

General overview

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

Volumes

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

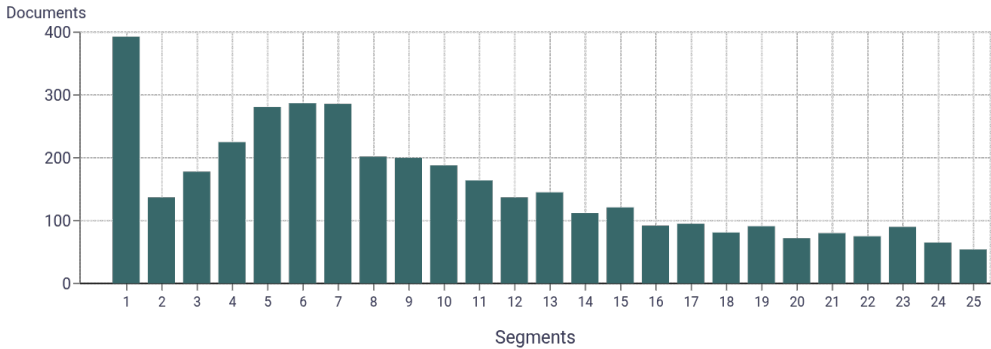
Top 10 domains

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

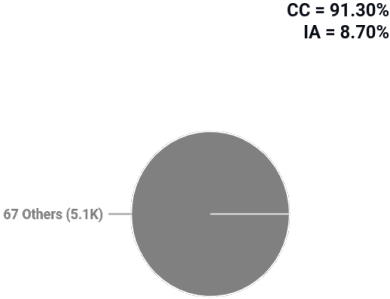
Domain	Docs	% of total
com	2.8K	54.83%
org	1.6K	30.78%
is	358	7.08%
ir	106	2.10%
net	86	1.70%
info	30	0.59%
sn	25	0.49%
fr	18	0.36%
download	15	0.30%
edu	12	0.24%

Documents size (in segments) ⓘ



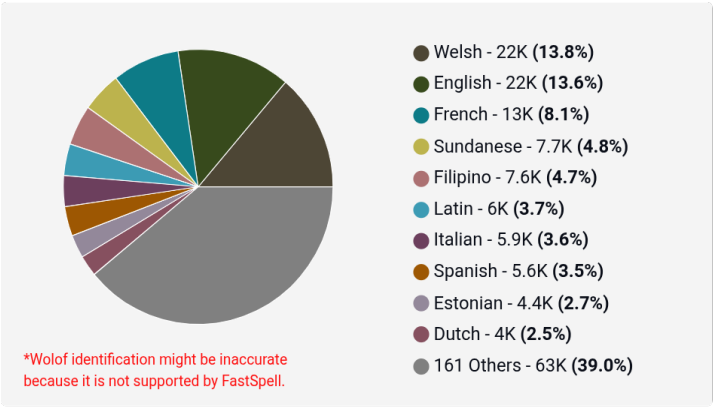
≤ 25 segments **76.17%** (3.9K documents)  
> 25 segments **23.83%** (1.2K documents)

Document collections

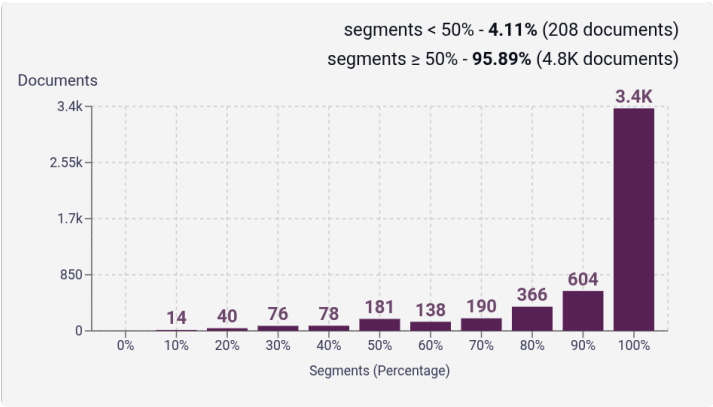


Language Distribution

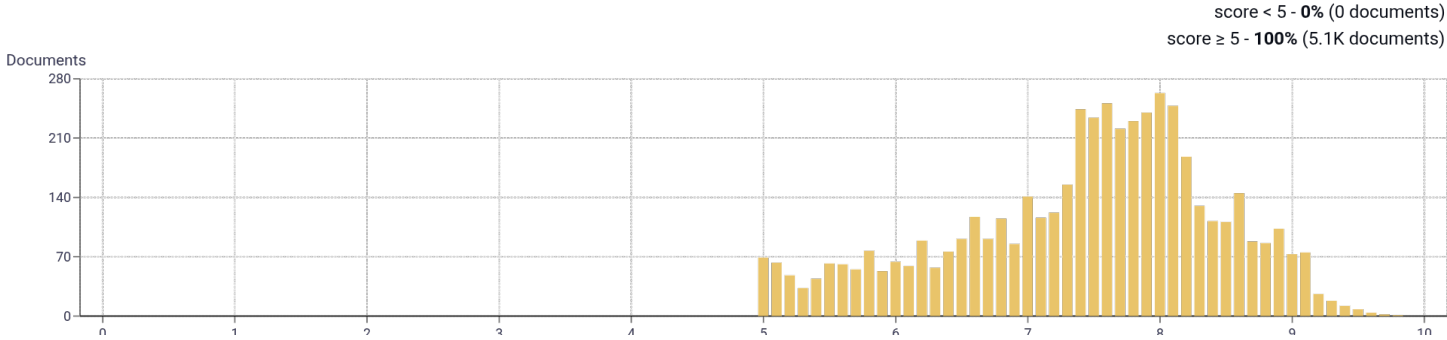
Number of segments in the Wolof corpus



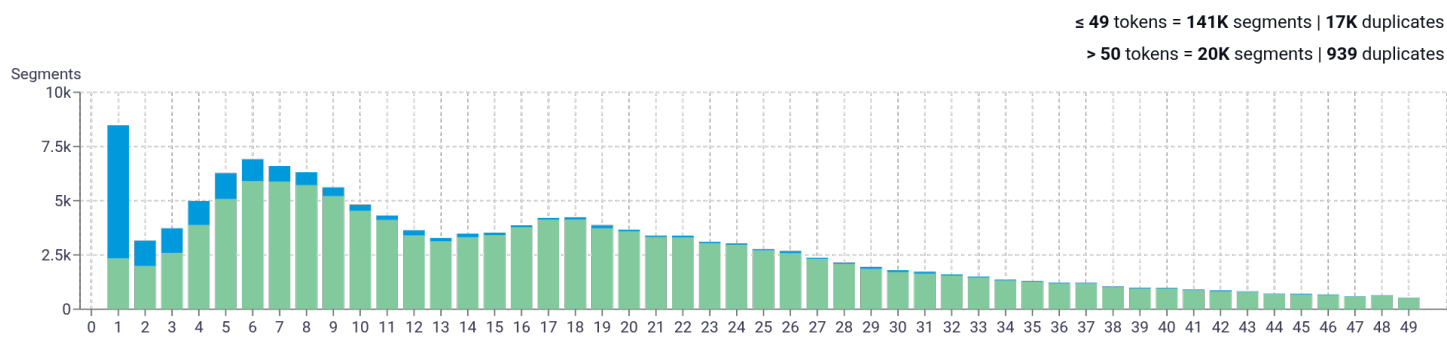
Percentage of segments in Wolof inside documents



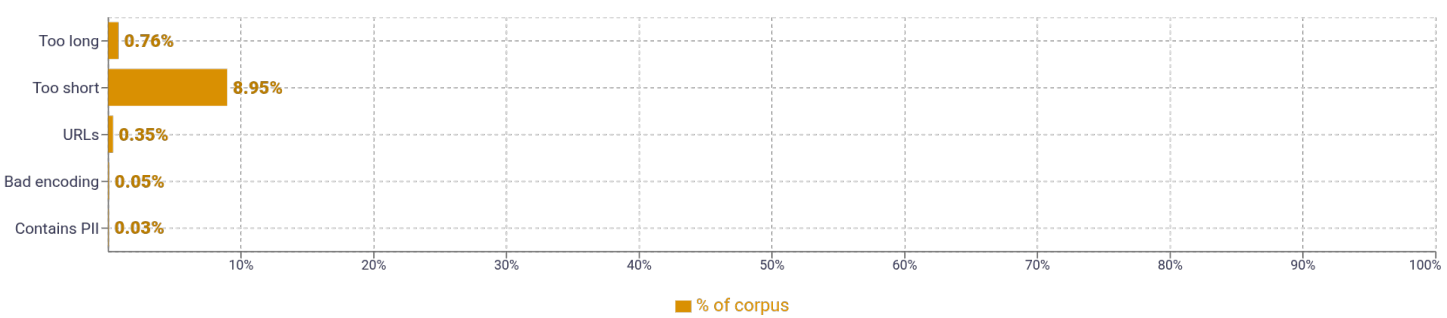
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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

# 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  $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$ , 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	ntp
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				