HPLT Analytics report

General overview

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

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

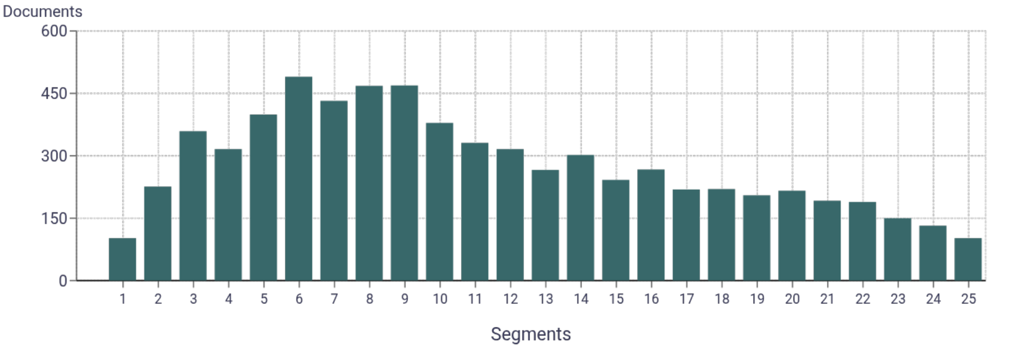
Domain	Docs	% of total
jw.org	4.1K	44.33%
southafrica.co.za	785	8.41%
biblesa.co.za	773	8.28%
mmegi.bw	565	6.05%
wikipedia.org	536	5.74%
dailynews.gov.bw	289	3.10%
dikgang24.news	271	2.90%
nwu.ac.za	176	1.89%
kutlwano.gov.bw	110	1.18%
www.gov.bw	93	1.00%

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%

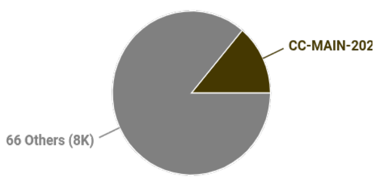
Documents size (in segments) ⓘ

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



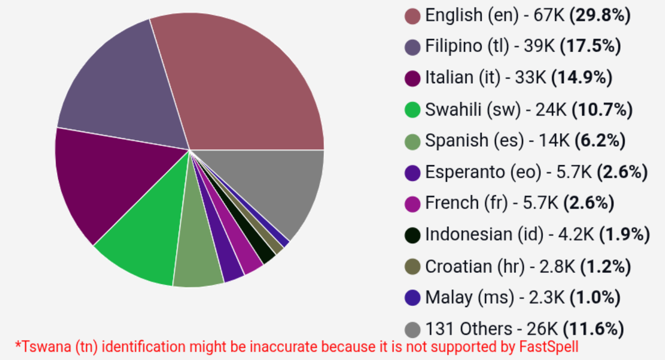
Document collections

CC = 91.83%
IA = 8.17%



Language Distribution

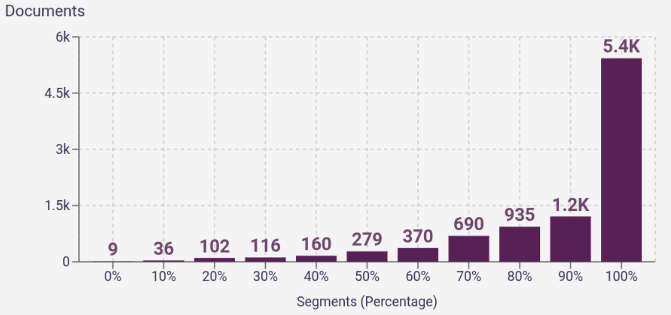
Number of segments in the Tswana (tn) corpus



*Tswana (tn) identification might be inaccurate because it is not supported by FastSpell

Percentage of segments in Tswana (tn) inside documents

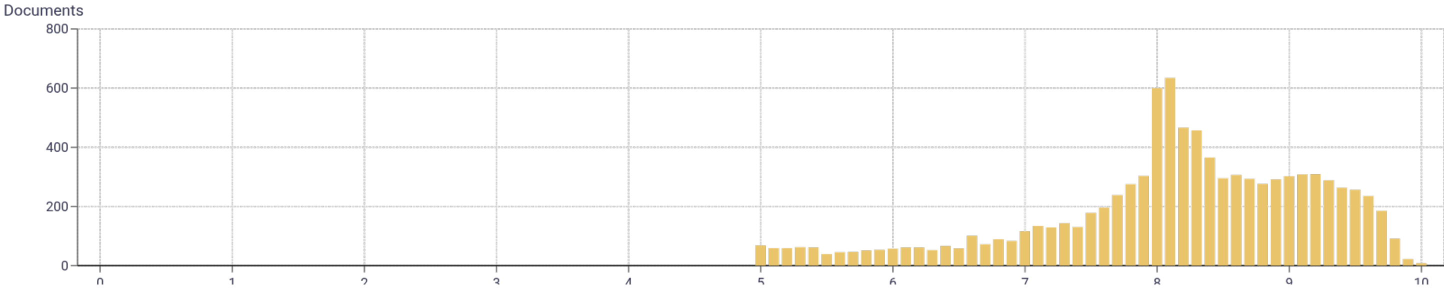
segments < 50% - **4.53%** (423 documents)
segments ≥ 50% - **95.47%** (8.9K documents)



Distribution of documents by document score

score < 5 - **0%** (0 documents)

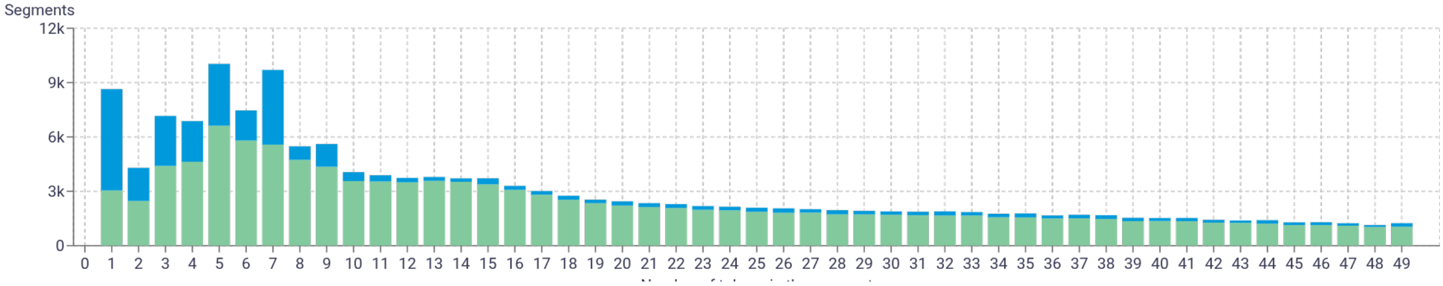
score ≥ 5 - **100%** (9.3K documents)



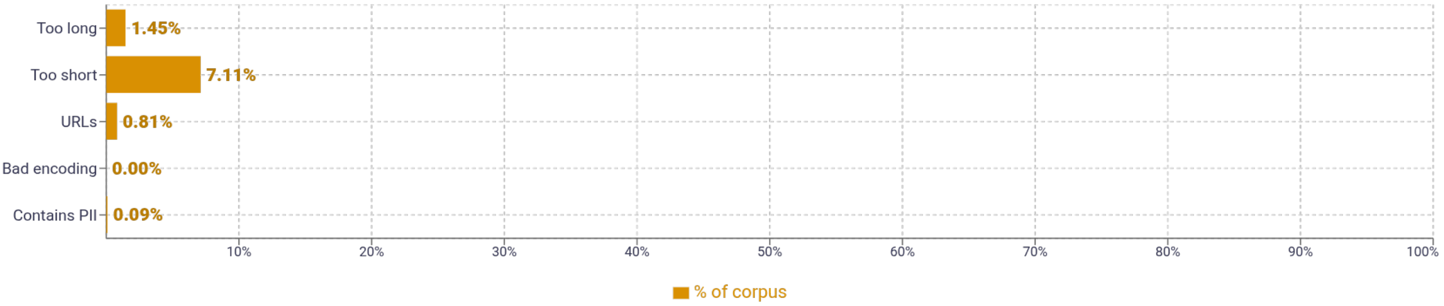
Segment length distribution by token

≤ **49** tokens = **153K** segments | **32K** duplicates

> **50** tokens = **71K** segments | **4.9K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	a | 447,009 ya | 232,732 wa | 97,093 tse | 86,389 mme | 81,956
2	a a | 22,159 jo bo | 11,235 neng a | 11,192 ya gagwe | 8,815 ntlha ya | 8,627
3	a ne a | 9,620 a neng a | 7,370 a bo a | 3,972 tse di molemo | 2,596 a ba a | 2,088
4	basupi ba ga jehofa | 3,632 jesu o ne a | 3,411 a se ka a | 2,147 jehofa o ne a | 2,039 batho ba le bantsi | 2,028
5	botshelo jo bo sa khutleng | 669 ya basupi ba ga jehofa | 472 moaposetoloi paulo o ne a | 431 thanolo ya lefatshe le lesha | 393 archived from the original on | 336

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				