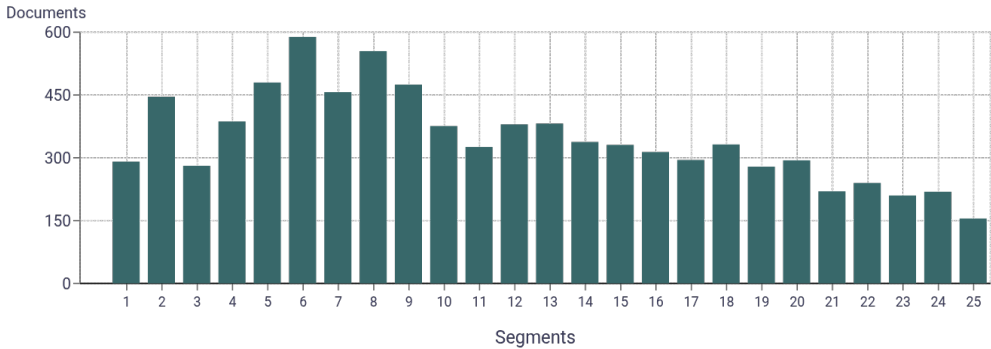
Top 10 domains

Domain	Docs	% of total
jw.org	4.9K	41.84%
southafrica.co.za	1.5K	12.79%
biblesa.co.za	1.5K	12.72%
bible.is	452	3.87%
wikipedia.org	427	3.66%
nthavela.co.za	377	3.23%
limpopomirror.c...	282	2.41%
vivmag.co.za	210	1.80%
tsongapop.co.za	131	1.12%
m2819.co.za	116	0.99%

Top 10 TLDs

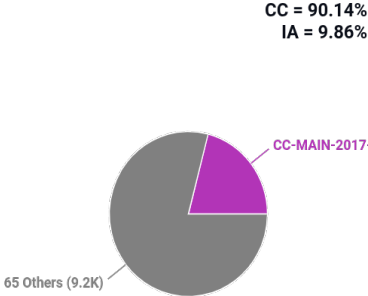
Domain	Docs	% of total
org	5.6K	48.20%
co.za	4.5K	38.82%
com	487	4.17%
is	452	3.87%
gov.za	123	1.05%
fm	118	1.01%
ru	70	0.60%
net	64	0.55%
org.za	40	0.34%
co.zw	40	0.34%

Documents size (in segments) ⓘ



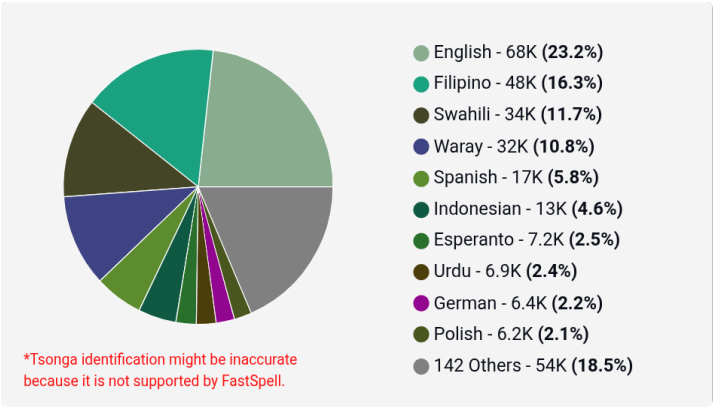
≤ 25 segments **74.08%** (8.7K documents)  
> 25 segments **25.92%** (3K documents)

Document collections

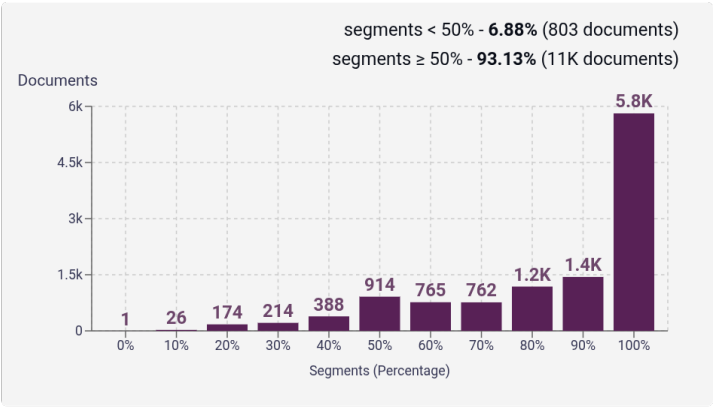


Language Distribution

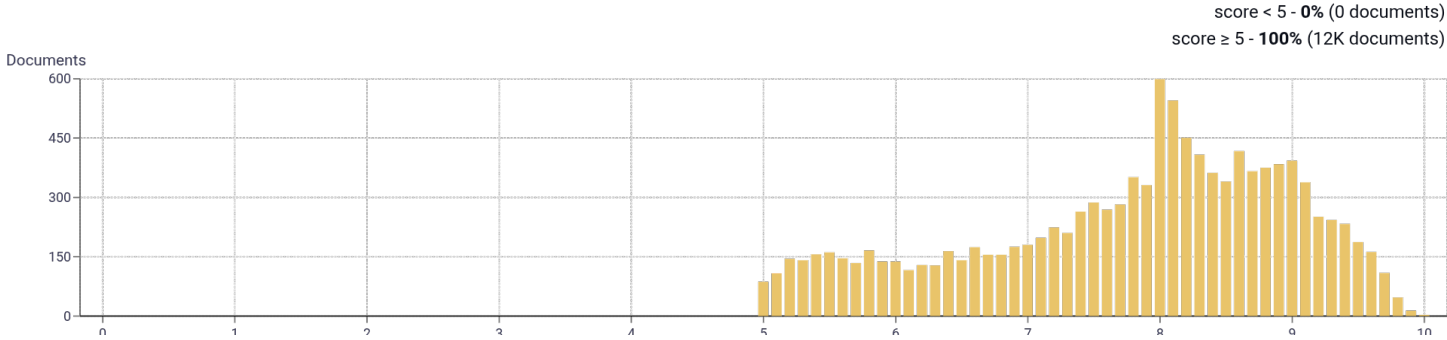
Number of segments in the Tsonga corpus



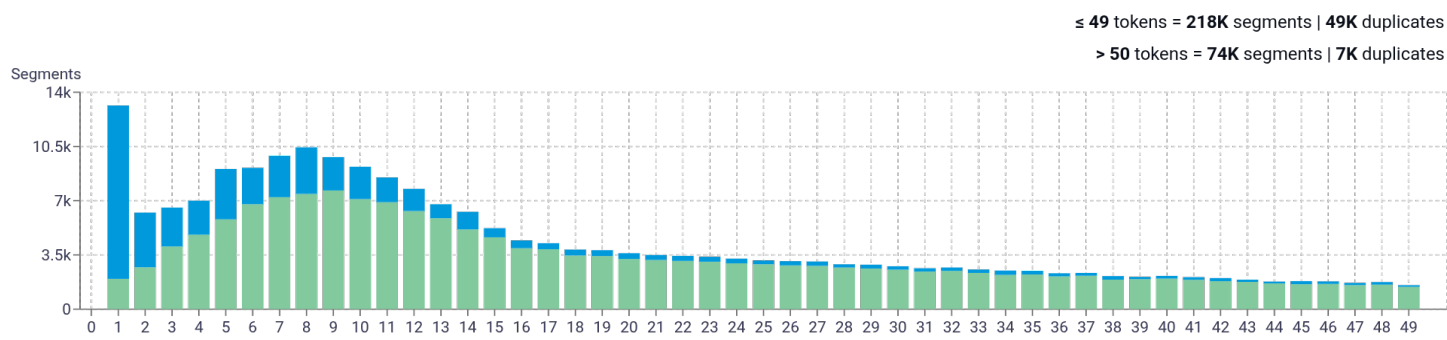
Percentage of segments in Tsonga inside documents



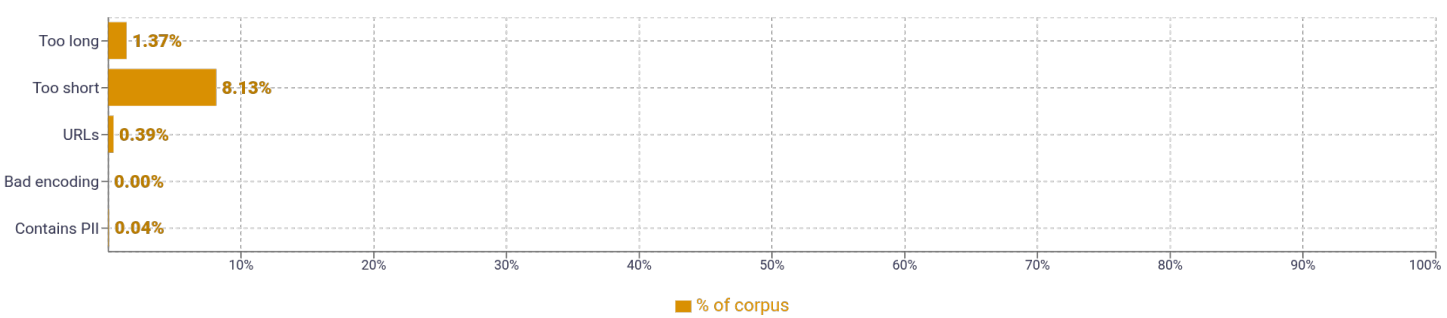
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	ni   174,540vha   100,266ha   67,959wana   60,882n   54,195	
2	ndlela leyi   9,445vha tshi   8,626ha yona   7,276tshi khou   6,176wana ni   6,085	
3	wana ni un   2,477timbhoni ta yehovha   2,470vha a tshi   1,877vanhu vo tala   1,873vha vha tshi   1,852	
4	o vha a tshi   1,269vho vha vha tshi   1,233vuhundzuluxeri bya misava leyintshwa   1,200hilaha ku nga heriki   1,107bya misava leyintshwa bya   1,044	
5	vuhundzuluxeri bya misava leyintshwa bya   1,044bya misava leyintshwa bya matsalwa   1,039misava leyintshwa bya matsalwa yo   1,026leyintshwa bya matsalwa yo kwetsima   1,026bibele eka web site leyi   961	

# 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				