HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-sag_Latn	9/18/2025	Sango

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
2,638	55,772	51,488 (92.32 %)	3.6M	14,026,771	13.87 MB

Top 10 domains

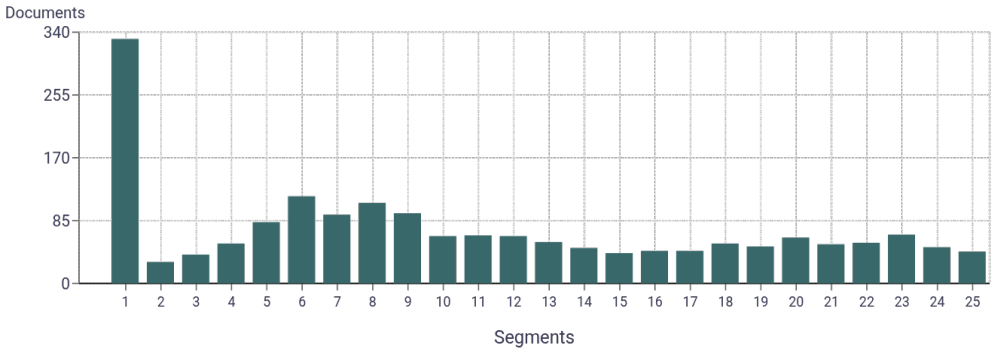
Domain	Docs	% of total
jw.org	2.1K	79.87%
bible.is	329	12.47%
islamhouse.com	40	1.52%
wiktionary.org	34	1.29%
icc-cpi.int	16	0.61%
wikipedia.org	14	0.53%
lueur.org	11	0.42%
gotquestions.org	9	0.34%
siriri.org	6	0.23%
bible.com	6	0.23%

Top 10 TLDs

Domain	Docs	% of total
org	2.2K	83.66%
is	329	12.47%
com	69	2.62%
int	16	0.61%
fr	5	0.19%
net	2	0.08%
info	2	0.08%
blog	2	0.08%
xyz	1	0.04%
ru	1	0.04%

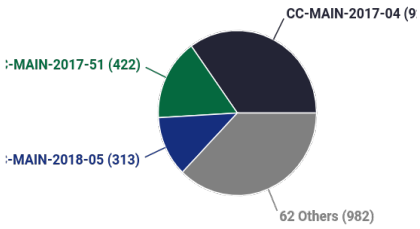
Documents size (in segments)

≤ 25 segments **68.57%** (1.8K documents)
> 25 segments **31.43%** (829 documents)



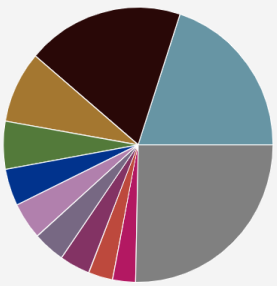
Document collections

CC = 95.75%
IA = 4.25%



Language Distribution

Number of segments in the Sango corpus

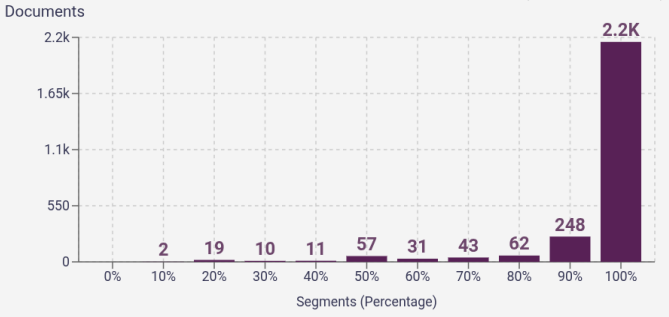


*Sango identification might be inaccurate because it is not supported by FastSpell.

- Iloko - 11K **(20.0%)**
- Filipino - 10K **(18.8%)**
- Welsh - 4.7K **(8.4%)**
- Albanian - 3.1K **(5.6%)**
- Sundanese - 2.4K **(4.3%)**
- Lojban - 2.4K **(4.3%)**
- Spanish - 2.1K **(3.8%)**
- Guarani - 2K **(3.7%)**
- null - 1.6K **(2.9%)**
- Sardinian - 1.5K **(2.8%)**
- 107 Others - 14K **(25.3%)**

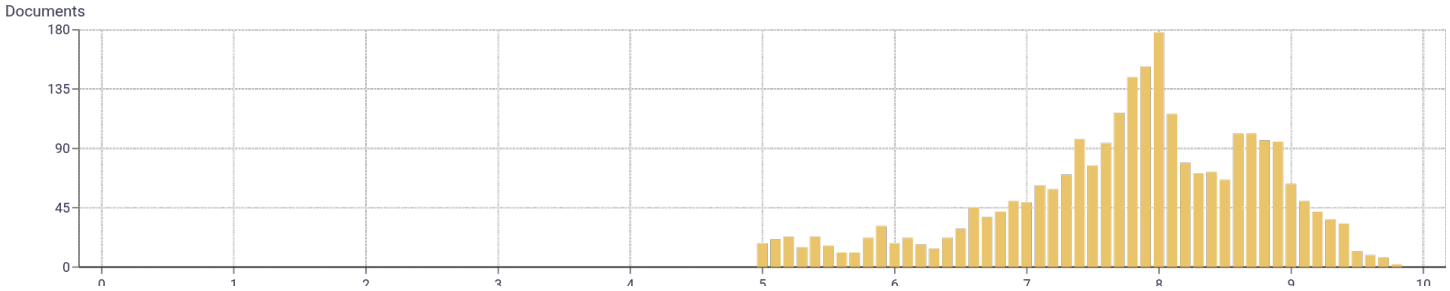
Percentage of segments in Sango inside documents

segments < 50% - **1.59%** (42 documents)
segments ≥ 50% - **98.41%** (2.6K documents)



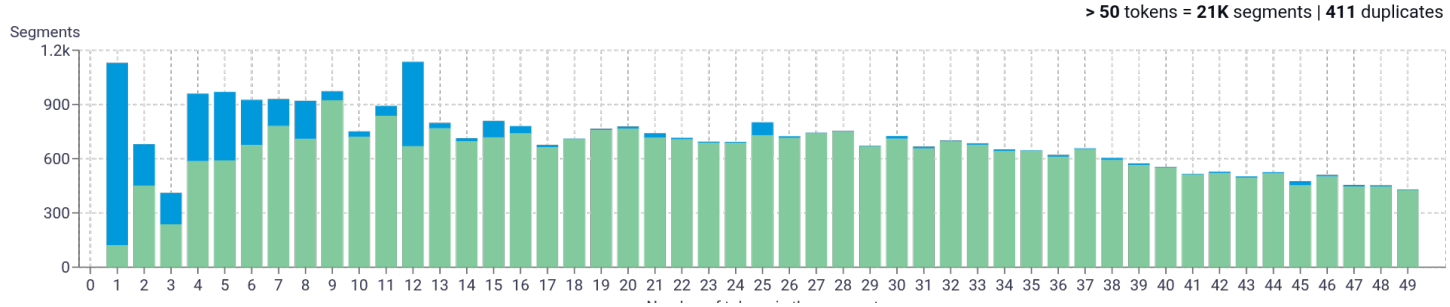
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (2.6K documents)

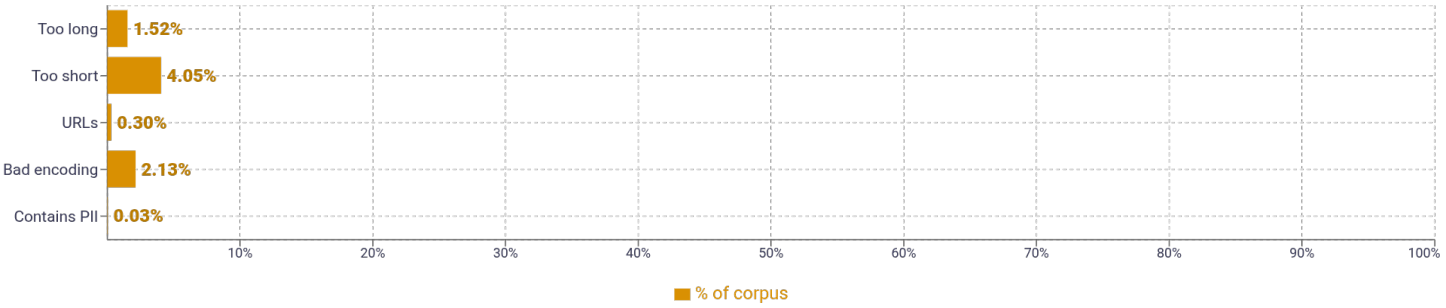


Segment length distribution by token

≤ 49 tokens = **35K** segments | **3.9K** duplicates
> 50 tokens = **21K** segments | **411** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	ala | 73,871 · ayeke | 61,728 · yeke | 48,405 · ye | 41,898 · mbi | 34,620
2	ala yeke | 8,484 · mbi yeke | 5,753 · tongana nyen | 5,106 · yeke sara | 4,820 · sara ye | 4,232
3	ngbanga ti nyen | 2,687 · ayeke na yâ | 1,887 · azo so ayeke | 1,880 · ye so ayeke | 1,876 · société biblique de | 1,830
4	société biblique de centrafrique | 1,830 · nzapa ti fini dunia | 685 · alingbi ti mû maboko | 522 · nyen la e lingbi | 492 · ndali ti so ala | 460
5	mbeti ti nzapa ti fini | 685 · bible na ndo ti internet | 584 · tongana nyen la e lingbi | 353 · ayeke na yâ ti bible | 328 · ala ti gue na yayu | 314

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				