HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-srd_Latn	9/18/2025	Sardinian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
66,660	791,941	602,167 (76.04 %)	35M	174,508,820	170.06 MB

Top 10 domains

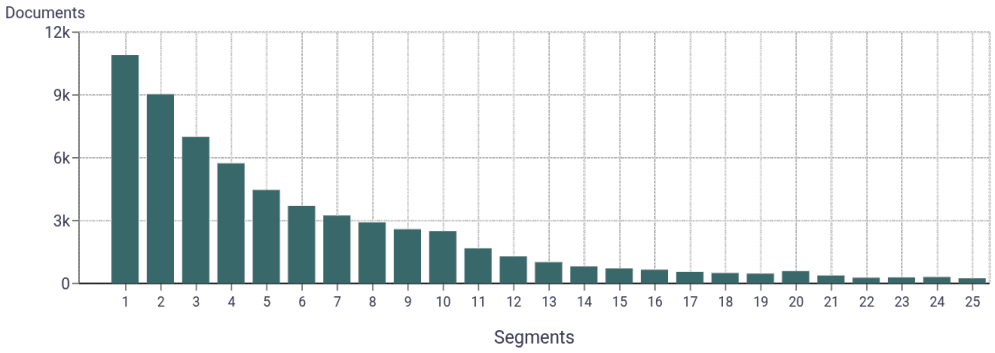
Domain	Docs	% of total
sardegnacultura.it	16K	24.54%
nor-web.eu	5.5K	8.22%
wikipedia.org	4.8K	7.15%
ilminuto.info	4.3K	6.47%
lacanas.it	2.1K	3.13%
salimbasarda.net	1.5K	2.32%
blogspot.com	1.2K	1.81%
wordpress.com	1.1K	1.65%
sagazeta.info	1K	1.54%
academiadesusar...	961	1.44%

Top 10 TLDs

Domain	Docs	% of total
it	33K	49.34%
org	7.2K	10.74%
com	6.6K	9.84%
eu	6.5K	9.79%
info	5.8K	8.72%
net	3.6K	5.36%
co	1.3K	1.93%
or.it	647	0.97%
xyz	277	0.42%
de	108	0.16%

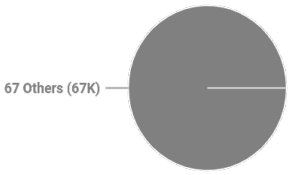
Documents size (in segments)

≤ 25 segments **92.71%** (62K documents)
> 25 segments **7.29%** (4.9K documents)



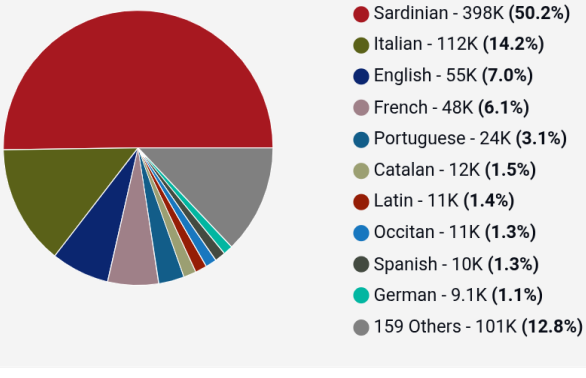
Document collections

CC = 96.33%
IA = 3.67%



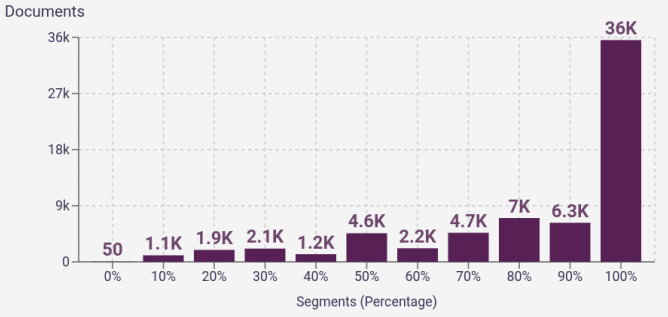
Language Distribution

Number of segments in the Sardinian corpus



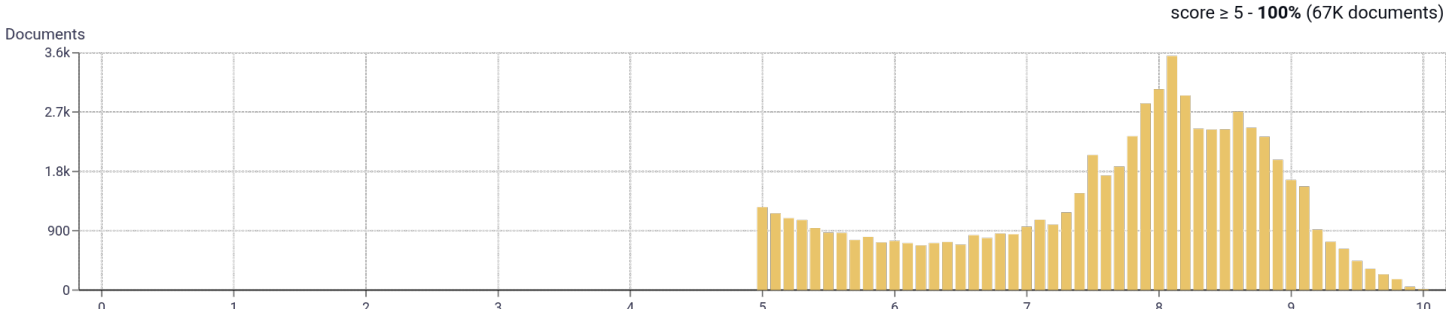
Percentage of segments in Sardinian inside documents

segments < 50% - **9.57%** (6.4K documents)
segments ≥ 50% - **90.43%** (60K documents)



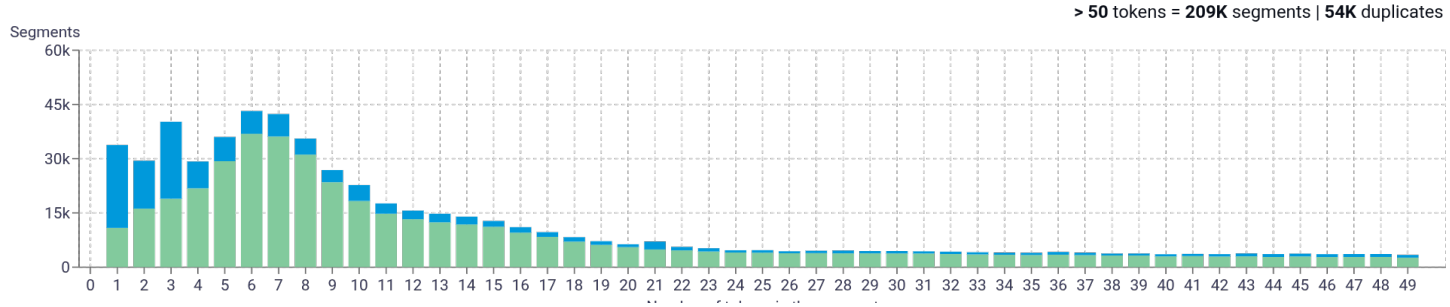
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (67K documents)

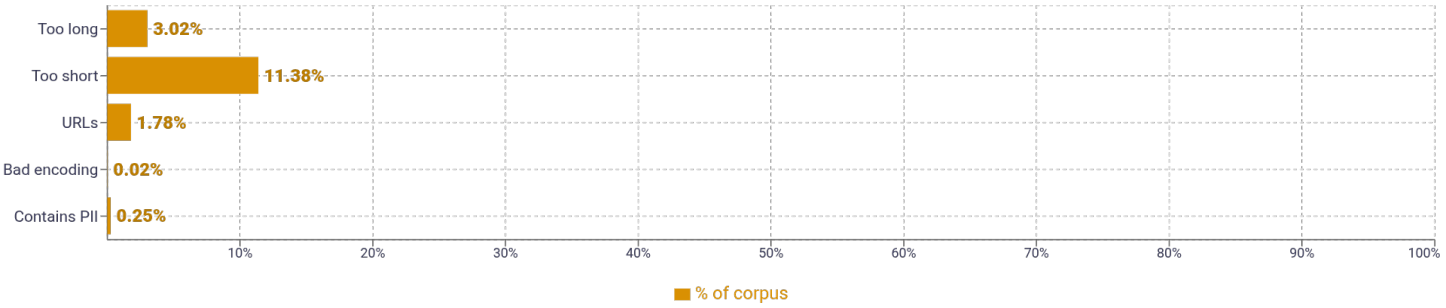


Segment length distribution by token

≤ 49 tokens = **583K** segments | **136K** duplicates
> 50 tokens = **209K** segments | **54K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	sa | 666,682 · s | 485,915 · est | 284,938 · si | 232,719 · sos | 197,588
2	tradutziones frantzesu | 41,342 · cun sa | 21,194 · ètimu srd | 18,232 · ingresu to | 16,839 · sa limba | 16,645
3	sinònimos e contràrios | 40,029 · synonyms e antonyms | 11,934 · sinonimi e contrari | 11,482 · sa limba sarda | 5,954 · còdighe de orìgine | 5,732
4	maneras de nàrrere csn | 4,916 · realizzata col contributo della | 2,777 · contributo della regione sardegna | 2,777 · col contributo della regione | 2,777 · attività realizzata col contributo | 2,776
5	modifica su còdighe de orìgine | 5,675 · realizzata col contributo della regione | 2,777 · col contributo della regione sardegna | 2,777 · attività realizzata col contributo della | 2,776 · acadèmia de su sardu onlus | 1,243

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				