HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
ita_Latn.jsonl.tsv	9/3/2025	Italian (it)

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
221,752,424	5,125,538,503	2,035,454,227 (39.71 %)	60.29%	151B	815,696,141,839	769.93 GB

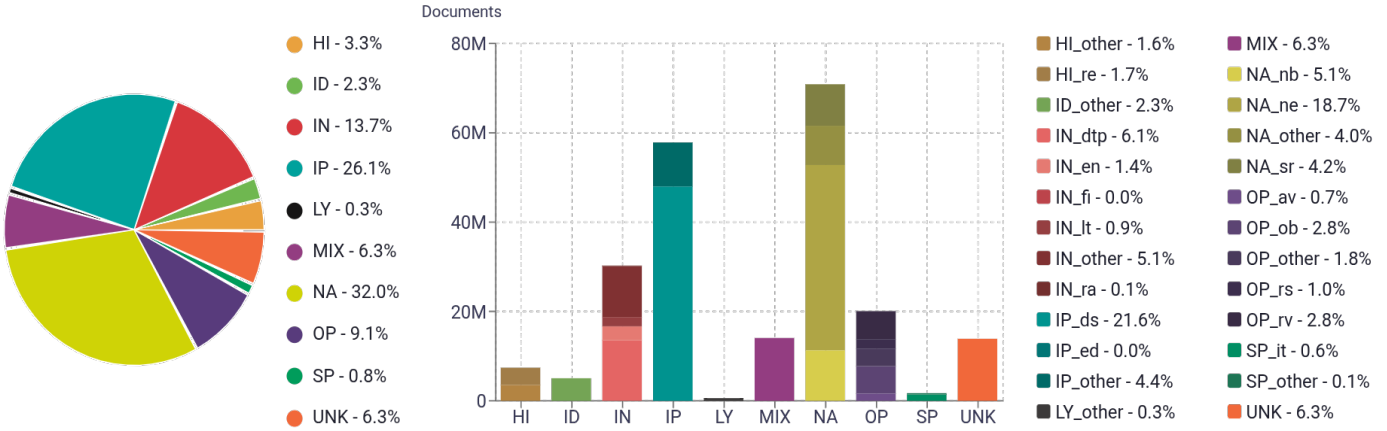
Top 10 domains

Domain	Docs	% of total
blogspot.com	6.2M	2.78%
blogspot.it	3.8M	1.71%
wordpress.com	2.3M	1.02%
wikipedia.org	2M	0.90%
kijiji.it	1.6M	0.70%
repubblica.it	956K	0.43%
altervista.org	814K	0.37%
corriere.it	759K	0.34%
tripadvisor.it	694K	0.31%
docplayer.it	637K	0.29%

Top 10 TLDs

Domain	Docs	% of total
it	125M	56.38%
com	59M	26.76%
org	9.9M	4.48%
net	7.5M	3.36%
eu	3.4M	1.52%
info	2.5M	1.13%
ch	2M	0.90%
tv	816K	0.37%
biz	424K	0.19%
de	412K	0.19%

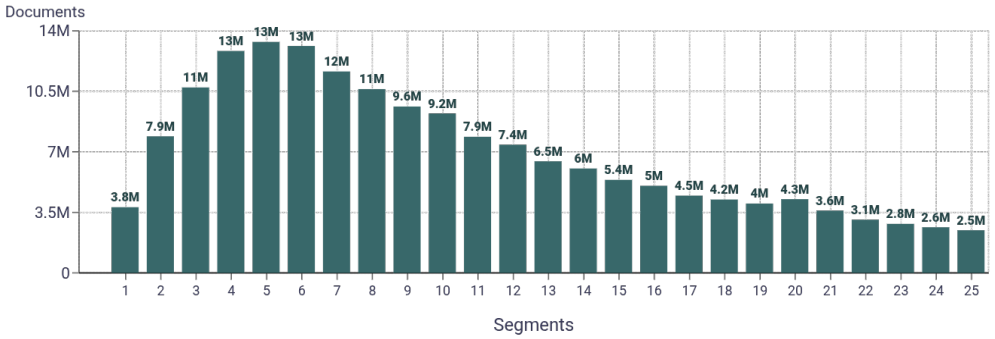
Register labels



MT:3.0% | 6.6M Documents

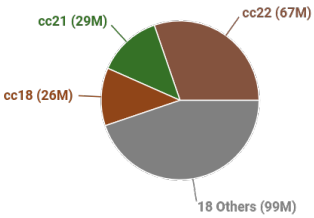
Documents size (in segments)

≤ 25 segments **77.86%** (173M documents)
> 25 segments **22.14%** (49M documents)



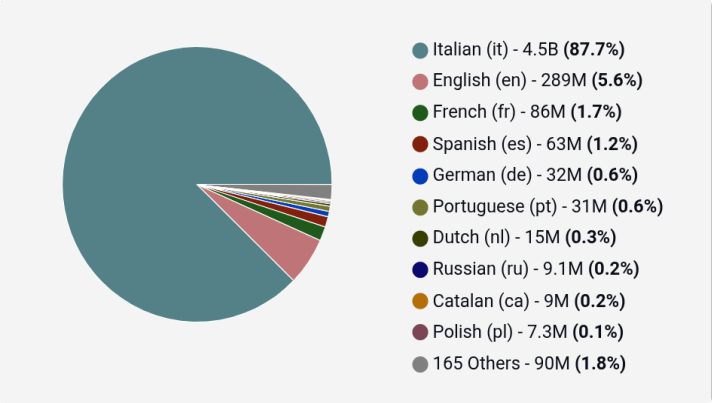
Document collections

CC = 67.63%
IA = 32.37%

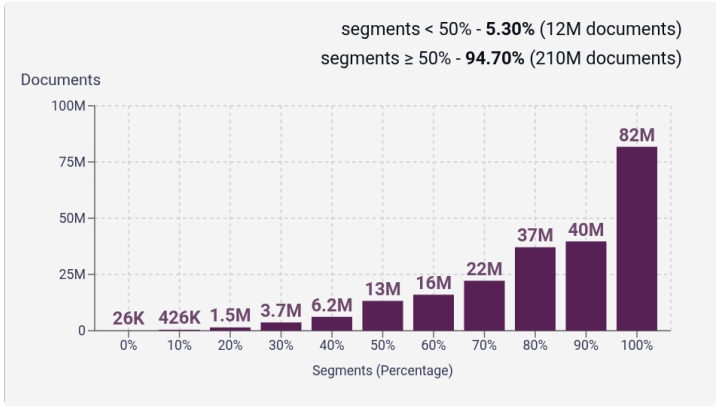


Language Distribution

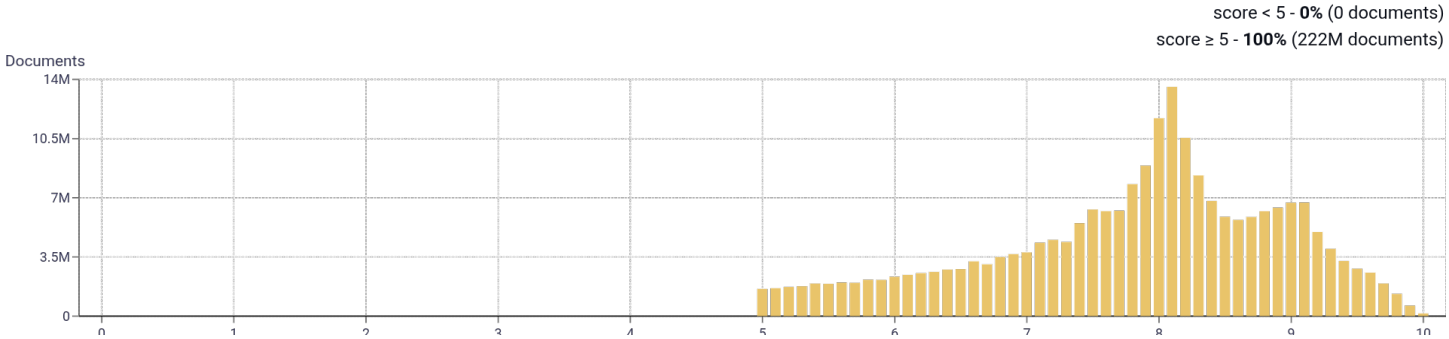
Number of segments in the Italian (it) corpus



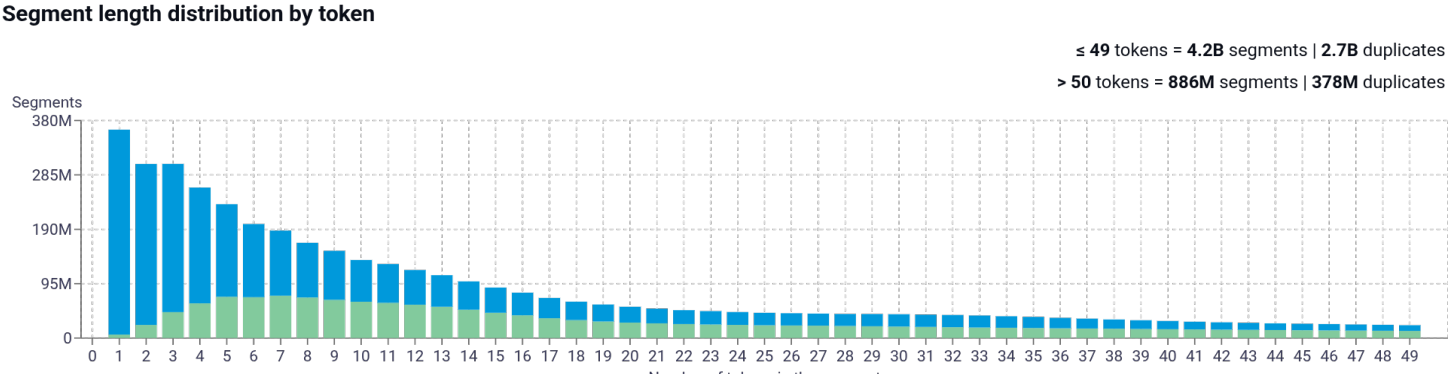
Percentage of segments in Italian (it) inside documents



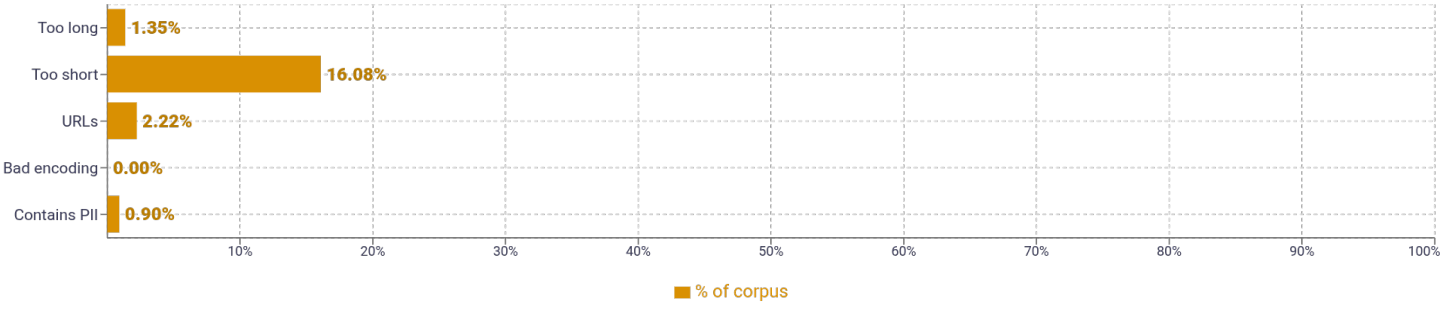
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	essere | 231,734,267 · solo | 175,187,073 · anni | 167,651,706 · due | 160,570,200 · prima | 155,603,051
2	può essere | 33,538,940 · possono essere | 18,532,851 · deve essere | 16,325,384 · dopo aver | 15,311,055 · stati uniti | 12,396,552
3	punto di vista | 11,354,749 · continua a leggere | 9,444,299 · milioni di euro | 7,339,613 · contatta l' utente | 6,167,234 · ancora una volta | 5,745,268
4	maggior parte dei casi | 1,161,914 · opzioni di acquistospedizione gratuita | 1,121,223 · link a questo post | 1,025,703 · olio extravergine di oliva | 905,787 · immagini o altri file | 828,953
5	italia che portano il nome | 1,213,272 · persone in italia che portano | 1,213,249 · contiene immagini o altri file | 827,989 · commons contiene immagini o altri | 827,393 · email è protetto dagli spambots | 761,867

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				