HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-wol_Latn	9/18/2025	Wolof (wo)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
5,056	161,394	143,895 (89.16 %)	4.9M	20,359,482	20.27 MB

Top 10 domains

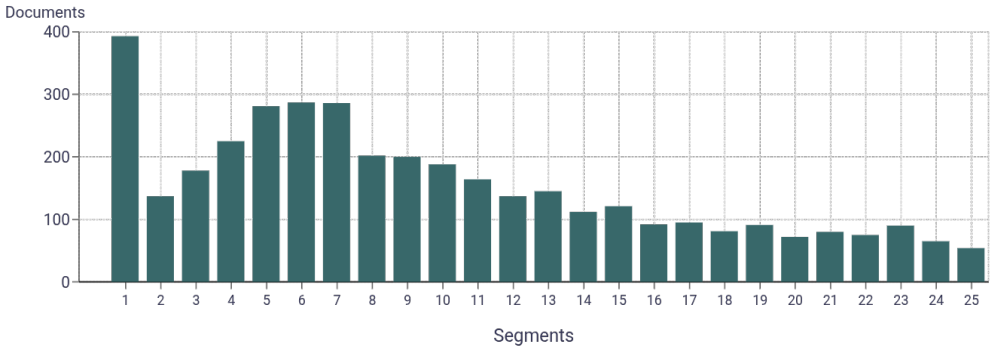
Domain	Docs	% of total
defuwaxu.com	1.4K	28.09%
wikipedia.org	487	9.63%
ebible.org	475	9.39%
breakeveryyoke.com	444	8.78%
bible.is	358	7.08%
jw.org	197	3.90%
ettubwolof.org	173	3.42%
wolof-online.com	125	2.47%
iqna.ir	101	2.00%
al-habdul-xadii...	94	1.86%

Top 10 TLDs

Domain	Docs	% of total
com	2.8K	54.83%
org	1.6K	30.78%
is	358	7.08%
ir	106	2.10%
net	86	1.70%
info	30	0.59%
sn	25	0.49%
fr	18	0.36%
download	15	0.30%
edu	12	0.24%

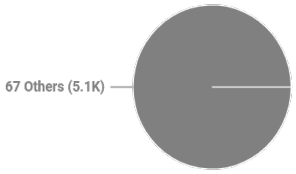
Documents size (in segments)

≤ 25 segments **76.17%** (3.9K documents)
> 25 segments **23.83%** (1.2K documents)



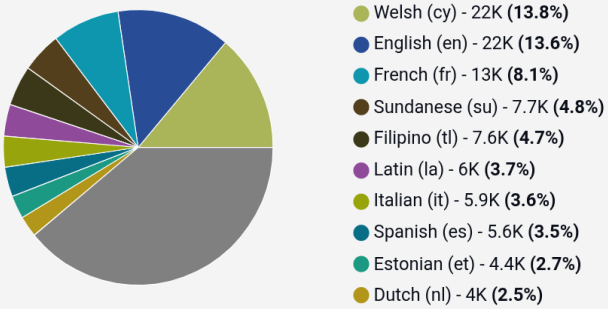
Document collections

CC = 91.30%
IA = 8.70%



Language Distribution

Number of segments in the Wolof (wo) corpus

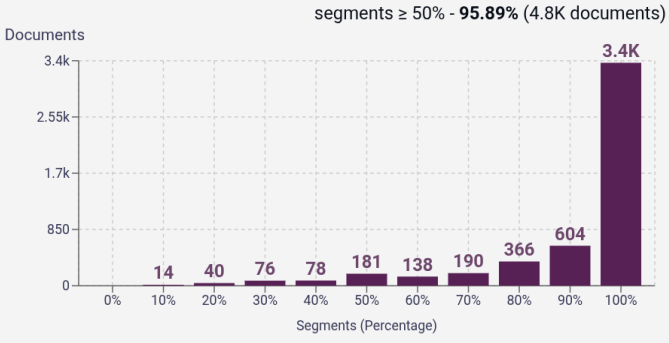


- Welsh (cy) - 22K **(13.8%)**
- English (en) - 22K **(13.6%)**
- French (fr) - 13K **(8.1%)**
- Sundanese (su) - 7.7K **(4.8%)**
- Filipino (tl) - 7.6K **(4.7%)**
- Latin (la) - 6K **(3.7%)**
- Italian (it) - 5.9K **(3.6%)**
- Spanish (es) - 5.6K **(3.5%)**
- Estonian (et) - 4.4K **(2.7%)**
- Dutch (nl) - 4K **(2.5%)**
- 161 Others - 63K **(39.0%)**

*Wolof (wo) identification might be inaccurate because it is not supported by FastSpell

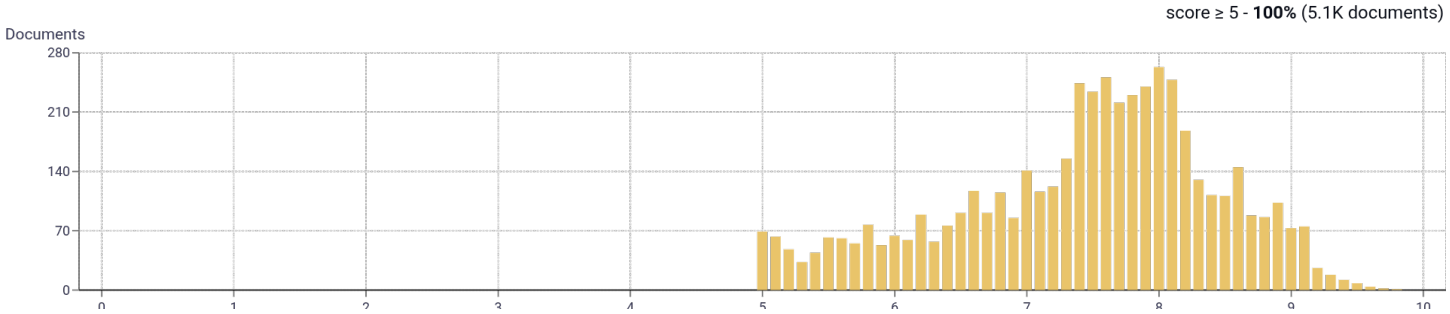
Percentage of segments in Wolof (wo) inside documents

segments < 50% - **4.11%** (208 documents)
segments ≥ 50% - **95.89%** (4.8K documents)



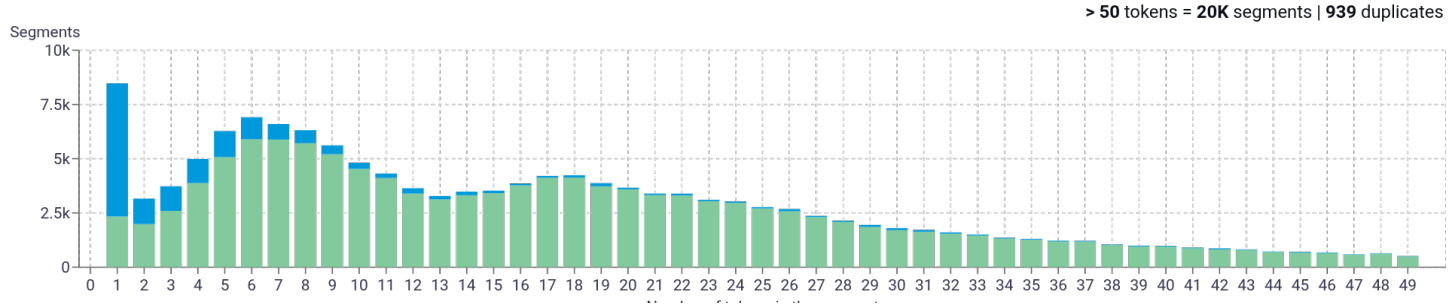
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (5.1K documents)

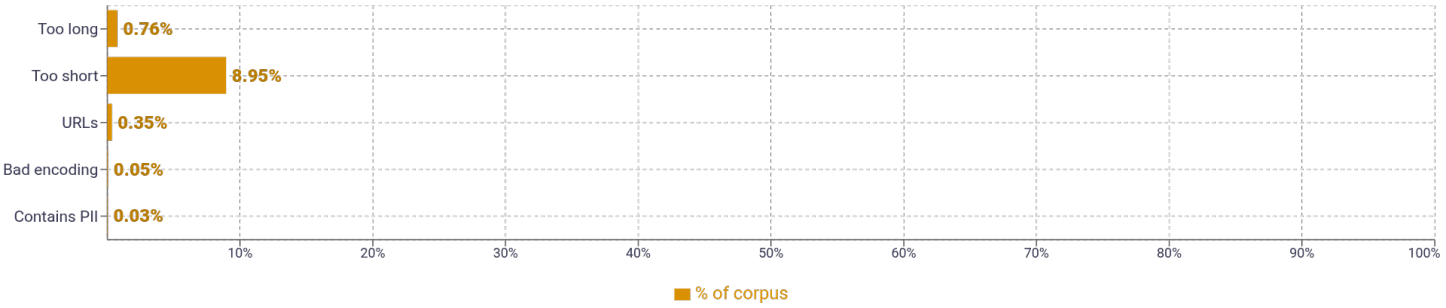


Segment length distribution by token

≤ 49 tokens = **141K** segments | **17K** duplicates
> 50 tokens = **20K** segments | **939** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	nii | 12,499 · xaa | 11,176 · aji | 9,538 · yeesu | 8,873 · lo | 8,165
2	lëë xaa | 2,572 · xaa nii | 1,740 · usmaan sonko | 1,541 · maki sàll | 1,395 · meñ nii | 1,280
3	dinañuy wax tamit | 1,577 · aan lëë xaa | 598 · aji sax jee | 574 · xel mu sell | 382 · fan ci weeru | 362
4	nii xniladzy xtiits dios | 237 · kàddug aji sax jee | 237 · dund gu dul jeex | 209 · meñ nii xniladzy xtiits | 202 · xaa nii xñabey lo | 163
5	meñ nii xniladzy xtiits dios | 177 · yal na ko yàlla dolli | 151 · aji sax ji boroom gàngoor | 116 · sunu boroom subhaanahu wa tahaalaa | 113 · njiitu réew mi maki sàll | 107

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				