HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
ast_Latn.jsonl.tsv	9/26/2024	Asturian (ast)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
273,237	7,426,182	3,418,322 (46.03 %)	248M	1,236,934,368	1.18 GB

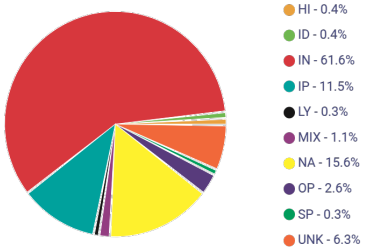
Top 10 domains

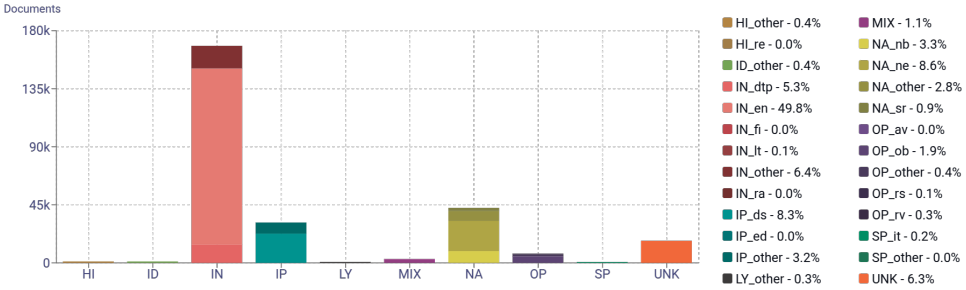
Domain	Docs	% of total
wikipedia.org	147K	53.77%
blogspot.com	8.8K	3.23%
asturies.com	7.6K	2.79%
wordpress.com	6.8K	2.50%
mp3xd.com	5.2K	1.91%
lasidra.as	4.9K	1.81%
blogspot.com.es	4.7K	1.73%
musicadevida.com	3.1K	1.12%
uniovi.es	2.8K	1.02%
mp3canciones.com	2.8K	1.02%

Top 10 TLDs

Domain	Docs	% of total
org	160K	58.44%
com	69K	25.38%
es	16K	5.92%
net	6K	2.20%
as	5.7K	2.07%
com.es	5.3K	1.94%
com.mx	1.4K	0.52%
info	1K	0.38%
com.ar	921	0.34%
de	636	0.23%

Register labels

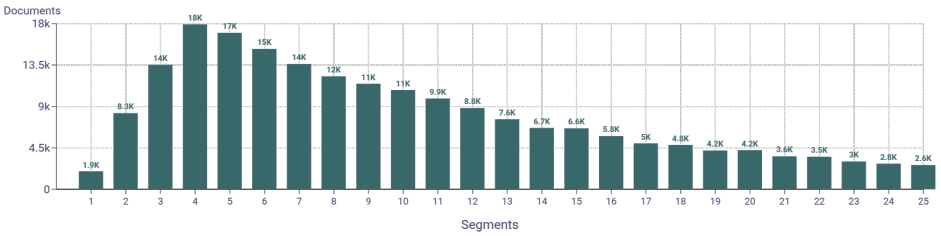




🤖 **MT**:1.8% | 4.9K Documents

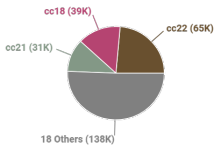
Documents size (in segments)

<= 25 segments **73.63%** (201K documents)
> 25 segments **26.37%** (72K documents)



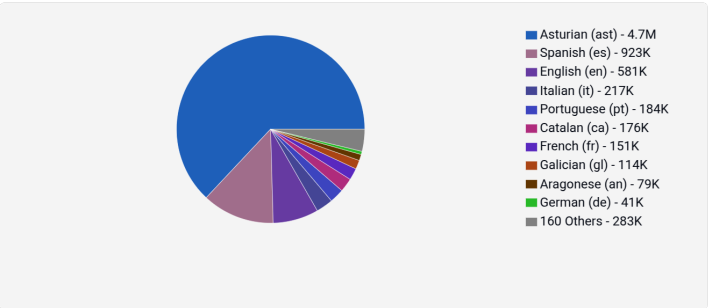
Documents by collection

CC = 67.73%
IA = 32.27%

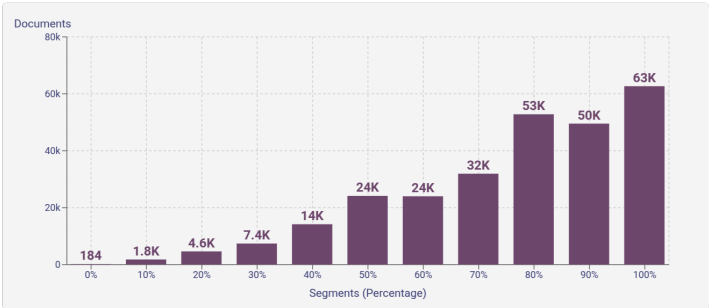


Language Distribution

Number of segments in the Asturian (ast) corpus

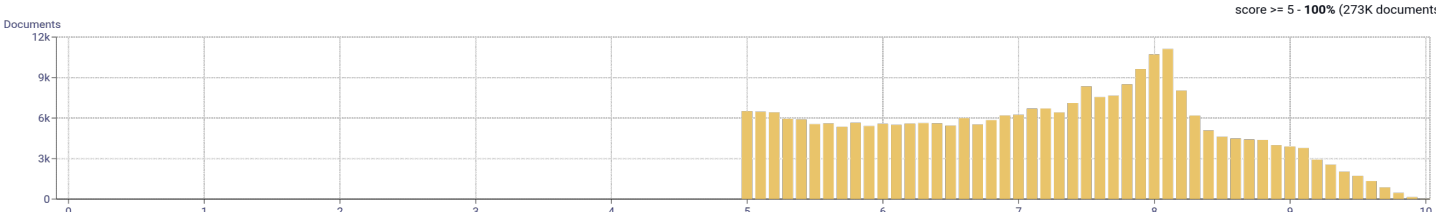


Percentage of segments in Asturian (ast) inside documents



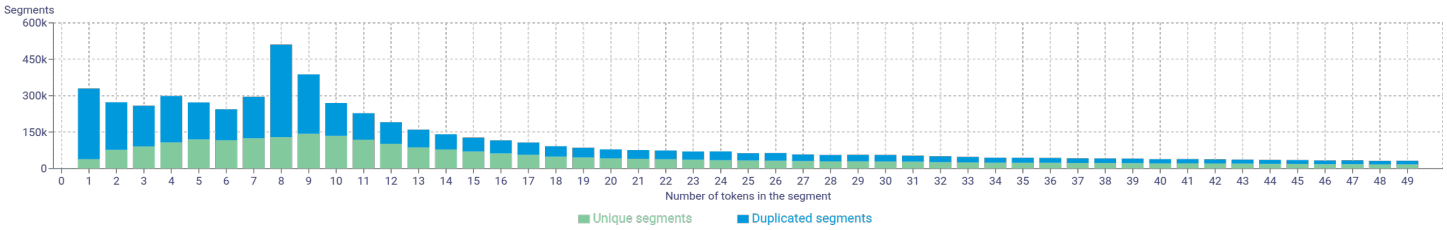
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (273K documents)

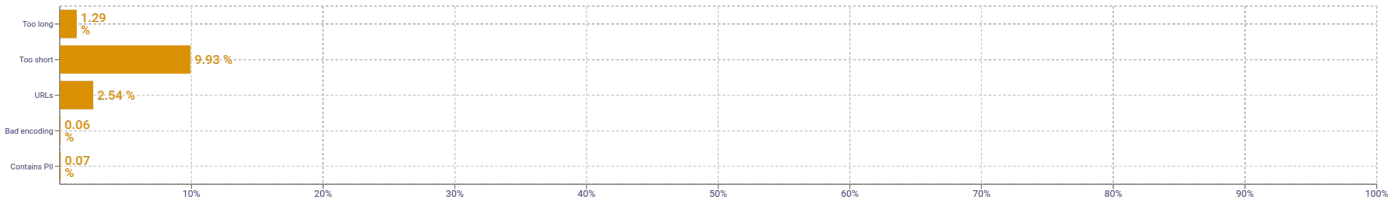


Segment length distribution by token

≤ 49 tokens = **2.5M** segments | **3.3M** duplicates
> 50 tokens = **1.6M** segments | **677K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	editar | 1265852 · fonte | 591614 · descargar | 347315 · the | 315007 · años | 232857
2	descargar reproducir | 71070 · compartir descargar | 71042 · escuchar descargar | 65407 · of the | 48663 · enllaces esternos | 47787
3	editar la fonte | 567301 · compartir descargar reproducir | 71042 · wikimedia commons acueye | 23677 · commons acueye conteníu | 23673 · acueye conteníu multimedia | 23673
4	wikimedia commons acueye conteníu | 23673 · commons acueye conteníu multimedia | 23673 · academia de la llingua | 17989 · llingua de la obra | 17241 · wikimedia commons tien conteníu | 8865
5	wikimedia commons acueye conteníu multimedia | 23673 · academia de la llingua asturiana | 15925 · wikimedia commons tien conteníu multimedia | 8865 · commons tien conteníu multimedia tocante | 8805 · grabs grabs grabs grabs grabs | 7694

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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av