HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
grn_Latn.jsonl.tsv	12/5/2024	Guarani (gn)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
73,423	1,712,841	830,361 (48.48 %)	41M	216,984,692	215.19 MB

Top 10 domains

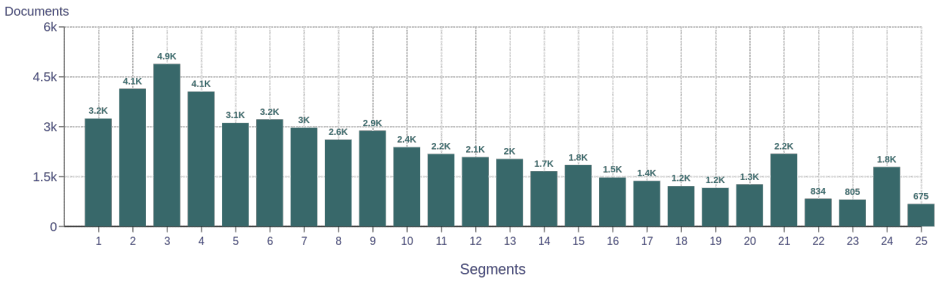
Domain	Docs	% of total
wikipedia.org	5.3K	7.20
blogspot.com	3K	4.15
doctoralia.es	2.9K	3.93
uma.es	2.3K	3.12
arqa.com	1.5K	2.11
bible.is	1.5K	2.05
blogspot.com.ar	1.1K	1.55
jw.org	1.1K	1.54
acceder.gov.ar	759	1.03
wordpress.com	629	0.86

Top 10 TLDs

Domain	Docs	% of total
com	21K	28.86
org	11K	14.43
es	10K	14.26
com.ar	4.6K	6.32
cl	2.3K	3.13
edu.ar	1.7K	2.32
mx	1.7K	2.30
net	1.5K	2.08
is	1.5K	2.06
gov.ar	1.3K	1.80

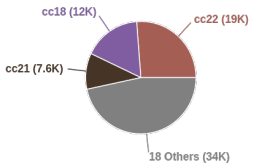
Documents size (in segments)

<= 25 segments **76.34%** (56K documents)
> 25 segments **23.66%** (17K documents)



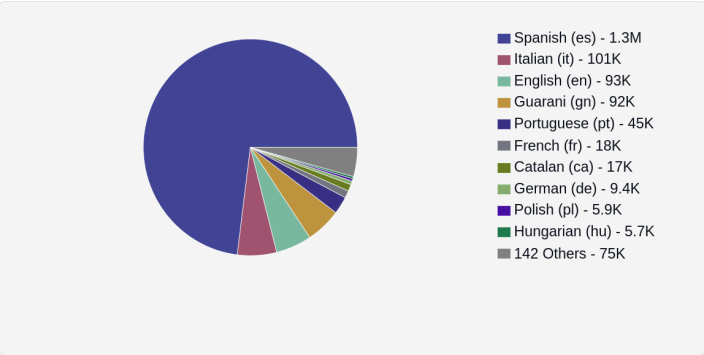
Documents by collection

CC = 67.99%
IA = 32.01%

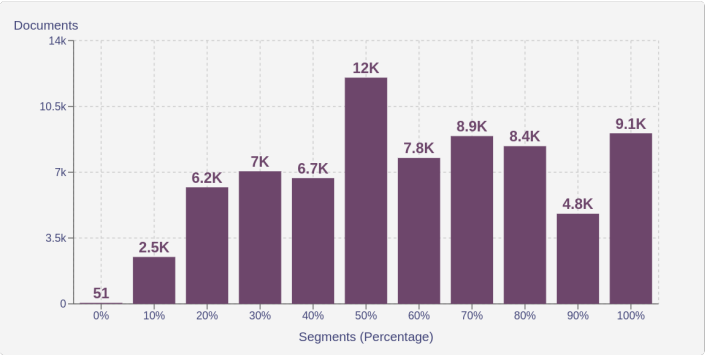


Language Distribution

Number of segments in the Guarani (gn) corpus

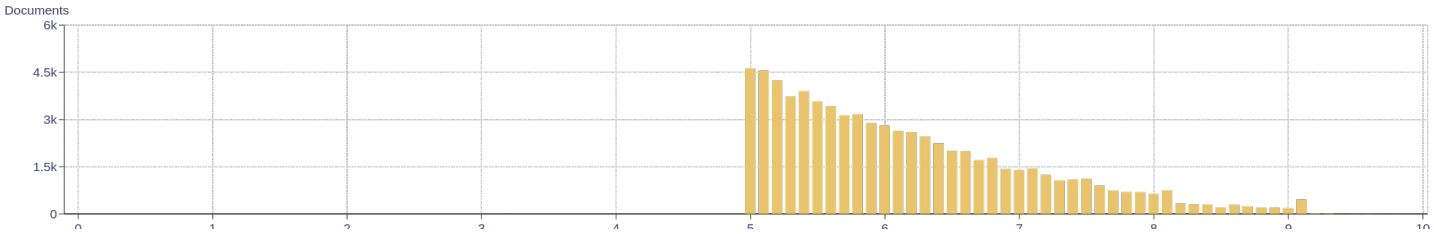


Percentage of segments in Guarani (gn) inside documents



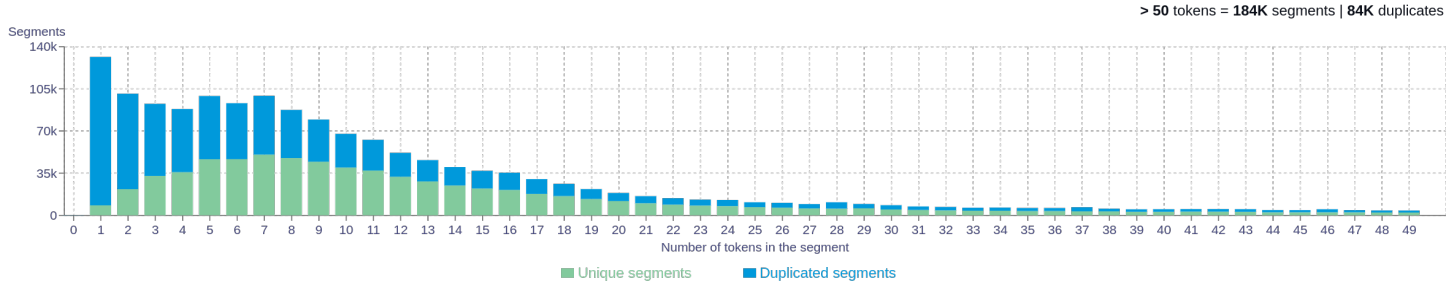
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (73K documents)

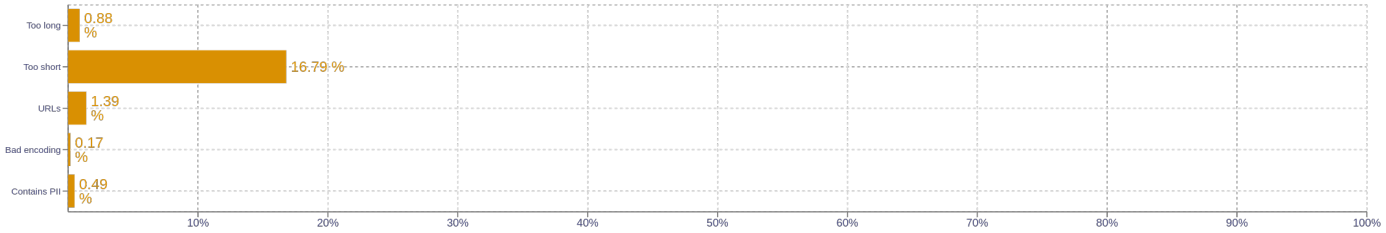


Segment length distribution by token

≤ 49 tokens = **729K** segments | **799K** duplicates
> 50 tokens = **184K** segments | **84K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	de | 1834124 · y | 580839 · la | 489353 · en | 411090 · del | 371574
2	de la | 267313 · buenos aires | 74758 · universidad de | 71576 · nacional de | 69920 · universidad nacional | 62773
3	universidad nacional de | 40143 · de buenos aires | 33131 · de la universidad | 25958 · urólogo ver más | 25612 · república bolivariana de | 22750
4	jurado de inglús en | 21711 · traductor jurado de inglús | 21710 · angiólogo y cirujano vascular | 15721 · medicina familiar y comunitaria | 14912 · del ministerio de defensa | 14528
5	traductor jurado de inglús en | 21710 · oficial del ministerio de defensa | 14473 · diario oficial del ministerio de | 14473 · de medicina familiar y comunitaria | 14145 · ciudad autónoma de buenos aires | 10875

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