

General overview

Corpus	Date	Language
hplt-v3-xho_Latn	9/18/2025	Xhosa

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
253,806	6,636,237	4,214,203 (63.50 %)	131M	856,657,669	819.02 MB

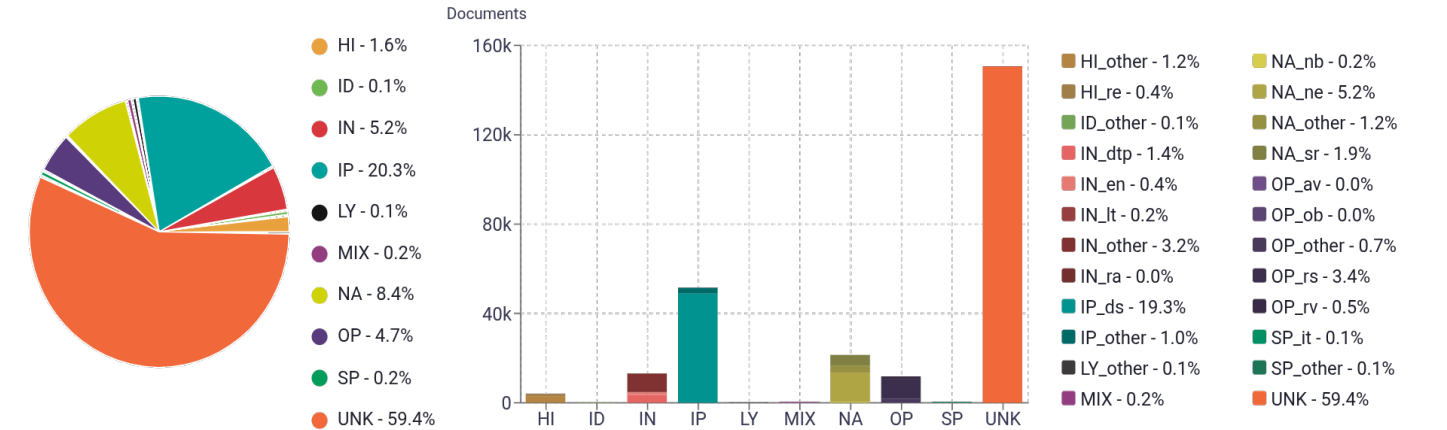
Top 10 domains

Domain	Docs	% of total
airbnb.com	63K	24.63%
isolezwelesixho...	14K	5.40%
martech.zone	5K	1.97%
airbnb.co.za	3.9K	1.54%
eferrit.com	3.7K	1.46%
jw.org	3.7K	1.44%
eturbonews.com	3.1K	1.22%
actualidadgadg...	2.5K	0.98%
creativosonline...	2.4K	0.95%
androidsis.com	2.4K	0.93%

Top 10 TLDs

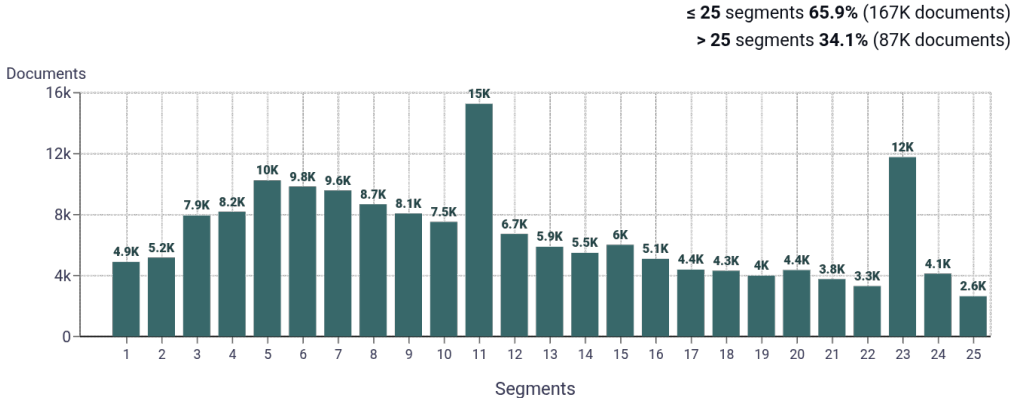
Domain	Docs	% of total
com	184K	72.31%
co.za	24K	9.54%
org	12K	4.60%
zone	5K	1.97%
net	3.2K	1.24%
pt	3.1K	1.20%
online	1.9K	0.73%
date	998	0.39%
tn	929	0.37%
top	801	0.32%

Register labels

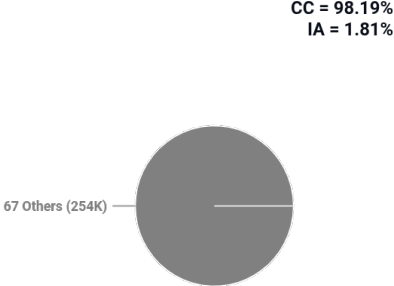


MT:73.0% | 185K Documents

Documents size (in segments) ⓘ

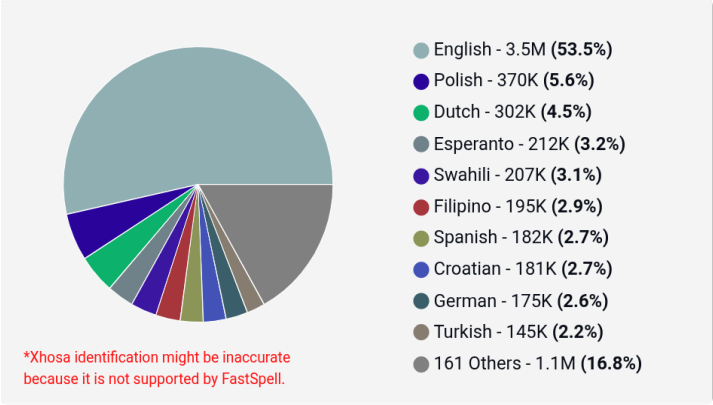


Document collections

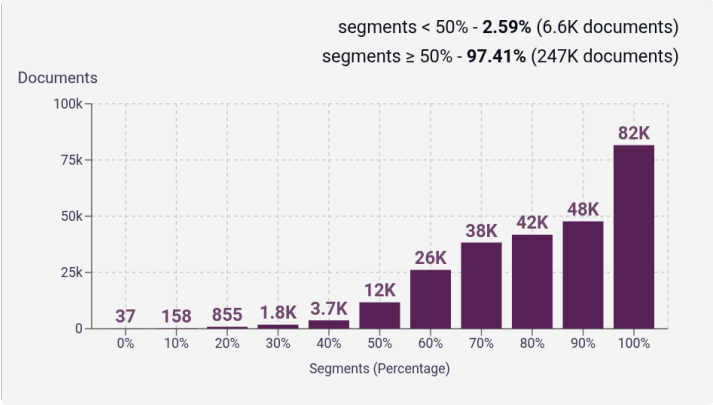


Language Distribution

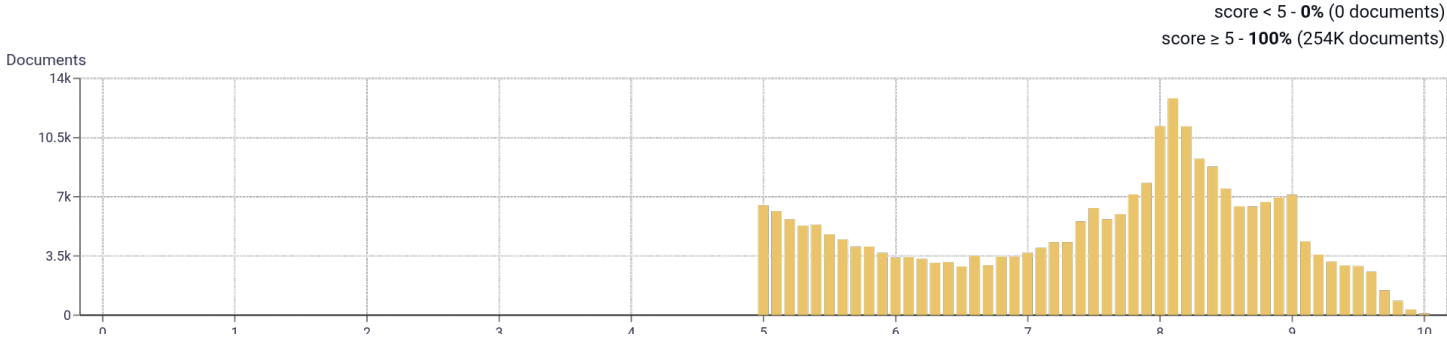
Number of segments in the Xhosa corpus



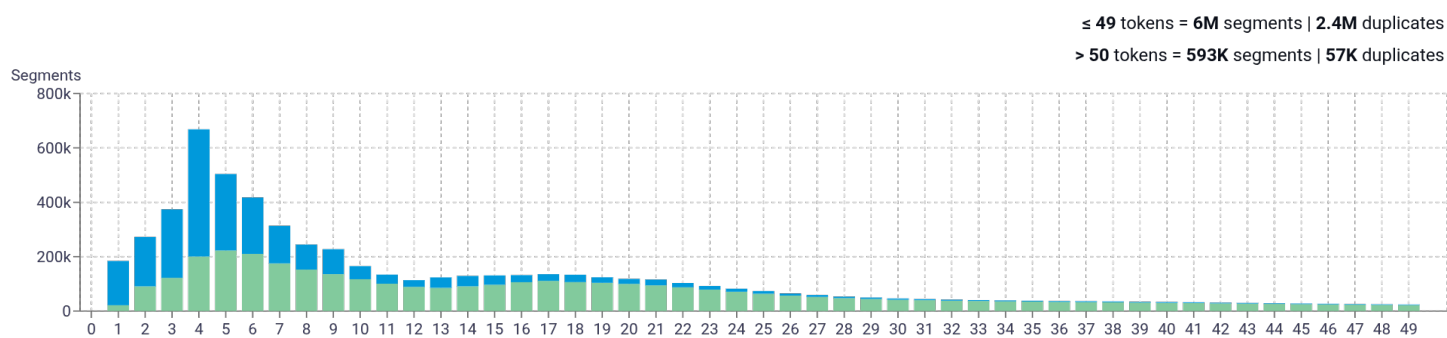
Percentage of segments in Xhosa inside documents



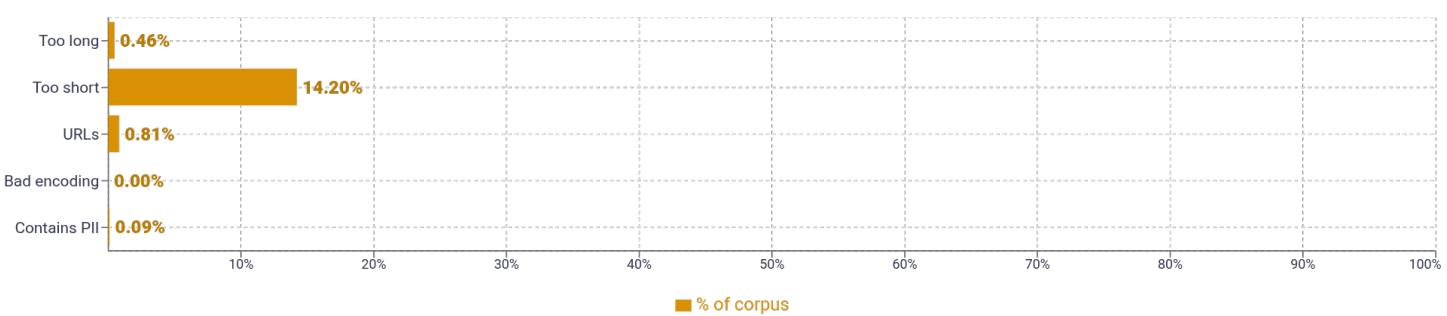
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>i 1,333,625</div> <div>e 739,544</div> <div>eziyi 665,798</div> <div>kwizimvo 509,192</div> <div>kumlinganiselo 509,112</div>	
2	<div>kwizimvo eziyi 508,466</div> <div>ongumyinge weziyi 508,465</div> <div>kumlinganiselo ongumyinge 508,465</div> <div>indlu e 199,799</div> <div>iindawo eziqeshisayo 134,859</div>	
3	<div>kumlinganiselo ongumyinge weziyi 508,465</div> <div>iflethi eqeshisayo e 74,535</div> <div>iflethi okanye indlu 54,628</div> <div>iindawo zeholide eziqeshisayo 46,953</div> <div>iindawo eziqeshisayo zeholide 41,020</div>	
4	<div>iflethi okanye indlu ekwi 42,982</div> <div>zeholide zazo zonke iintlobo 38,603</div> <div>iindawo eziqeshisayo zeholide zazo 38,603</div> <div>fumana uze ubhukishe iindawo 34,921</div> <div>igumbi lakho lokulala e 30,402</div>	
5	<div>eziqeshisayo zeholide zazo zonke iintlobo 38,603</div> <div>uze ubhukishe iindawo zokuhlala ezikhethekileyo 18,663</div> <div>fumana uze ubhukishe iindawo zokuhlala 18,663</div> <div>ubhukishe iindawo zokuhlala ezikhethekileyo kuairbnb 17,204</div> <div>zokuhlala ziconywa kakhulu ngendawo ezikuyo 17,201</div>	

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 $\text{*number of types (uniques)/number of tokens*}$, 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>, , , 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 stopwords. 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				