HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-nso_Latn	9/18/2025	Northern Sotho (nso)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
8,183	234,057	172,408 (73.66 %)	9.3M	42,025,444	41.23 MB

Top 10 domains

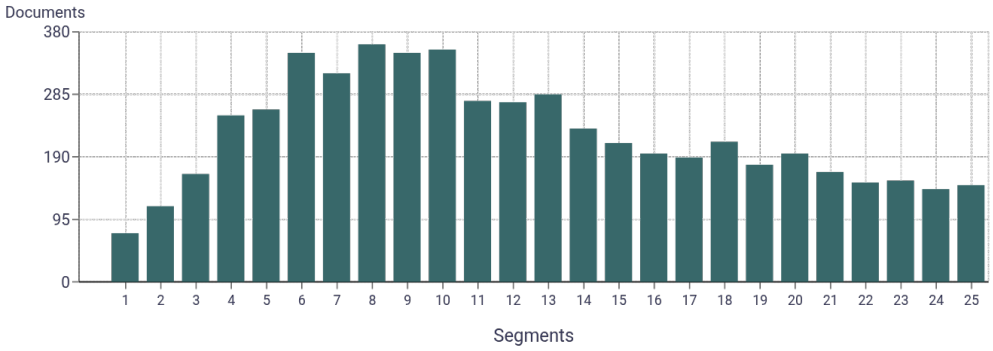
Domain	Docs	% of total
jw.org	3.7K	45.35%
biblesa.co.za	1.8K	22.30%
southafrica.co.za	806	9.85%
fundza.mobi	185	2.26%
seiponemadireng...	159	1.94%
nalibali.org	115	1.41%
wikipedia.org	93	1.14%
oxforddictionar...	81	0.99%
sars.gov.za	54	0.66%
sekhukhunetimes...	47	0.57%

Top 10 TLDs

Domain	Docs	% of total
org	4K	49.40%
co.za	3.1K	38.47%
com	385	4.70%
mobi	186	2.27%
gov.za	88	1.08%
org.za	48	0.59%
ru	46	0.56%
net	46	0.56%
fm	41	0.50%
ac.za	28	0.34%

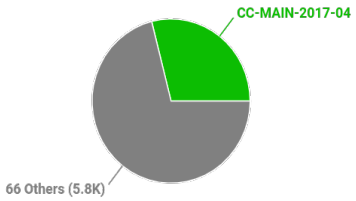
Documents size (in segments)

≤ 25 segments **68.46%** (5.6K documents)
> 25 segments **31.54%** (2.6K documents)



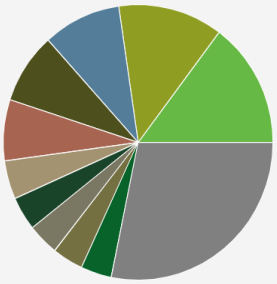
Document collections

CC = 97.03%
IA = 2.97%



Language Distribution

Number of segments in the Northern Sotho (nso) corpus

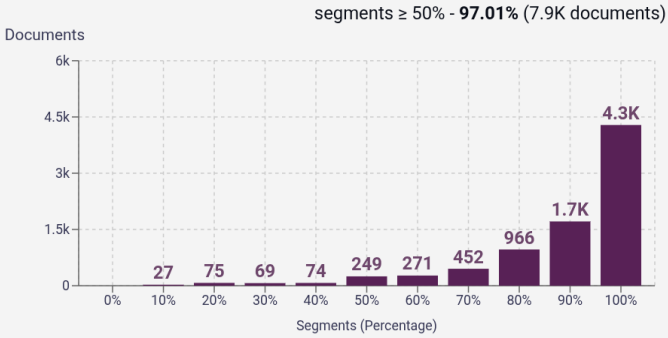


- English (en) - 35K **(14.8%)**
- Filipino (tl) - 29K **(12.6%)**
- Swahili (sw) - 22K **(9.3%)**
- Slovenian (sl) - 19K **(8.3%)**
- Finnish (fi) - 17K **(7.2%)**
- Spanish (es) - 11K **(4.5%)**
- Cornish (kw) - 9.3K **(4.0%)**
- Italian (it) - 8.7K **(3.7%)**
- French (fr) - 8.7K **(3.7%)**
- Croatian (hr) - 8.7K **(3.7%)**
- 136 Others - 66K **(28.2%)**

*Northern Sotho (nso) identification might be inaccurate because it is not supported by FastSpell

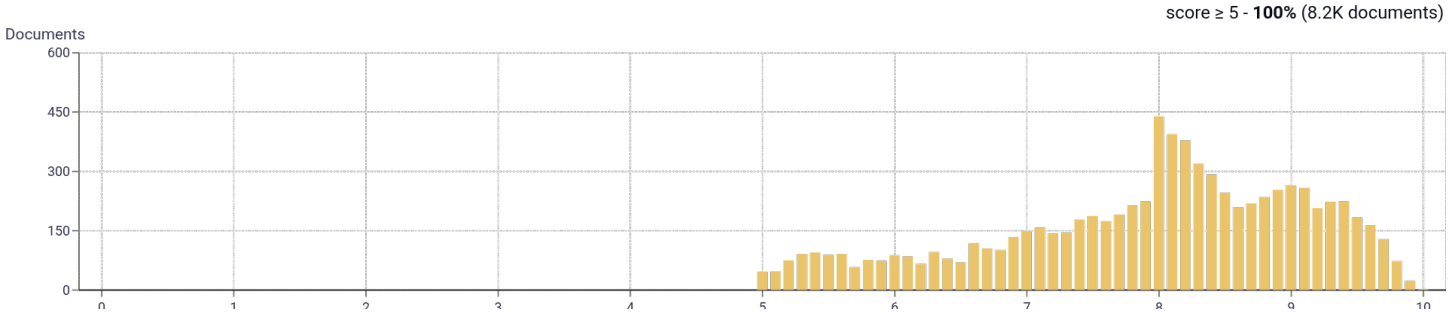
Percentage of segments in Northern Sotho (nso) inside documents

segments < 50% - **2.99%** (245 documents)
segments ≥ 50% - **97.01%** (7.9K documents)



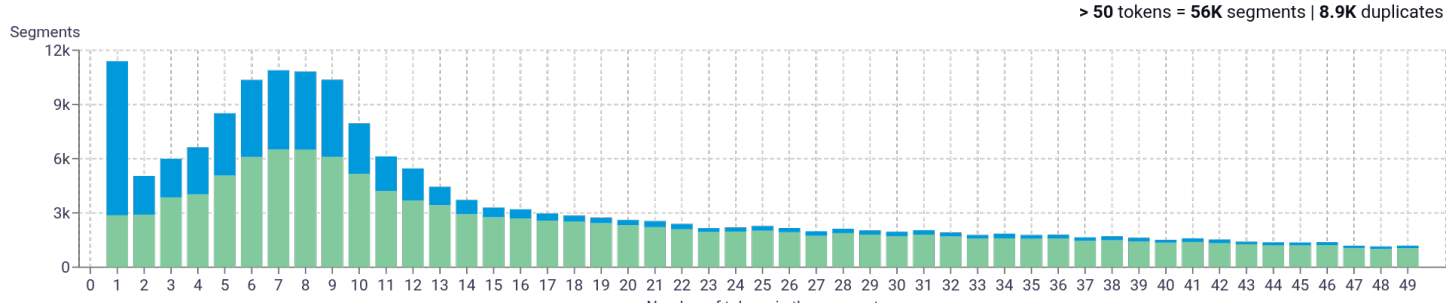
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (8.2K documents)

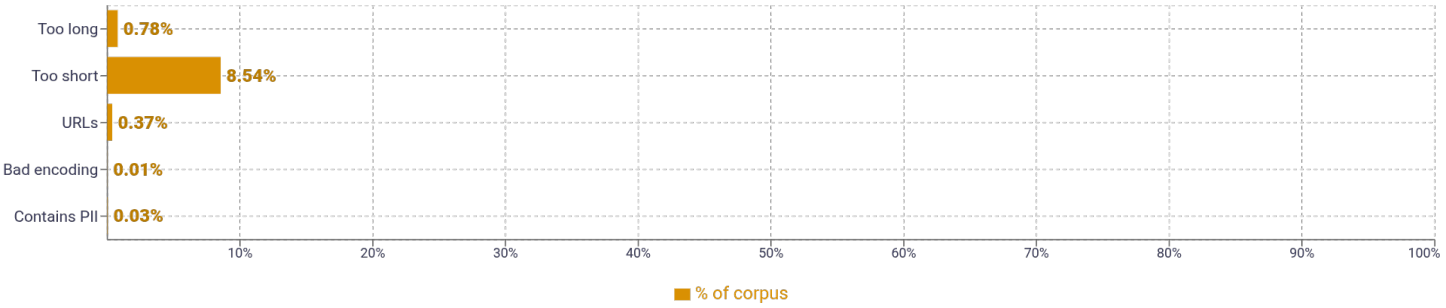


Segment length distribution by token

≤ 49 tokens = **178K** segments | **53K** duplicates
> 50 tokens = **56K** segments | **8.9K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	ya | 201,284 · wa | 107,305 · re | 105,061 · tša | 88,829 · ge | 79,792
2	yo mongwe | 10,205 · wa gagwe | 8,171 · ya gagwe | 7,220 · tše dingwe | 7,116 · baka la | 6,459
3	dihlatse tša jehofa | 2,771 · yo mongwe wa | 2,523 · mongwe le yo | 2,505 · bjalo ka ge | 2,388 · bao ba bego | 2,300
4	lega go le bjalo | 3,486 · mongwe le yo mongwe | 2,500 · yo mongwe le yo | 2,453 · ge e le gabotse | 1,411 · sengwe le se sengwe | 1,259
5	yo mongwe le yo mongwe | 2,448 · phetolelo ya lefase le lefsa | 1,168 · ya lefase le lefsa ya | 1,030 · lefase le lefsa ya mangwalo | 1,017 · lefsa ya mangwalo a makgethwa | 1,004

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				