HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-ast_Latn	9/16/2025	Asturian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
247,531	5,075,057	3,867,365 (76.20 %)	201M	1,004,466,606	983.83 MB

Top 10 domains

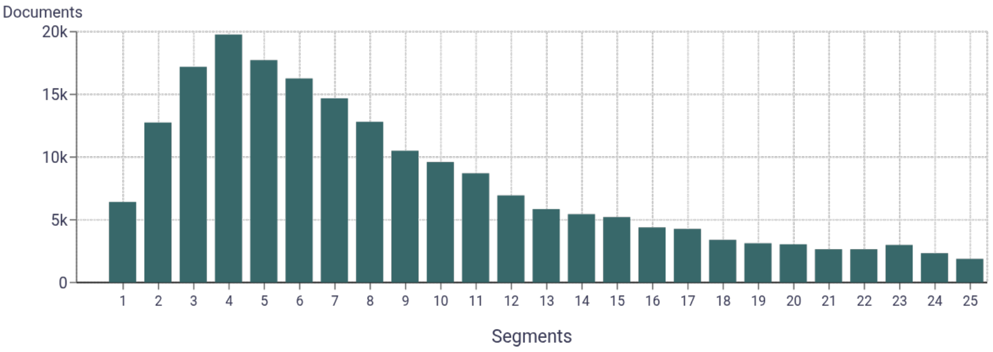
Domain	Docs	% of total
wikipedia.org	80K	32.38%
blogspot.com	12K	5.03%
asturies.com	11K	4.50%
wordpress.com	8.2K	3.30%
lasidra.as	5.6K	2.28%
infoasturies.com	3.9K	1.56%
agapea.com	3.2K	1.30%
infoasturies.net	2.5K	1.02%
blogspot.com.es	2.5K	1.01%
asturias.es	2.3K	0.94%

Top 10 TLDs

Domain	Docs	% of total
org	97K	39.33%
com	88K	35.36%
es	30K	12.27%
net	8.4K	3.41%
as	5.9K	2.39%
com.es	3K	1.20%
info	1.7K	0.68%
com.ar	1.3K	0.53%
nf	868	0.35%
click	689	0.28%

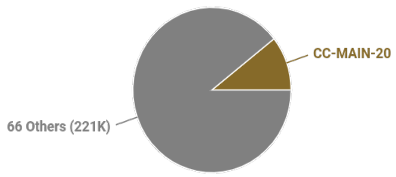
Documents size (in segments)

≤ 25 segments **81.09%** (201K documents)
> 25 segments **18.91%** (47K documents)



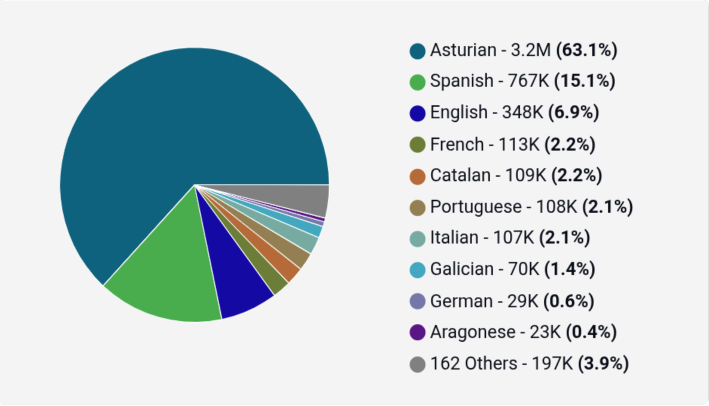
Document collections

CC = 92.45%
IA = 7.55%

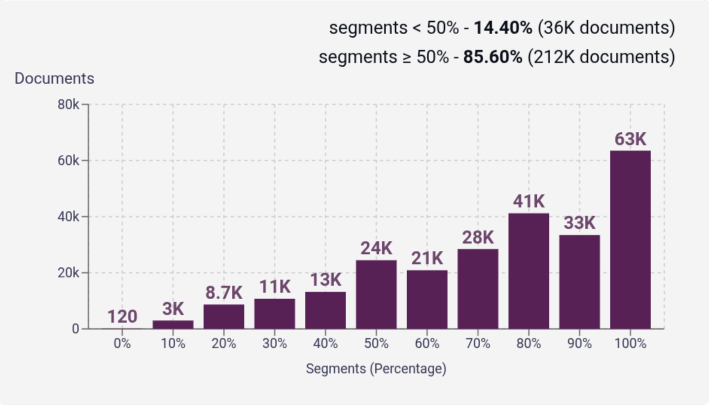


Language Distribution

Number of segments in the Asturian corpus

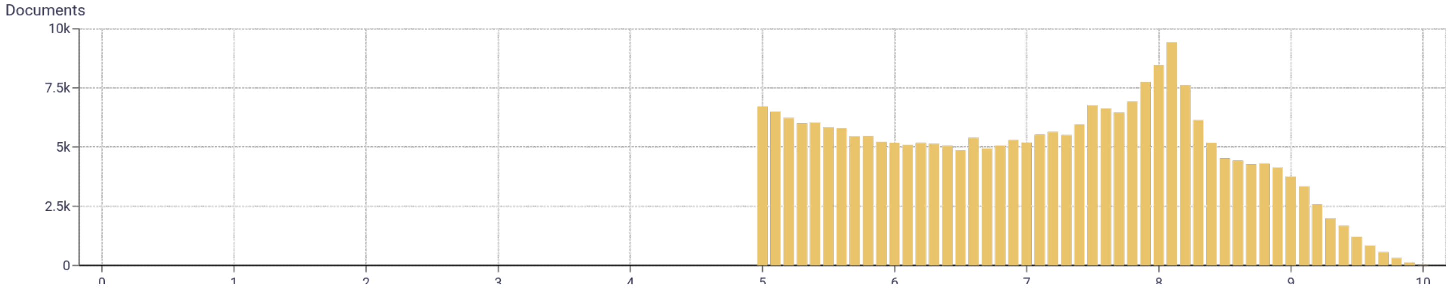


Percentage of segments in Asturian inside documents


segments < 50% - **14.40%** (36K documents)
segments ≥ 50% - **85.60%** (212K documents)

Distribution of documents by document score

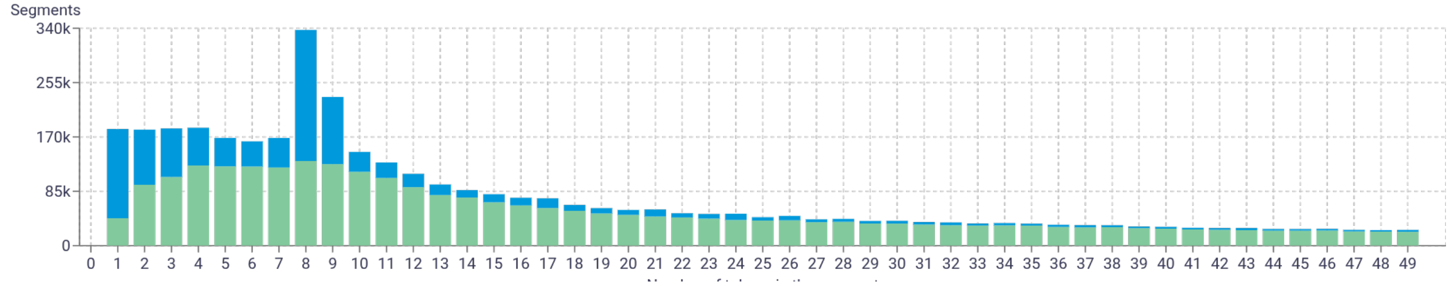
score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (248K documents)



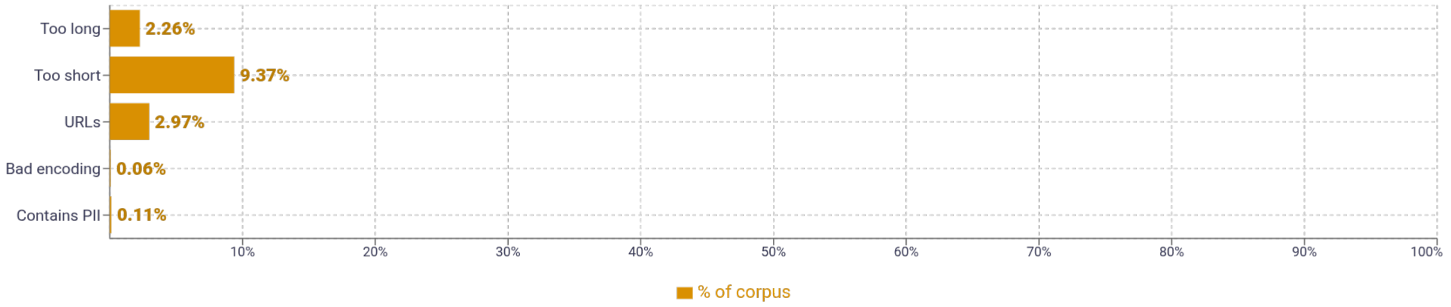
Segment length distribution by token

≤ 49 tokens = **3.8M** segments | **1.1M** duplicates

> 50 tokens = **1.3M** segments | **118K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	editar | 846,876 fonte | 418,675 e | 377,647 i | 369,310 san | 287,855
2	llingua asturiana | 39,620 enllaces esternos | 36,477 e r | 36,081 of the | 34,821 r e | 31,378
3	editar la fonte | 402,878 acueye conteníu multimedia | 19,628 e n t | 18,886 q u e | 14,496 p a r | 10,513
4	guía de la industria | 11,416 academia de la llingua | 11,071 llingua de la obra | 7,802 e n t e | 7,044 m e n t | 6,970
5	academia de la llingua asturiana | 9,546 wikimedia commons acueye conteníu multimedia | 6,021 ubicación y con la ayuda | 4,909 servicio y asi poder repostar | 4,909 poder repostar al menor coste | 4,909

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				