HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
afr_Latn.json.tsv	9/6/2024	Afrikaans (af)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,457,165	37,737,319	18,802,427 (49.82 %)	1.2B	5,910,906,748	5.56 GB

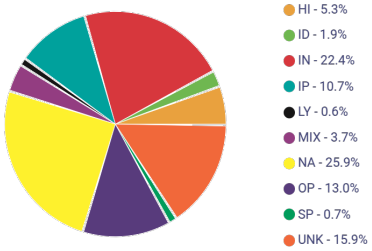
Top 10 domains

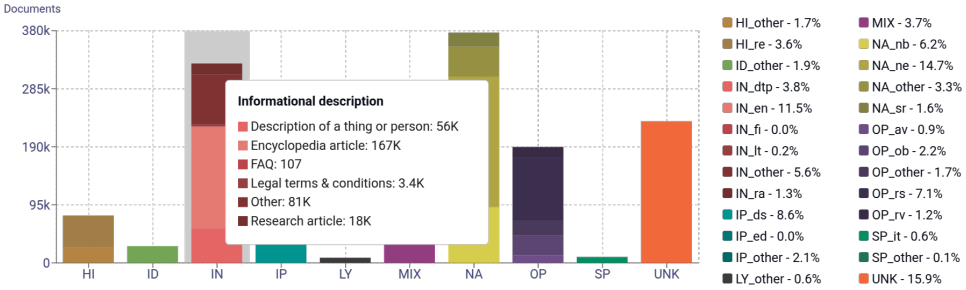
Domain	Docs	% of total
wikipedia.org	164K	11.23%
maroelamedia.co.za	68K	4.64%
netwerk24.com	42K	2.87%
landbou.com	35K	2.38%
praag.co.za	33K	2.24%
wordpress.com	32K	2.19%
litnet.co.za	32K	2.17%
sarie.com	28K	1.90%
softoware.net	20K	1.40%
androware.net	20K	1.39%

Top 10 TLDs

Domain	Docs	% of total
com	529K	36.33%
co.za	451K	30.96%
org	239K	16.38%
net	69K	4.74%
org.za	30K	2.06%
ac.za	28K	1.90%
com.na	19K	1.27%
ca	11K	0.76%
info	9.3K	0.64%
pt	4K	0.28%

Register labels

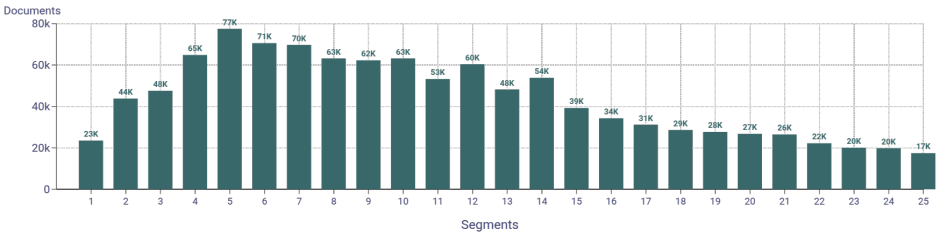




MT:12.8% | 186K Documents

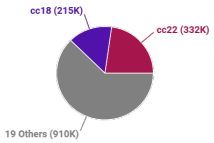
Documents size (in segments)

<= 25 segments **75.18%** (1.1M documents)
> 25 segments **24.82%** (362K documents)



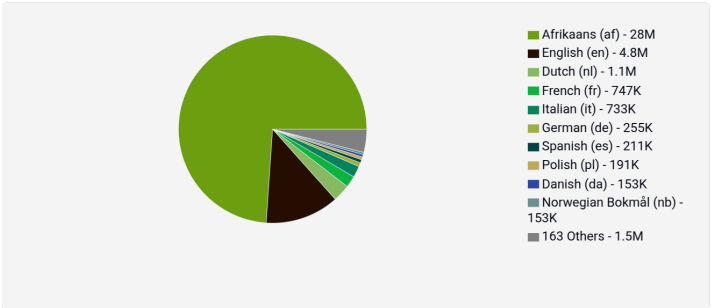
Documents by collection

CC = 59.43%
IA = 40.57%

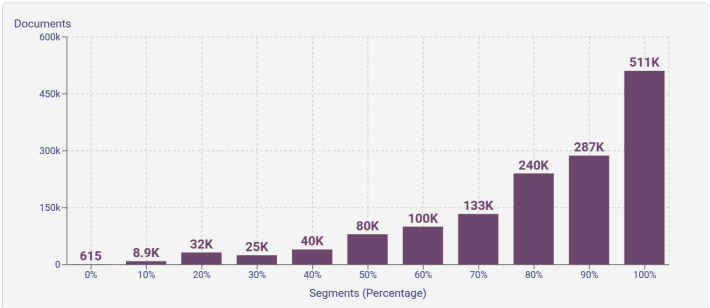


Language Distribution

Number of segments in the Afrikaans (af) corpus

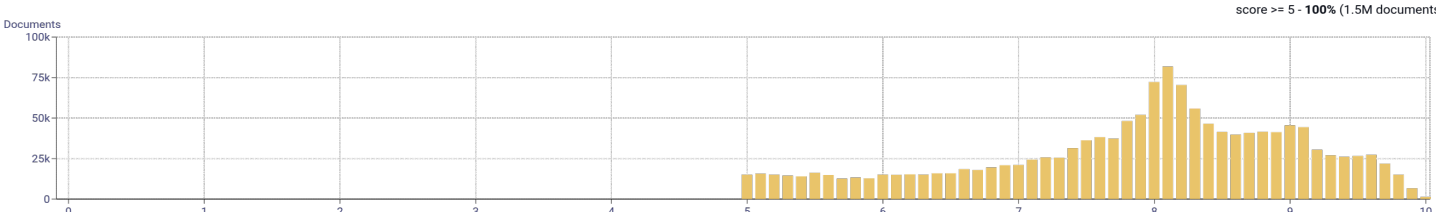


Percentage of segments in Afrikaans (af) inside documents



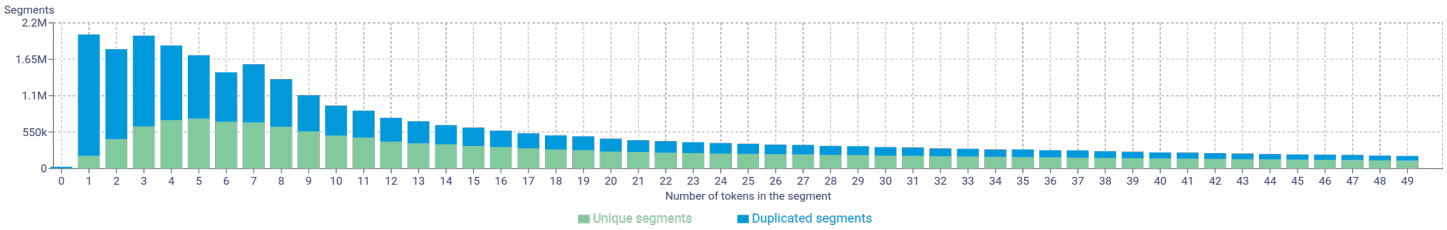
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (1.5M documents)

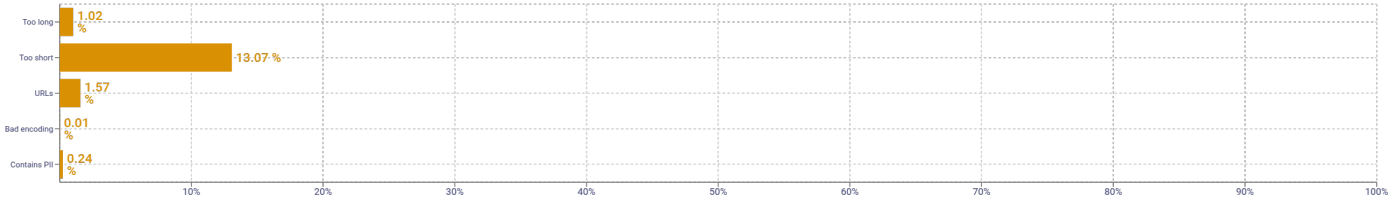


Segment length distribution by token

≤ 49 tokens = **14M** segments | **16M** duplicates
> 50 tokens = **7.3M** segments | **2.8M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	word | 6406818 · of | 6184511 · deur | 3733665 · hierdie | 3584126 · ook | 3389903
2	of the | 474554 · wysig bron | 345356 · word deur | 234204 · gebruik word | 180407 · moet word | 173996
3	vanaf die oorspronklike | 48240 · oor die algemeen | 43175 · sout en peper | 36126 · voor te berei | 30939 · woord van god | 30722
4	geargiveer vanaf die oorspronklike | 43348 · wikimedia commons het meer | 24986 · commons het meer media | 24983 · we are searching data | 24886 · searching data for your | 24886
5	wikimedia commons het meer media | 24983 · we are searching data for | 24886 · searching data for your request | 24886 · are searching data for your | 24886 · speletjie saam met jou beste | 24549

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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av