HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-sot_Latn	9/18/2025	Southern Sotho (st)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
152,062	3,618,844	2,830,790 (78.22 %)	128M	601,224,893	576.39 MB

Top 10 domains

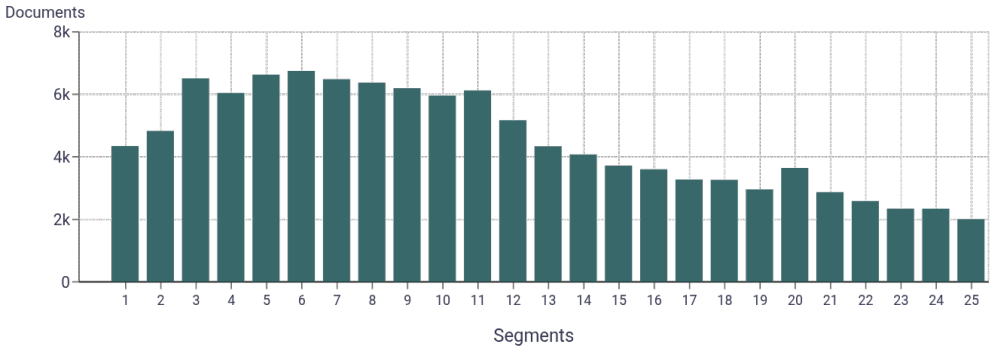
Domain	Docs	% of total
eturbonews.com	5.5K	3.62%
martech.zone	5.3K	3.46%
jw.org	3.2K	2.08%
actualidadgadge...	2K	1.33%
bitemybun.com	1.9K	1.23%
comme-un-pro.fr	1.4K	0.93%
actualidadviaje...	1.4K	0.91%
actualidadiphon...	1.3K	0.84%
desdelinux.net	1.2K	0.78%
hombresconestil...	1.2K	0.76%

Top 10 TLDs

Domain	Docs	% of total
com	116K	75.99%
org	7.7K	5.08%
zone	5.3K	3.46%
net	3.8K	2.49%
co.za	2.5K	1.67%
ru	2.3K	1.54%
fr	1.5K	0.96%
news	1.3K	0.84%
info	1.2K	0.81%
co.ls	947	0.62%

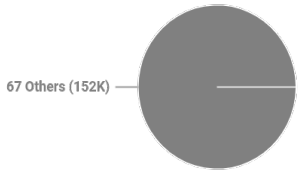
Documents size (in segments)

≤ 25 segments **73.94%** (112K documents)
> 25 segments **26.06%** (40K documents)



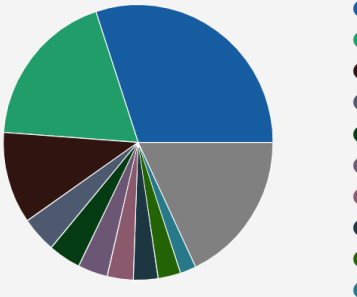
Document collections

CC = 95.63%
IA = 4.37%



Language Distribution

Number of segments in the Southern Sotho (st) corpus

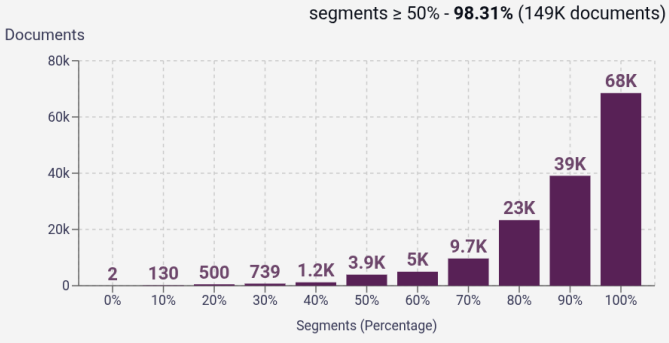


- English (en) - 1.1M **(30.1%)**
- Filipino (tl) - 679K **(18.8%)**
- Italian (it) - 390K **(10.8%)**
- French (fr) - 153K **(4.2%)**
- Croatian (hr) - 140K **(3.9%)**
- Spanish (es) - 129K **(3.6%)**
- German (de) - 113K **(3.1%)**
- Indonesian (id) - 107K **(3.0%)**
- Breton (br) - 98K **(2.7%)**
- Finnish (fi) - 69K **(1.9%)**
- 155 Others - 651K **(18.0%)**

*Southern Sotho (st) identification might be inaccurate because it is not supported by FastSpell

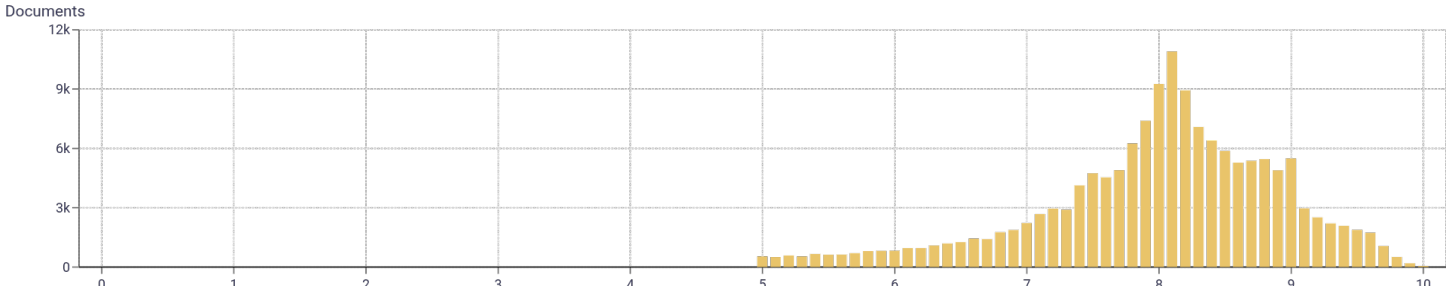
Percentage of segments in Southern Sotho (st) inside documents

segments < 50% - **1.69%** (2.6K documents)
segments ≥ 50% - **98.31%** (149K documents)



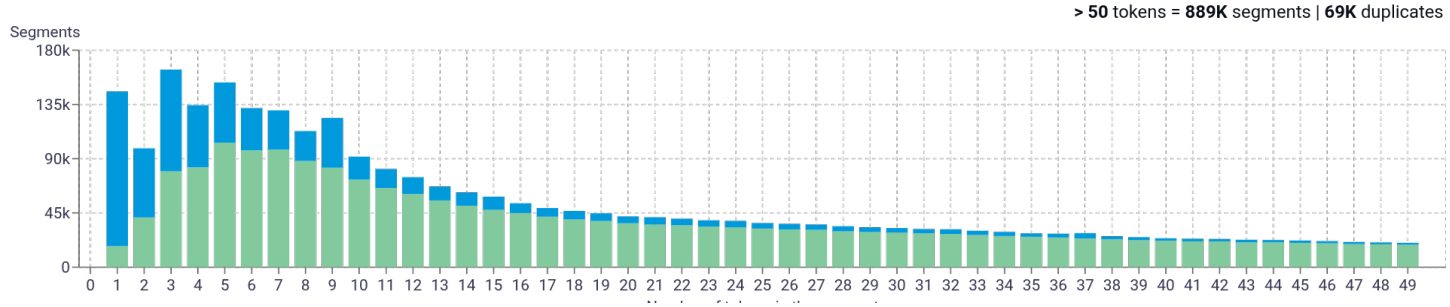
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (152K documents)

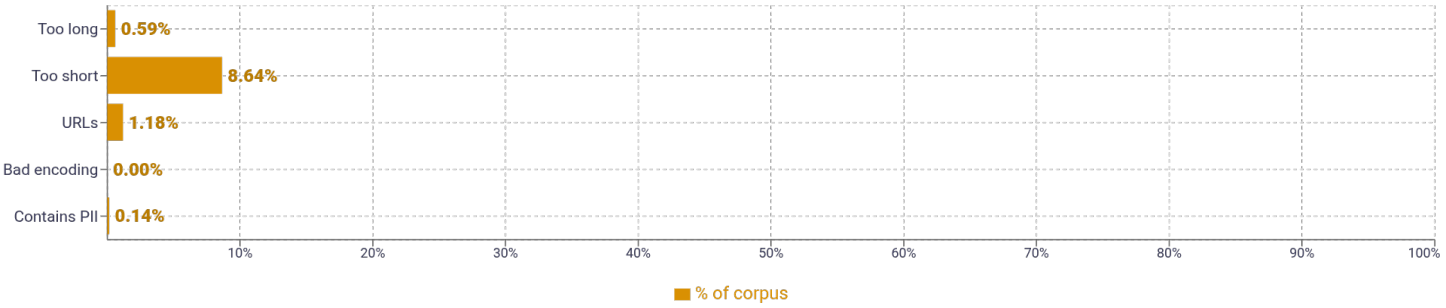


Segment length distribution by token

≤ 49 tokens = **2.7M** segments | **721K** duplicates
> 50 tokens = **889K** segments | **69K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	u | 927,095 · na | 729,609 · kapa | 527,249 · tla | 516,190 · bakeng | 436,607
2	haeba u | 87,588 · boleng bo | 52,469 · u tla | 47,777 · bo botle | 42,100 · hona joale | 41,541
3	nako e telele | 40,078 · efe kapa efe | 27,448 · boleng bo holimo | 26,495 · bophelo bo botle | 25,470 · letsatsi le letsatsi | 18,805
4	leha ho le joalo | 56,230 · mong le e mong | 26,935 · ntle ka ho fetisisa | 19,698 · u se ke ua | 18,470 · molemo ka ho fetisisa | 16,737
5	pele ho fana ka maikutlo | 10,018 · town town town town town | 8,714 · etsa bonnete ba hore u | 4,841 · sebaka sa hau sa marang | 4,021 · na le mefuta e mengata | 3,430

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				