

General overview

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

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

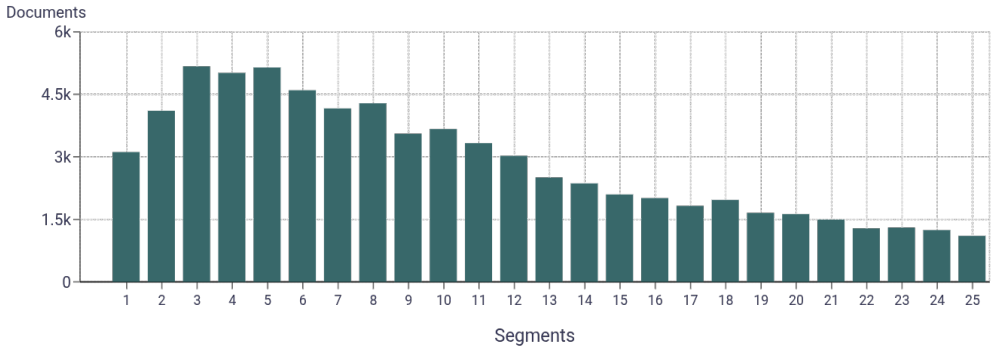
Top 10 domains

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%

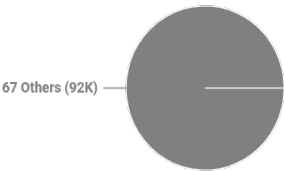
Documents size (in segments) ⓘ



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

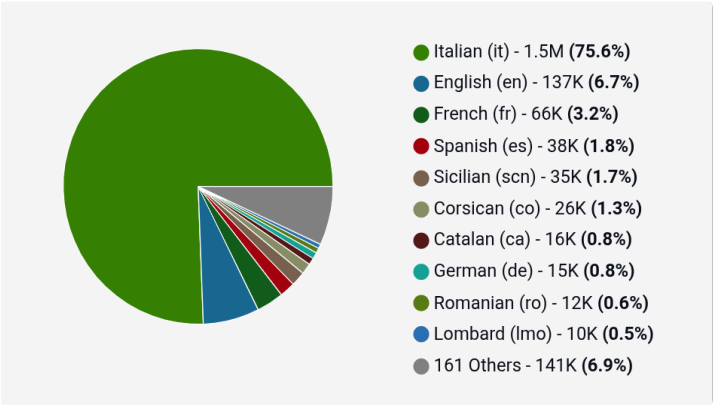
Document collections

CC = 94.60%  
IA = 5.40%

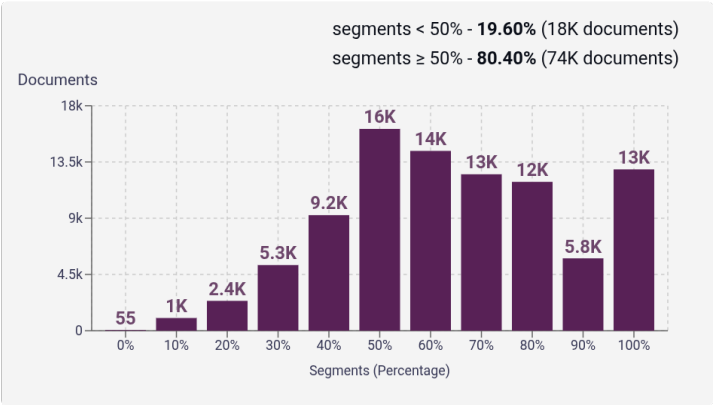


Language Distribution

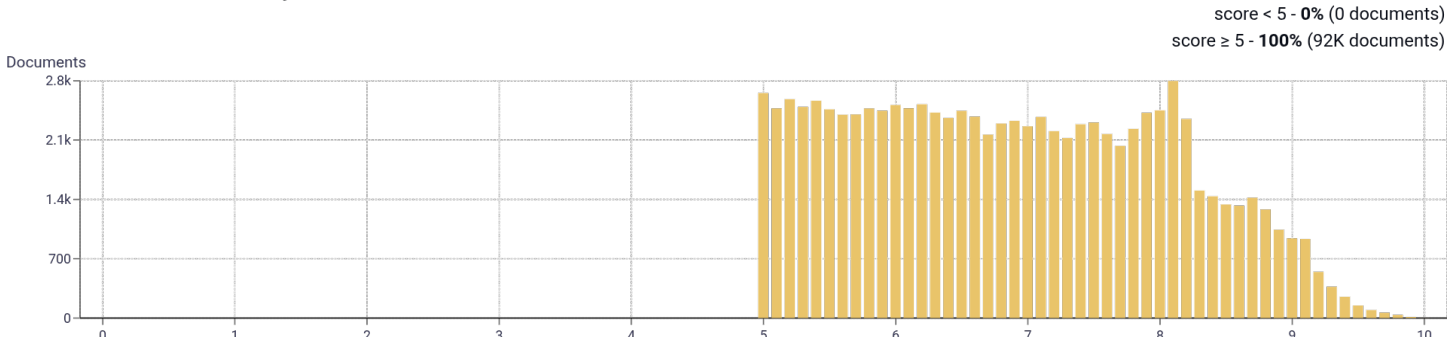
Number of segments in the Sicilian (scn) corpus



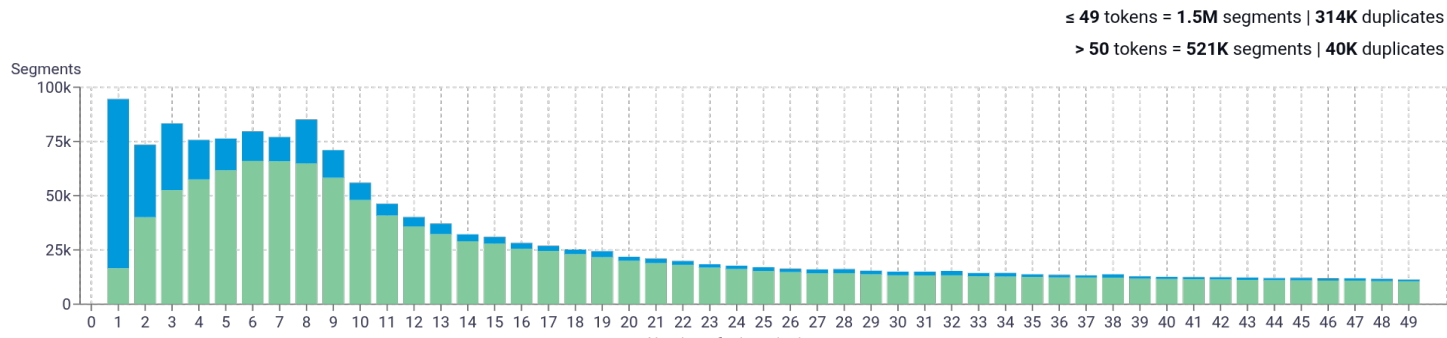
Percentage of segments in Sicilian (scn) inside documents



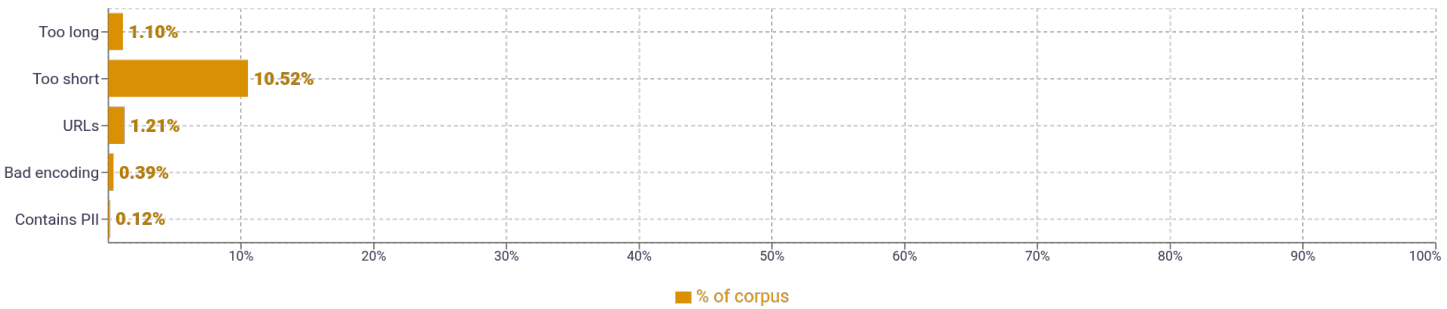
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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

# 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  $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$ , 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	ntp
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				