HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-vec_Latn	9/18/2025	Venetian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
102,317	1,316,709	976,562 (74.17 %)	55M	274,934,839	265.5 MB

Top 10 domains

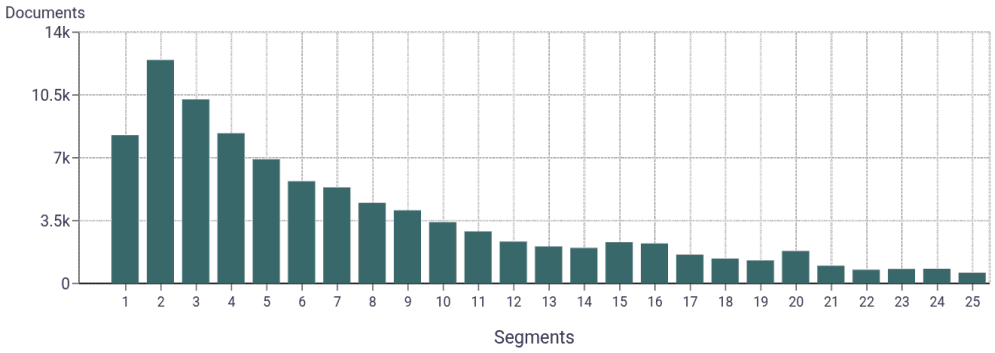
Domain	Docs	% of total
wikipedia.org	10K	9.95%
giuffre.it	1.7K	1.64%
gelocal.it	1.5K	1.49%
locatemyname.com	1.4K	1.41%
wikisource.org	1.4K	1.40%
wordpress.com	1K	0.99%
blogspot.com	846	0.83%
kiao.net	528	0.52%
larenadomila.it	514	0.50%
paperzz.com	489	0.48%

Top 10 TLDs

Domain	Docs	% of total
it	49K	48.18%
com	22K	21.28%
org	16K	15.49%
net	3.4K	3.37%
eu	1.7K	1.71%
info	1.4K	1.33%
de	537	0.52%
venezia.it	460	0.45%
la	453	0.44%
blog	431	0.42%

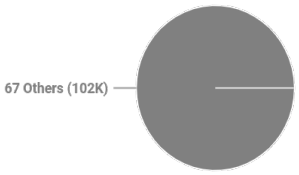
Documents size (in segments)

≤ 25 segments **90.95%** (93K documents)
> 25 segments **9.05%** (9.3K documents)



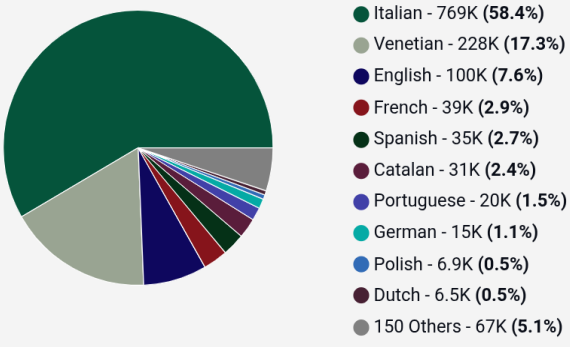
Document collections

CC = 88.71%
IA = 11.29%



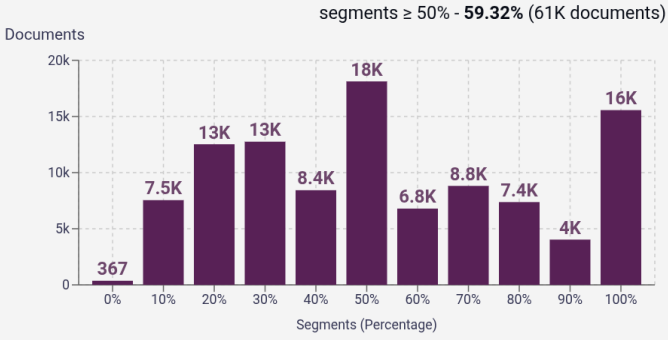
Language Distribution

Number of segments in the Venetian corpus



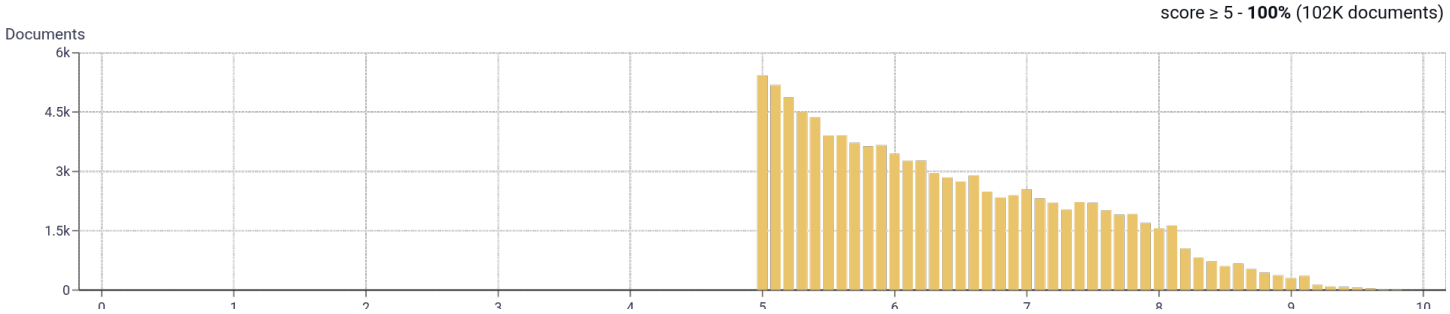
Percentage of segments in Venetian inside documents

segments < 50% - **40.68%** (42K documents)
segments ≥ 50% - **59.32%** (61K documents)



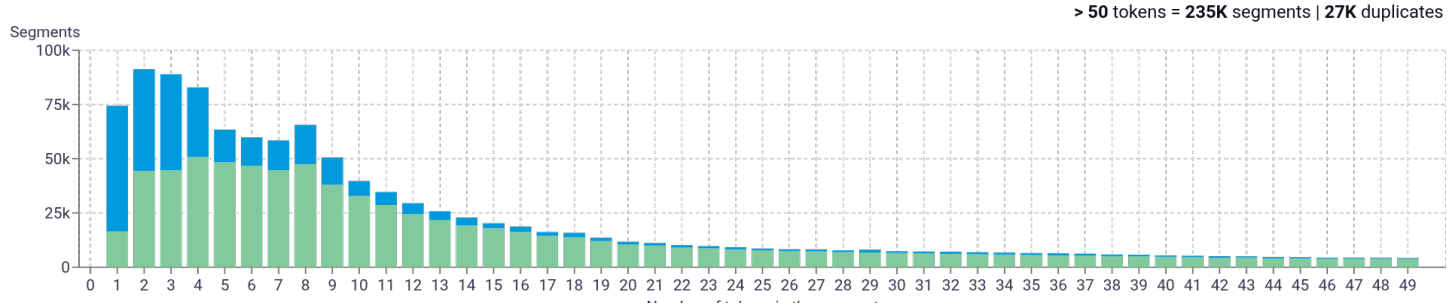
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (102K documents)

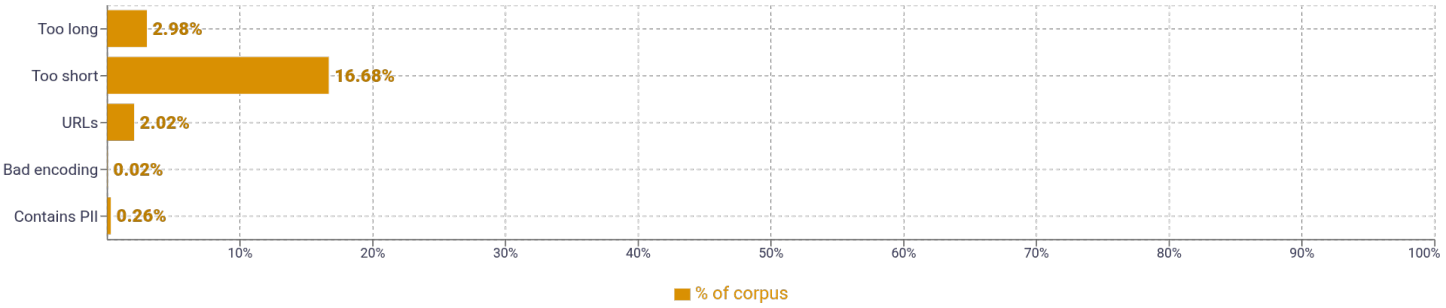


Segment length distribution by token

≤ 49 tokens = **1.1M** segments | **314K** duplicates
> 50 tokens = **235K** segments | **27K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	d | 148,960 · maria | 125,956 · francesco | 107,769 · s | 107,162 · giovanni | 105,294
2	santa maria | 12,210 · canbia sorxente | 8,831 · candidato sindaco | 7,428 · anna maria | 7,374 · of the | 5,128
3	canbia el còdaxe | 14,955 · canbia el còdexe | 4,827 · x x x | 3,731 · università degli studi | 3,106 · carta da parati | 3,015
4	x x x x | 3,030 · popular firstnames for surname | 1,445 · rank in italy is | 1,443 · popularity rank in italy | 1,443 · chiesa di santa maria | 1,146
5	x x x x x | 2,673 · the popularity rank in italy | 1,443 · popularity rank in italy is | 1,443 · trovati usando i seguenti criteri | 1,088 · stati trovati usando i seguenti | 1,088

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				