HPLT Analytics report

General overview

Corpus	Date	Language
hplt-v3-kmb_Latn	9/17/2025	Kimbundu

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,178	20,703	19,076 (92.14 %)	1.1M	4,882,349	4.76 MB

Top 10 domains

Domain	Docs	% of total
jw.org	1.1K	97.54%
sapo.ao	7	0.59%
unicode.org	2	0.17%
mweloweto.com	2	0.17%
contafrica.org	2	0.17%
watchtower.org	1	0.08%
seghea.com	1	0.08%
scribd.com	1	0.08%
newchudaikikaha...	1	0.08%
namendb.com	1	0.08%

Top 10 TLDs

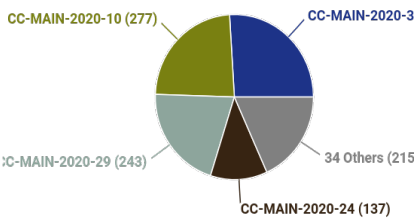| Domain	Docs	% of total
org	1.2K	98.22%
com	7	0.59%
ao	7	0.59%
com.br	2	0.17%
pt	1	0.08%
mus.br	1	0.08%
info	1	0.08%
co.uk	1	0.08%
club	1	0.08%

Documents size (in segments) ⓘ
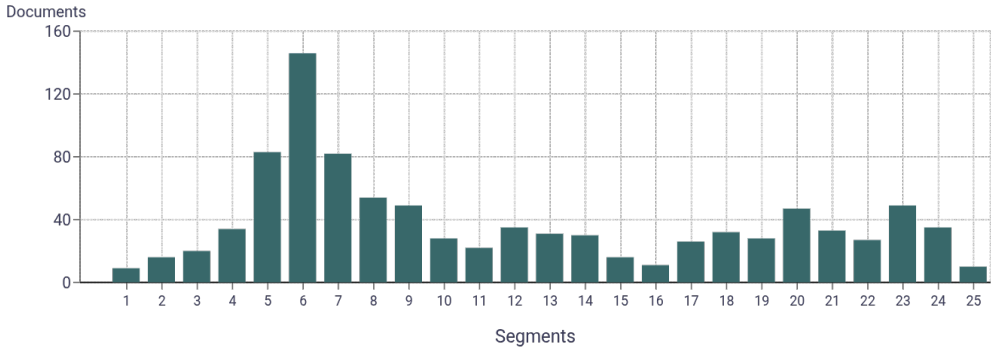
≤ 25 segments **80.9%** (953 documents)
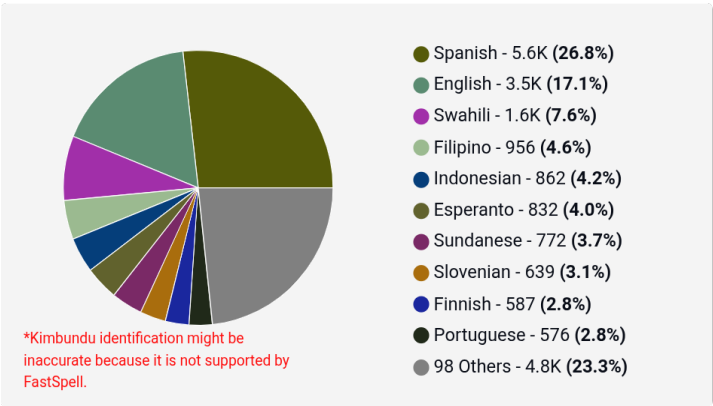> 25 segments **19.1%** (225 documents)
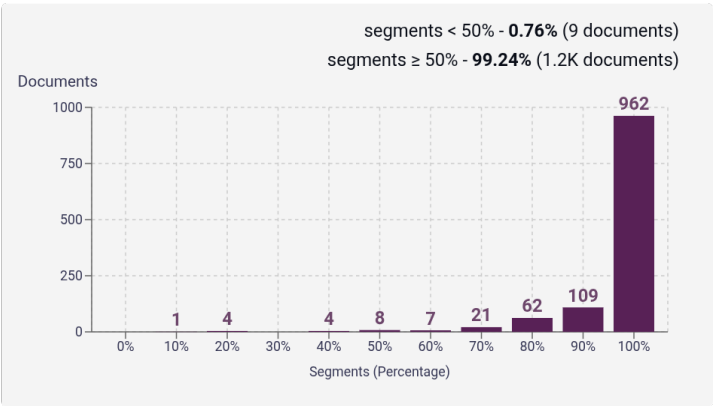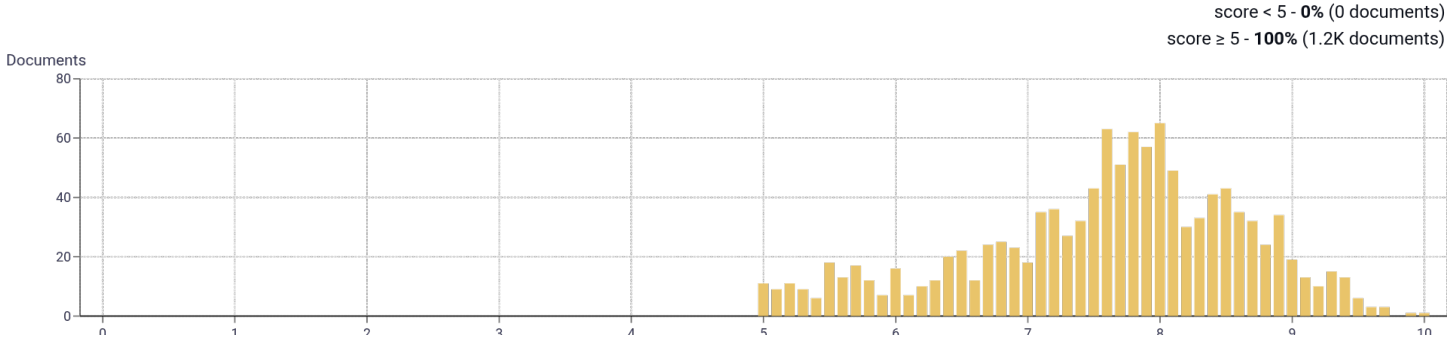

Document collections

CC = 93.46%
IA = 6.54%

CC-MAIN-2020-10 (277)
CC-MAIN-2020-3
CC-MAIN-2020-29 (243)
34 Others (215
CC-MAIN-2020-24 (137)

Language Distribution

Number of segments in the Kimbundu corpus

- Spanish - 5.6K **(26.8%)**
- English - 3.5K **(17.1%)**
- Swahili - 1.6K **(7.6%)**
- Filipino - 956 **(4.6%)**
- Indonesian - 862 **(4.2%)**
- Esperanto - 832 **(4.0%)**
- Sundanese - 772 **(3.7%)**
- Slovenian - 639 **(3.1%)**
- Finnish - 587 **(2.8%)**
- Portuguese - 576 **(2.8%)**
- 98 Others - 4.8K **(23.3%)**

*Kimbundu identification might be inaccurate because it is not supported by FastSpell.

Percentage of segments in Kimbundu inside documents

segments < 50% - **0.76%** (9 documents)
segments ≥ 50% - **99.24%** (1.2K documents)

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (1.2K documents)

Documents

Segment length distribution by token

≤ **49** tokens = **13K** segments | **1.6K** duplicates
> **50** tokens = **7.9K** segments | **16** duplicates

Segments

Segment noise distribution

Too long	**0.81%**
Too short	**3.26%**
URLs	**0.47%**
Bad encoding	**0.74%**
Contains PII	**0.00%**

■ % of corpus
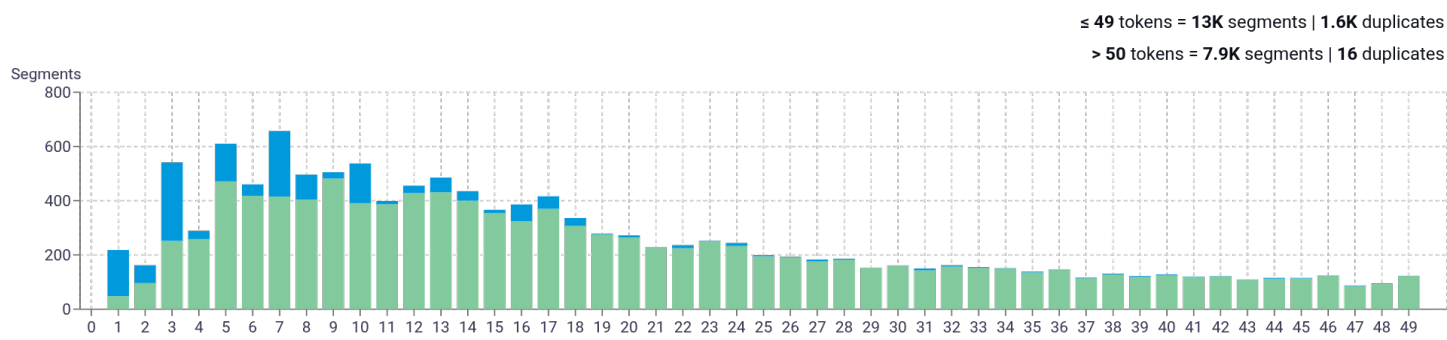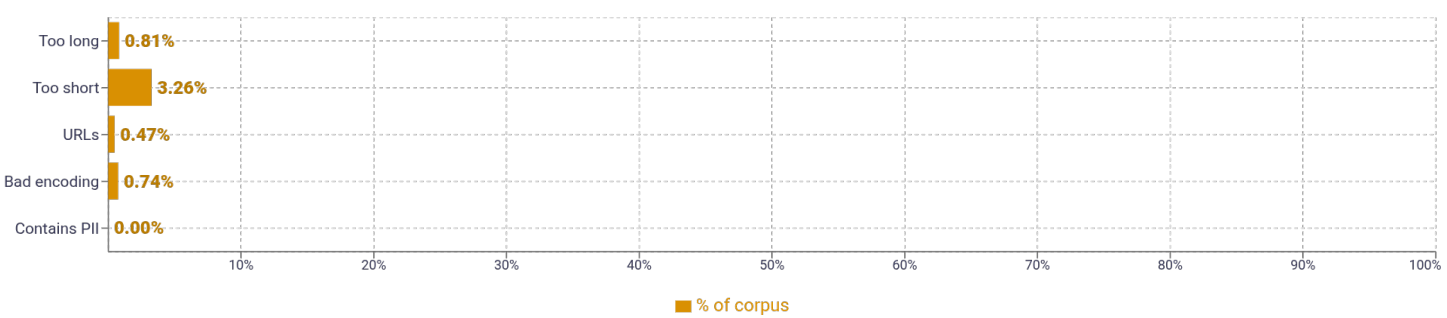
Frequent n-grams

SIZE	N-GRAMS					
1	o | 55,419	mu | 41,368	ni | 27,421	ua | 21,164	ia | 18,665	
2	lungu ni | 2,580	o athu | 2,130	kexile mu | 2,127	o bibidia | 1,988	muéne ua | 1,787	
3	ia lungu ni | 1,564	bhu kaxi ka | 612	jimbangi ja jihova | 577	ua kexile mu | 531	utuminu ua nzambi | 496	
4	kiebhi ki tu tena | 337	o utuminu ua nzambi | 275	o jimbangi ja jihova | 253	o bibidia i tu | 158	jikidistá a a undu | 147	
5	ihi i longa o bibidia | 105	o jikidistá a a undu | 98	o bibidia i tu tangela | 85	o kidi kia lungu ni | 81	mu kitangelu kya bibidya kya | 80	

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				