HPLT Analytics report

General overview

Corpus	Date	Language
hplt-v3-gug_Latn	10/3/2025	Guarani (gug)

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
98,974	1,699,445	1,258,836 (74.07 %)	25.93%	46M	239,598,746	237.71 MB

Top 10 domains

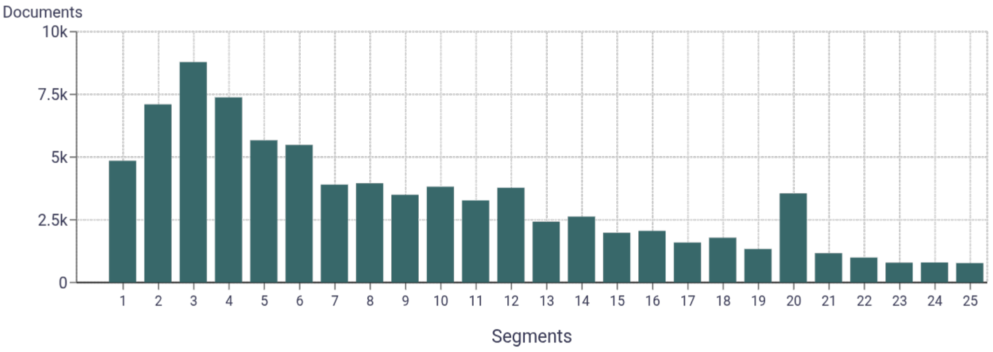
Domain	Docs	% of total
uma.es	7.8K	7.92%
blogspot.com	2.7K	2.77%
wikipedia.org	2.1K	2.14%
ucm.es	1.7K	1.71%
jw.org	1.2K	1.25%
spl.gov.py	1.2K	1.20%
abc.com.py	1.2K	1.18%
misaguarani.com	894	0.90%
wordpress.com	816	0.82%
cultura.gob.mx	757	0.76%

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%

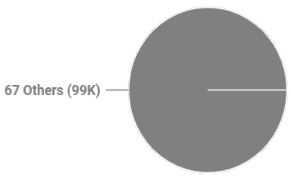
Documents size (in segments) ⓘ

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



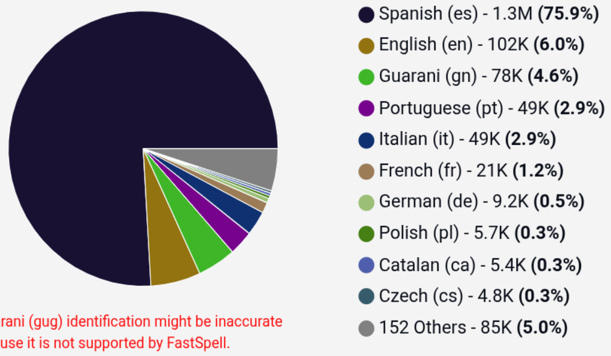
Document collections

CC = 90.08%
IA = 9.92%



Language Distribution

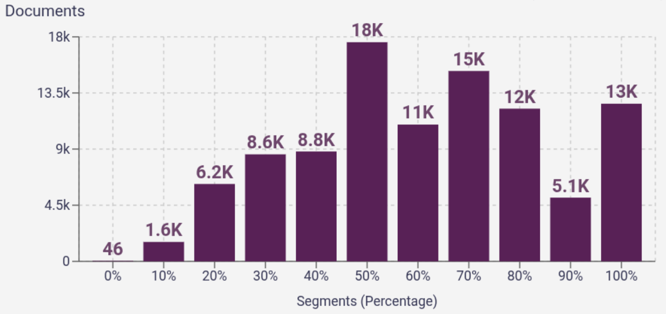
Number of segments in the Guarani (gug) corpus



*Guarani (gug) identification might be inaccurate because it is not supported by FastSpell.

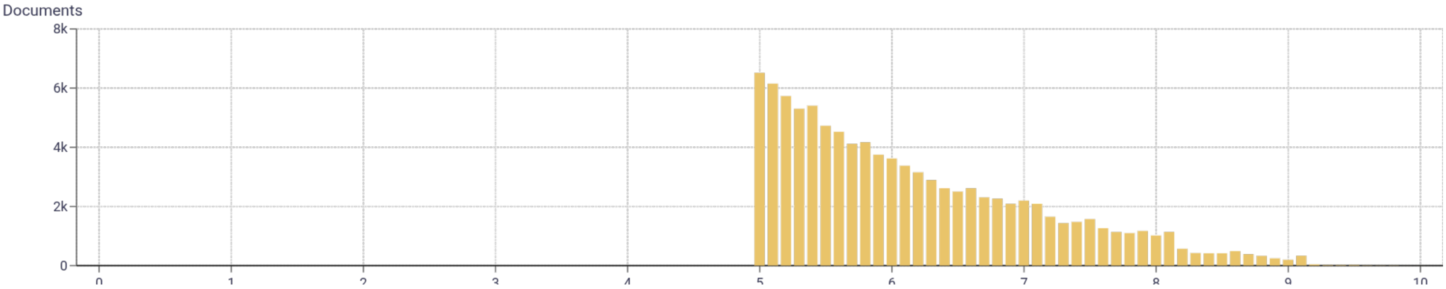
Percentage of segments in Guarani (gug) inside documents

segments < 50% - **25.46%** (25K documents)
segments ≥ 50% - **74.54%** (74K documents)



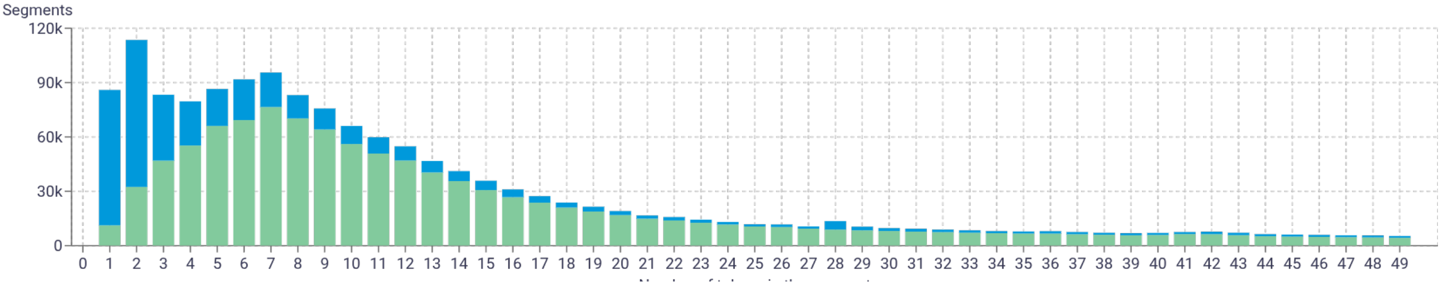
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (99K documents)

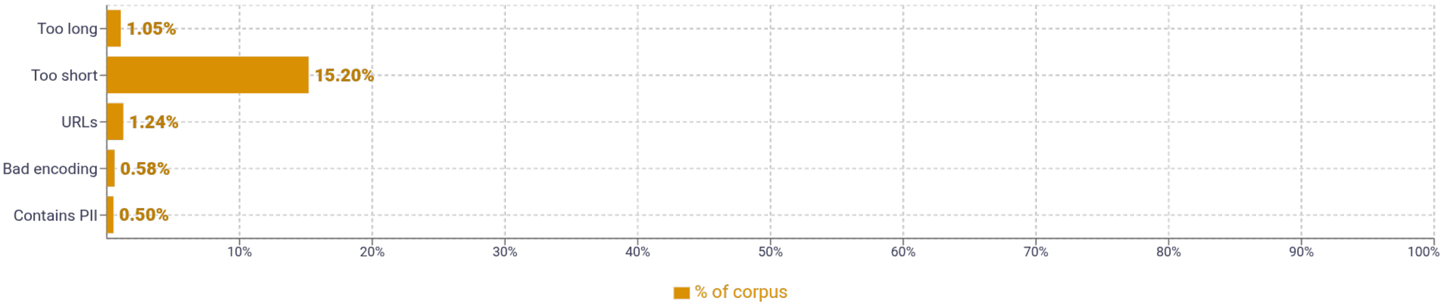


Segment length distribution by token

≤ **49** tokens = **1.5M** segments | **404K** duplicates
> **50** tokens = **210K** segments | **38K** duplicates



Segment noise distribution



Frequent n-grams

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

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				