HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-scn_Latn	9/18/2025	Sicilian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
91,611	2,037,136	1,684,258 (82.68 %)	76M	367,108,035	361.86 MB

Top 10 domains

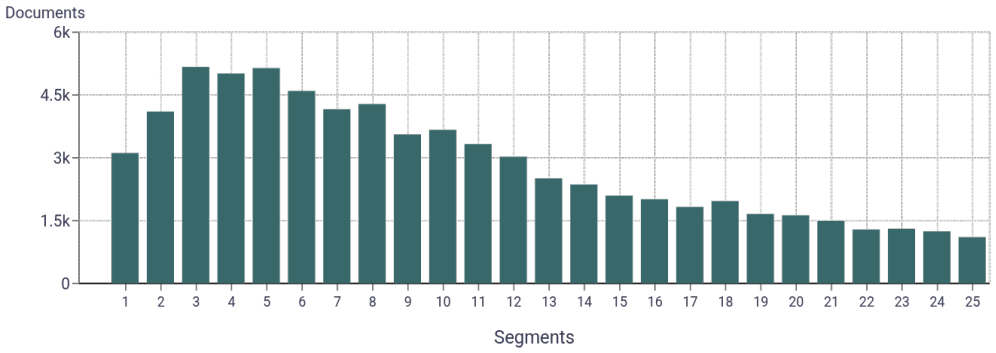
Domain	Docs	% of total
wikipedia.org	8.5K	9.31%
eturbonews.com	2.6K	2.84%
vsaduidoma.com	1.5K	1.68%
ihorror.com	1.4K	1.57%
julinse.com	1.2K	1.35%
martech.zone	1.2K	1.28%
tempicorsica.com	1.2K	1.27%
rayhaber.com	938	1.02%
cuncezzione.com	900	0.98%
blogspot.com	770	0.84%

Top 10 TLDs

Domain	Docs	% of total
com	60K	65.46%
org	12K	13.01%
it	3.3K	3.57%
corsica	2.9K	3.14%
net	2.1K	2.32%
pt	1.9K	2.02%
zone	1.2K	1.28%
fr	981	1.07%
ru	767	0.84%
pro	575	0.63%

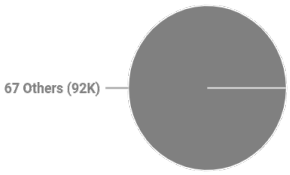
Documents size (in segments)

≤ 25 segments **78.23%** (72K documents)
> 25 segments **21.77%** (20K documents)



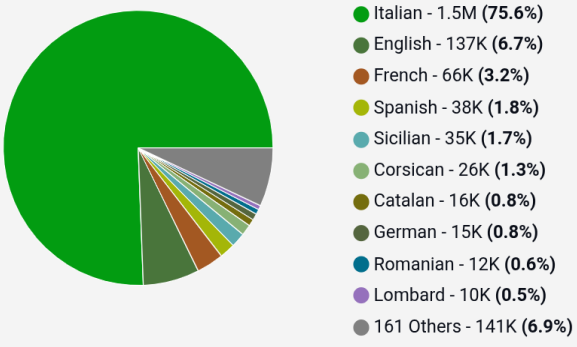
Document collections

CC = 94.60%
IA = 5.40%



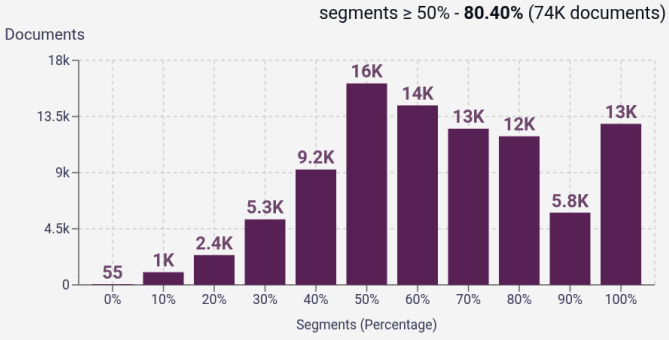
Language Distribution

Number of segments in the Sicilian corpus



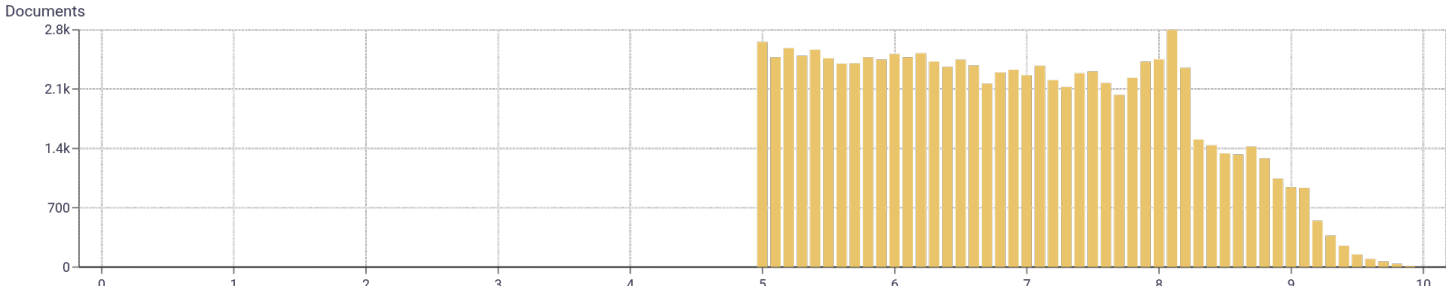
Percentage of segments in Sicilian inside documents

segments < 50% - **19.60%** (18K documents)
segments ≥ 50% - **80.40%** (74K documents)



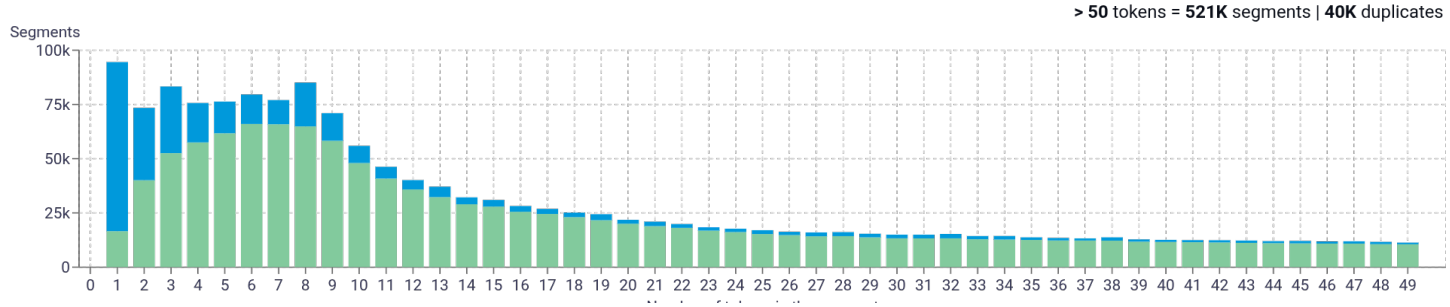
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (92K documents)

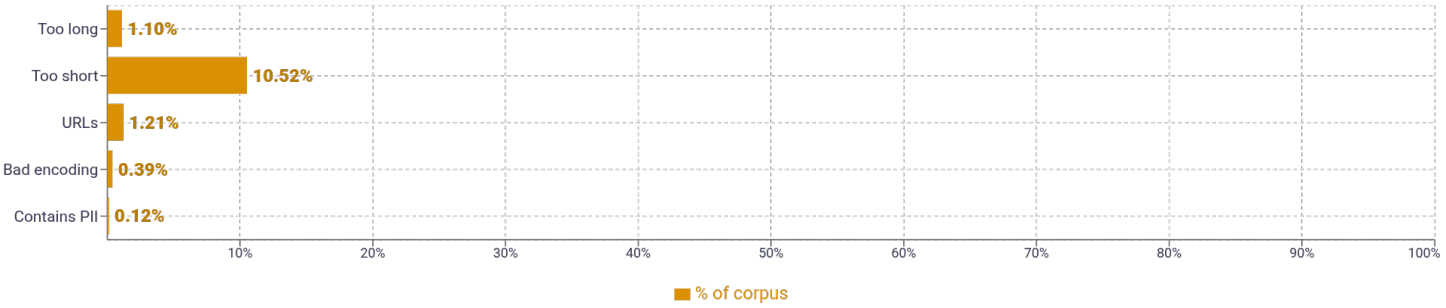


Segment length distribution by token

≤ 49 tokens = **1.5M** segments | **314K** duplicates
> 50 tokens = **521K** segments | **40K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	à | 1,236,223 · hè | 840,102 · chì | 803,085 · cù | 311,031 · hà | 304,483
2	nantu à | 154,266 · chì hè | 54,320 · ùn hè | 49,353 · ciò chì | 44,985 · hè micca | 43,531
3	ùn hè micca | 36,989 · ùn sò micca | 15,688 · cancia la surgenti | 14,428 · ùn ci hè | 12,086 · ùn hà micca | 8,925
4	ùn ci hè micca | 4,923 · chì ùn hè micca | 3,021 · chì ùn sò micca | 2,535 · ùn hè micca solu | 2,532 · the first to comment | 2,313
5	sianu the first to comment | 2,309 · story plus untold biography facts | 1,052 · childhood story plus untold biography | 1,044 · maiò parte di a ghjente | 698 · nantu à u situ web | 626

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				