

General overview

Corpus	Date	Language
hplt-v3-gug_Latn	9/17/2025	Guarani (gn)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
98,974	1,699,445	1,258,836 (74.07 %)	46M	239,598,746	237.71 MB

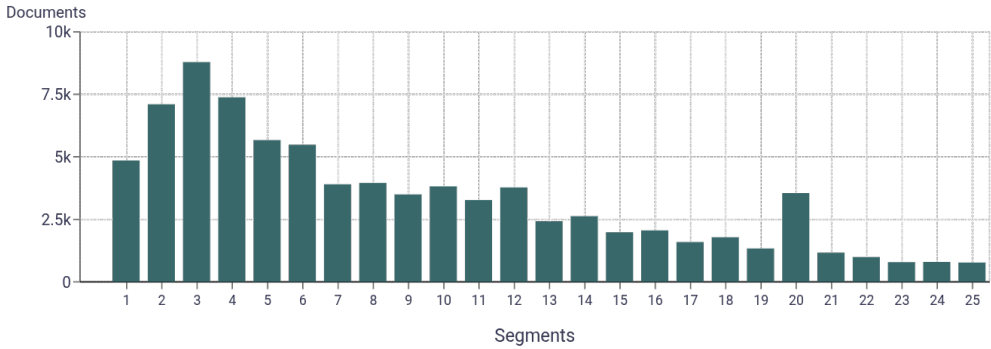
Top 10 domains

Domain	Docs	% of total
uma.es	7.8K	7.92%
blogspot.com	2.7K	2.77%
wikipedia.org	2.1K	2.14%
ucm.es	1.7K	1.71%
jw.org	1.2K	1.25%
spl.gov.py	1.2K	1.20%
abc.com.py	1.2K	1.18%
misaguarani.com	894	0.90%
wordpress.com	816	0.82%
cultura.gob.mx	757	0.76%

Top 10 TLDs

Domain	Docs	% of total
com	25K	24.91%
es	17K	16.68%
org	8.9K	9.00%
com.ar	6.3K	6.34%
cl	3.8K	3.86%
gov.py	3K	3.04%
com.py	2.5K	2.56%
edu.co	2.4K	2.43%
edu.ar	2.4K	2.40%
org.ar	2.1K	2.09%

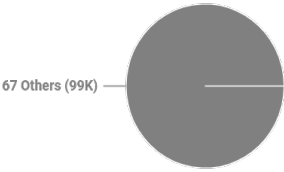
Documents size (in segments) ⓘ



≤ 25 segments **84.25%** (83K documents)
> 25 segments **15.75%** (16K documents)

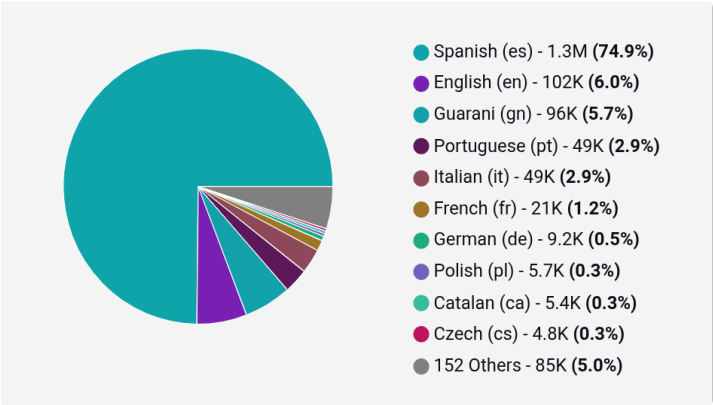
Document collections

CC = **90.08%**
IA = **9.92%**

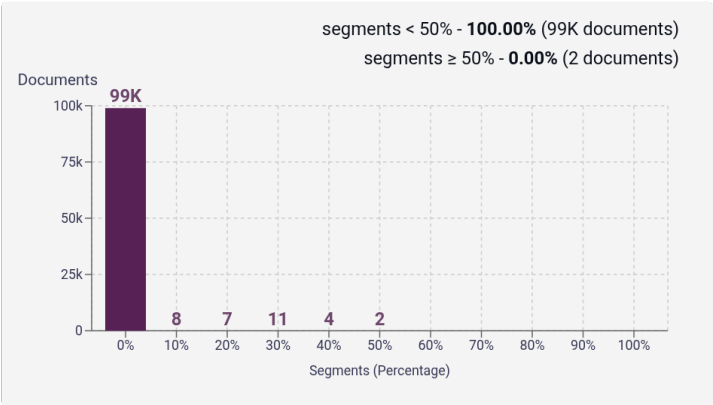


Language Distribution

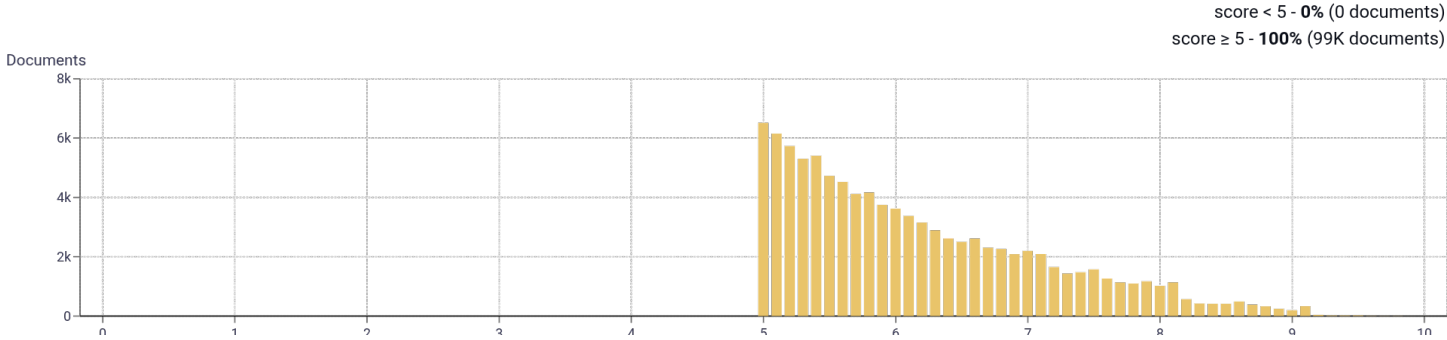
Number of segments in the Guarani (gn) corpus



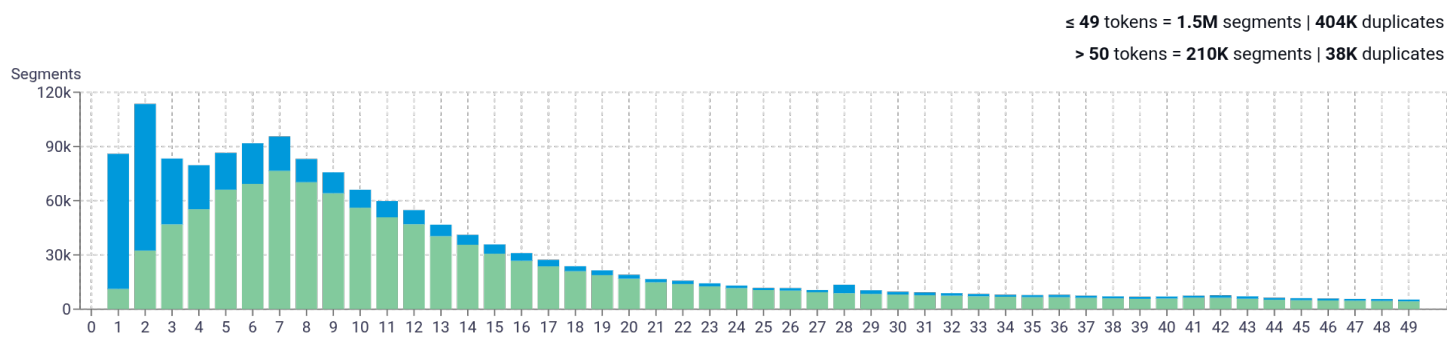
Percentage of segments in Guarani (gn) inside documents



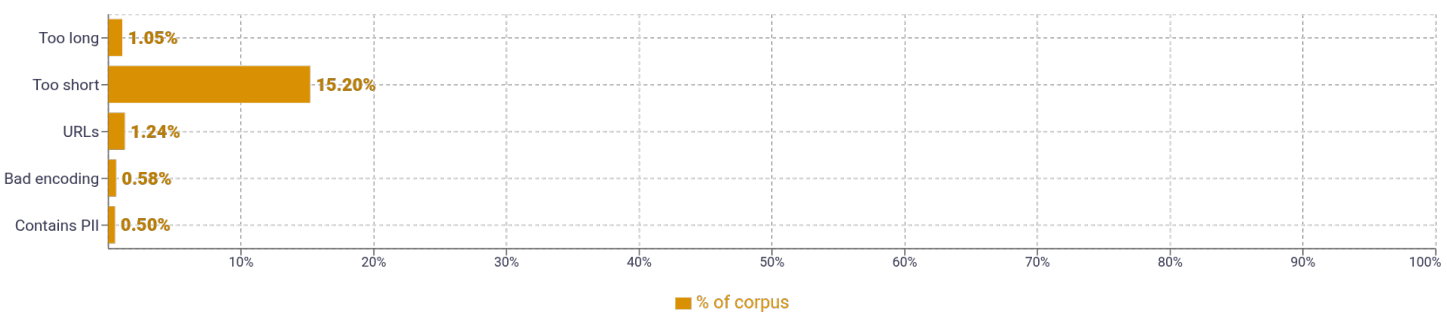
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	de 2,047,032y 698,104la 584,857del 414,024en 336,006	
2	de la 314,938nacional de 75,170universidad de 69,521universidad nacional 54,325facultad de 51,429	
3	universidad nacional de 36,039de la lengua 27,650de buenos aires 23,529de la universidad 22,893traductor jurado de 21,749	
4	jurado de inglés en 21,711traductor jurado de inglés 21,710de la lengua española 12,492ã ã ã ã 12,379universidad de buenos aires 9,843	
5	traductor jurado de inglés en 21,710ã ã ã ã ã 10,496nacional de antropología e historia 6,692instituto mexicano del seguro social 6,389d e f g h 4,662	

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