HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-cat_Latn	9/18/2025	Catalan (ca)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
26,411,847	459,934,979	269,507,645 (58.60 %)	15B	74,994,335,215	71.98 GB

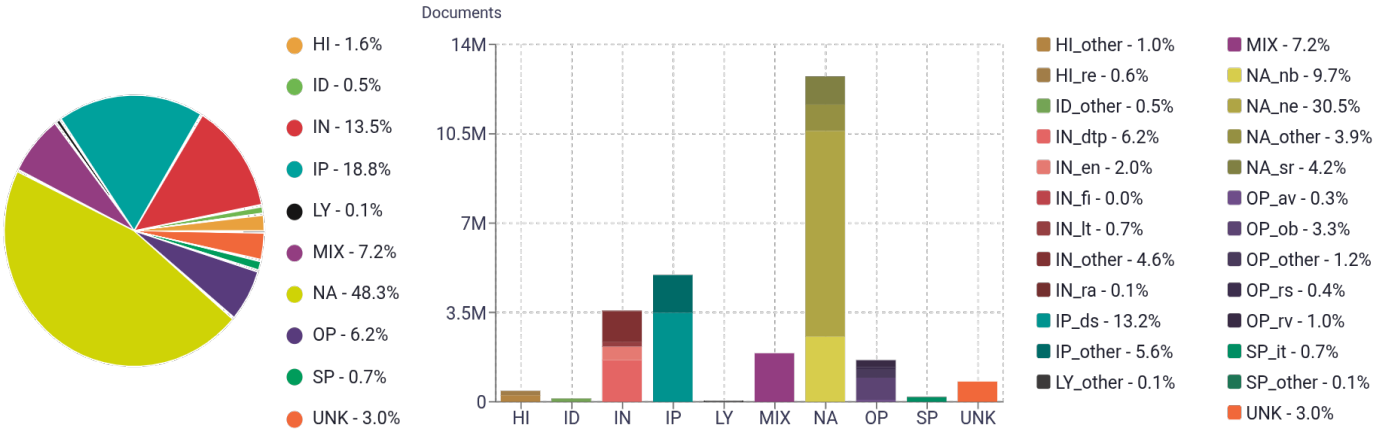
Top 10 domains

Domain	Docs	% of total
blogspot.com	1.8M	6.75%
wordpress.com	548K	2.08%
wikipedia.org	371K	1.40%
ara.cat	343K	1.30%
diaridegirona.cat	322K	1.22%
blogspot.com.es	302K	1.14%
regio7.cat	271K	1.03%
xtec.cat	262K	0.99%
ccma.cat	196K	0.74%
elpuntavui.cat	188K	0.71%

Top 10 TLDs

Domain	Docs	% of total
cat	12M	43.66%
com	8.6M	32.68%
org	2M	7.71%
es	1.6M	5.94%
net	676K	2.56%
ad	311K	1.18%
com.es	304K	1.15%
edu	298K	1.13%
info	234K	0.89%
eu	123K	0.47%

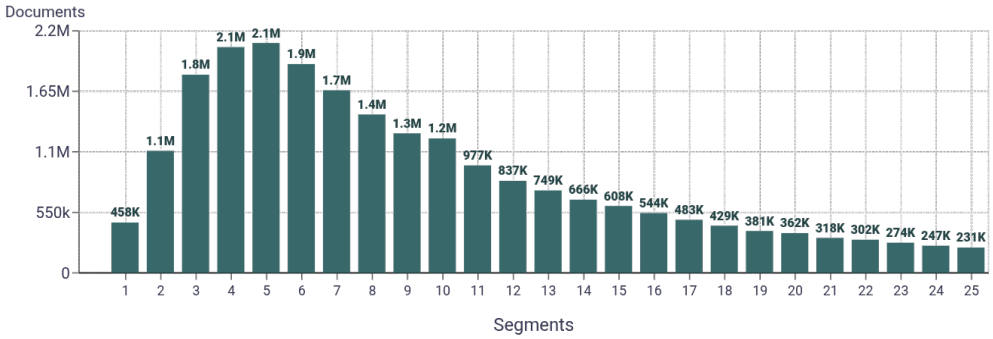
Register labels



MT:0.9% | 234K Documents

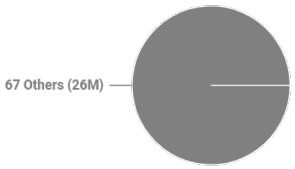
Documents size (in segments)

≤ 25 segments **84.82%** (22M documents)
> 25 segments **15.18%** (4M documents)



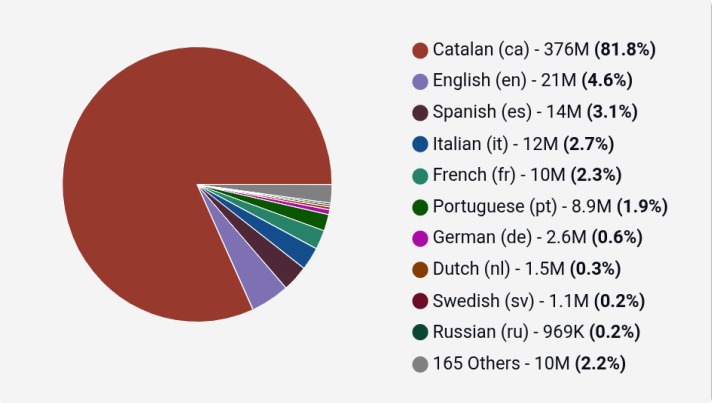
Document collections

CC = 93.18%
IA = 6.82%

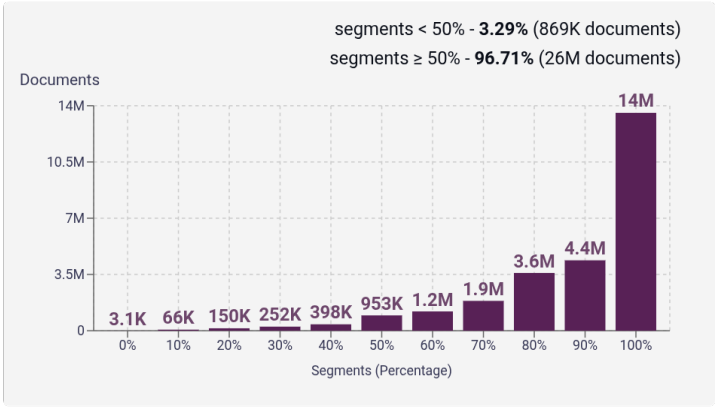


Language Distribution

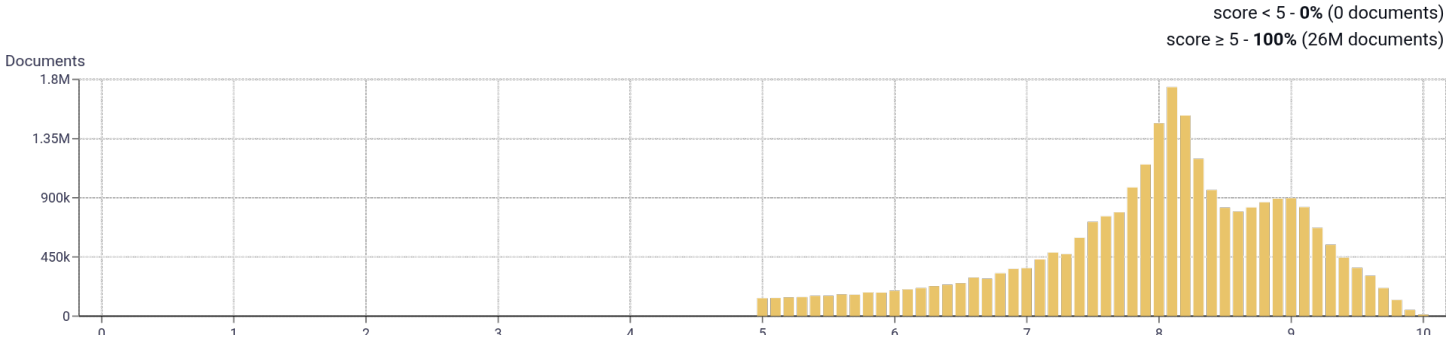
Number of segments in the Catalan (ca) corpus



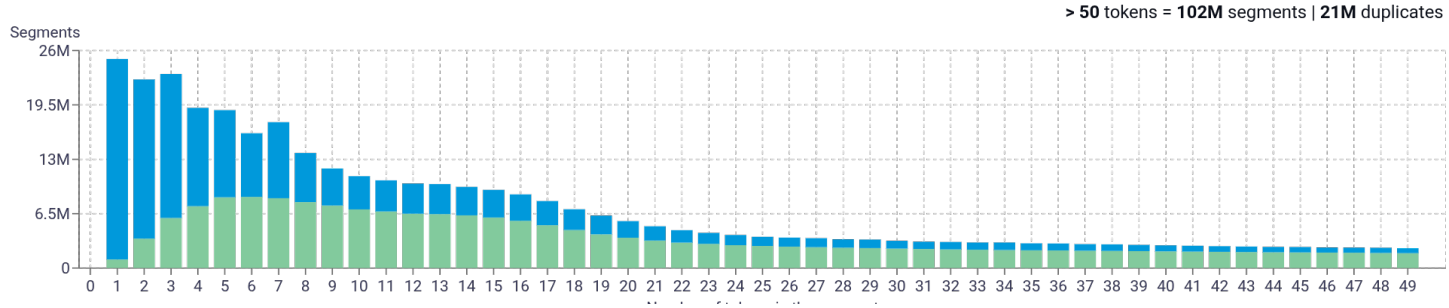
Percentage of segments in Catalan (ca) inside documents



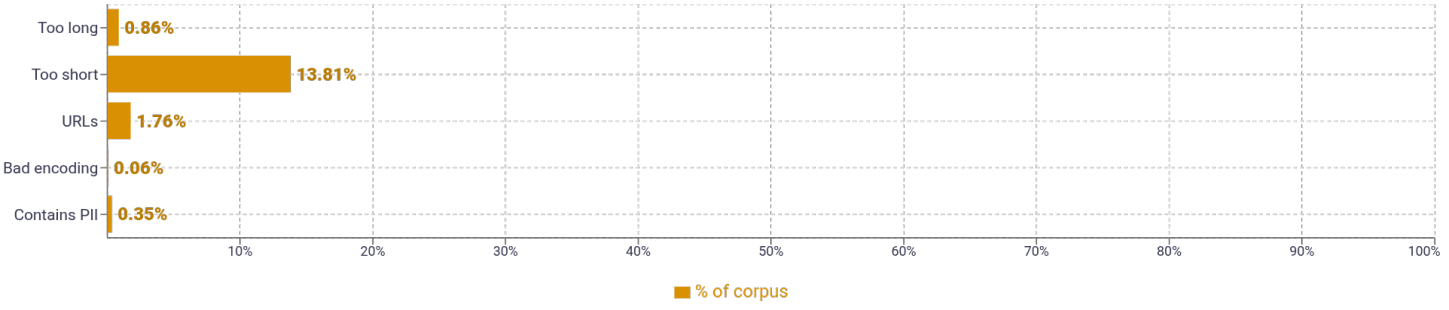
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	anys | 17,732,760 · cap | 17,341,491 · dia | 14,629,500 · any | 14,310,747 · persones | 12,735,896
2	medi ambient | 1,078,673 · estats units | 1,074,978 · xarxes socials | 1,059,995 · any passat | 1,033,816 · tindrà lloc | 975,017
3	cap de setmana | 1,635,133 · publica un comentari | 1,094,027 · generalitat de catalunya | 905,681 · dur a terme | 759,788 · nens i nenes | 677,899
4	president de la generalitat | 341,814 · té com a objectiu | 266,338 · universitat autònoma de barcelona | 212,400 · vilanova i la geltrú | 205,516 · centre de la ciutat | 204,915
5	universitat de les illes balears | 84,488 · través de les xarxes socials | 83,116 · afegit la noticia a favorits | 78,199 · millorar la qualitat de vida | 69,296 · superior de justícia de catalunya | 67,181

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				