HPLT Analytics report

HPLTAnalytics

General overview

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

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%

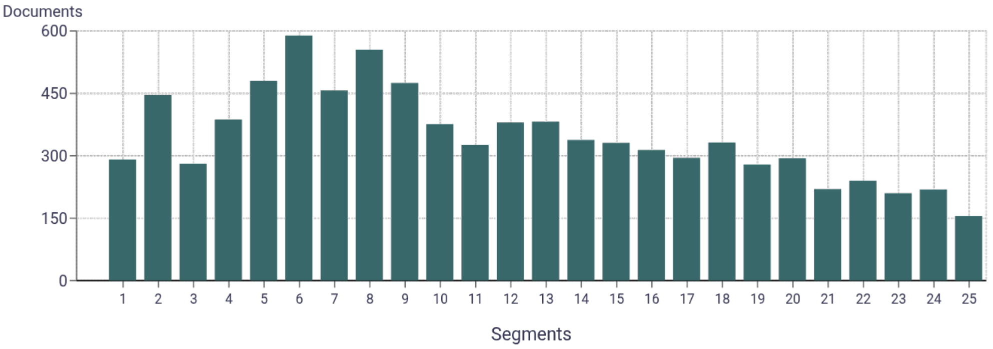
Top 10 TLDs

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)

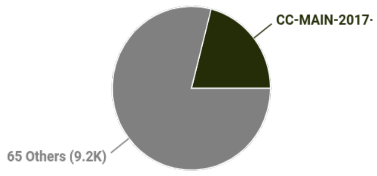
≤ 25 segments **74.08%** (8.7K documents)

> 25 segments 25.92% (3K documents)



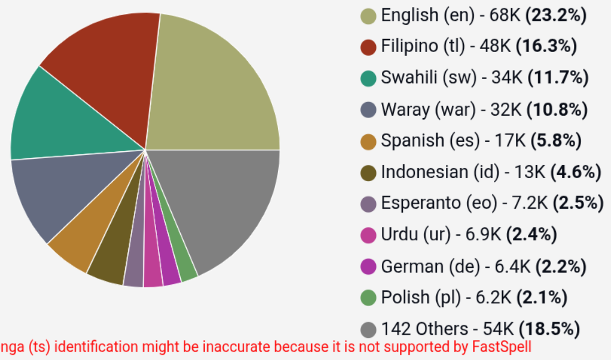
Document collections

CC = 90.14%
IA = 9.86%



Language Distribution

Number of segments in the Tsonga (ts) corpus

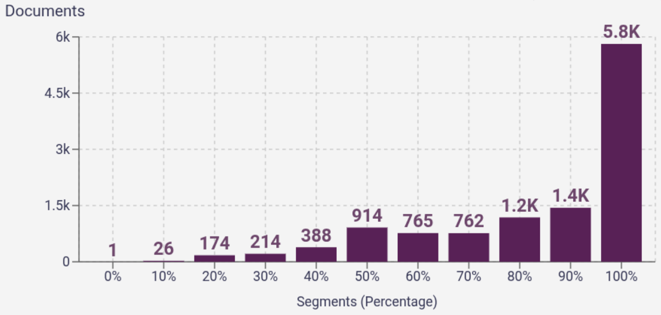


*Tsonga (ts) identification might be inaccurate because it is not supported by FastSpell

Percentage of segments in Tsonga (ts) inside documents

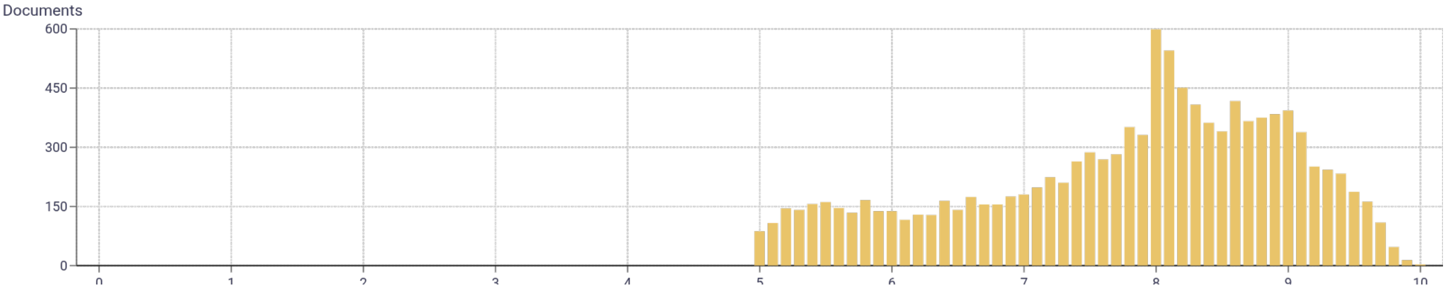
segments < 50% - **6.88%** (803 documents)

segments ≥ 50% - **93.13%** (11K documents)



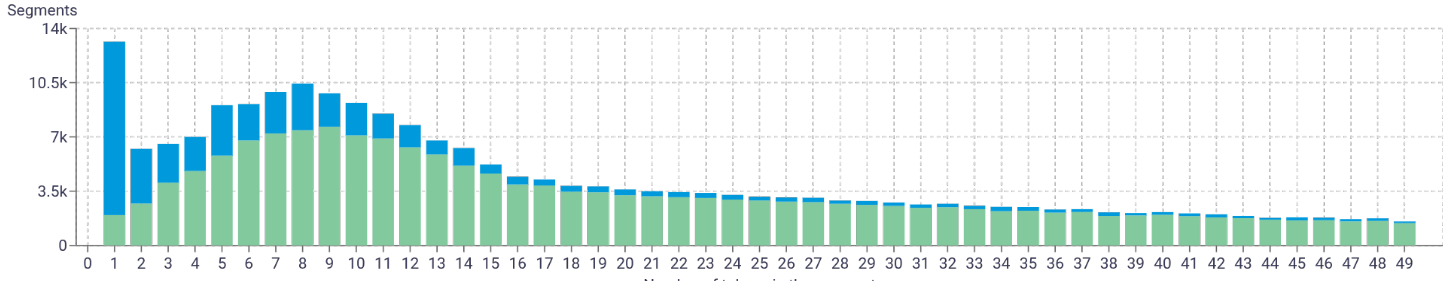
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (12K documents)

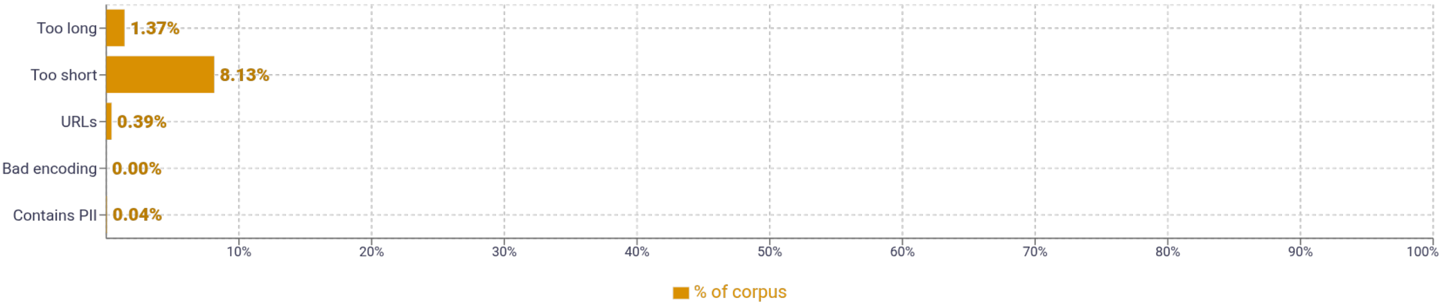


Segment length distribution by token

≤ **49** tokens = **218K** segments | **49K** duplicates
> **50** tokens = **74K** segments | **7K** duplicates



Segment noise distribution



Frequent n-grams

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