

General overview

Corpus	Date	Language
hplt-v3-mit_Latn	9/18/2025	Maltese (mt)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
752,735	17,217,418	10,768,498 (62.54 %)	408M	2,445,294,159	2.37 GB

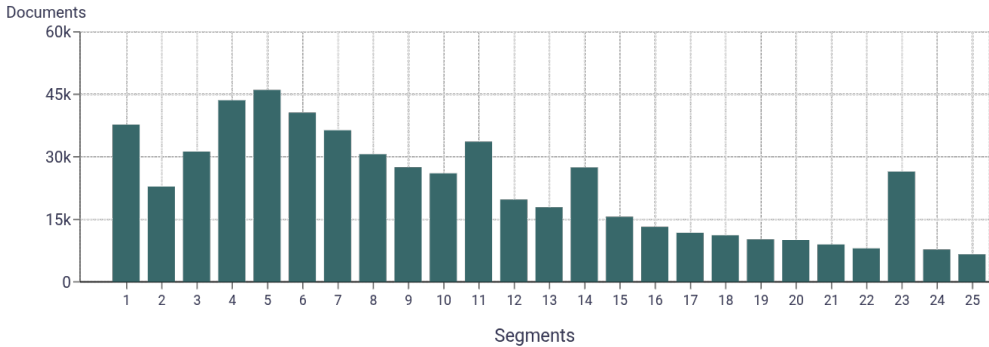
Top 10 domains

Domain	Docs	% of total
airbnb.com	101K	13.41%
newsbook.com.mt	46K	6.16%
europa.eu	39K	5.23%
jiffyrando.com	19K	2.57%
tvm.com.mt	19K	2.55%
one.com.mt	19K	2.51%
inewsmalta.com	16K	2.12%
talk.mt	14K	1.81%
itsmygame.org	11K	1.52%
laikosblog.org	11K	1.51%

Top 10 TLDs

Domain	Docs	% of total
com	413K	54.91%
com.mt	112K	14.92%
org	61K	8.16%
eu	54K	7.15%
mt	39K	5.22%
net	13K	1.72%
org.mt	8.4K	1.12%
co	5.6K	0.75%
zone	4.6K	0.61%
de	3.8K	0.51%

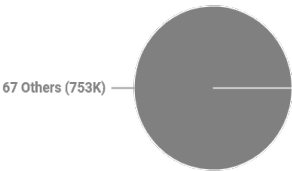
Documents size (in segments) ⓘ



≤ 25 segments 75.91% (571K documents)
> 25 segments 24.09% (181K documents)

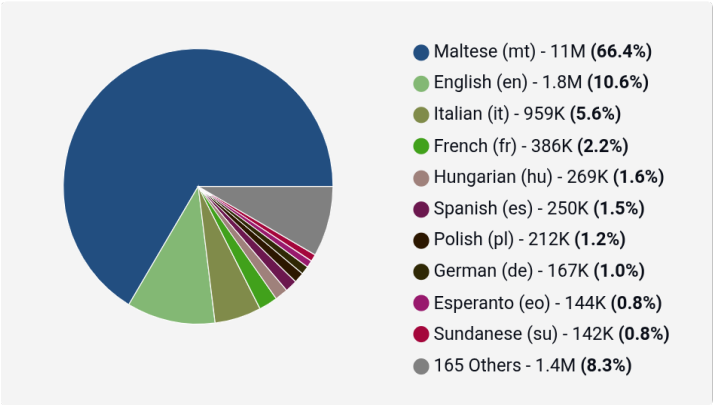
Document collections

CC = 85.34%
IA = 14.66%

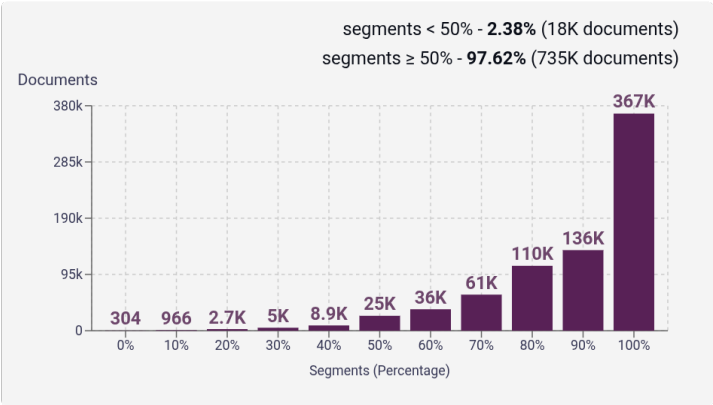


Language Distribution

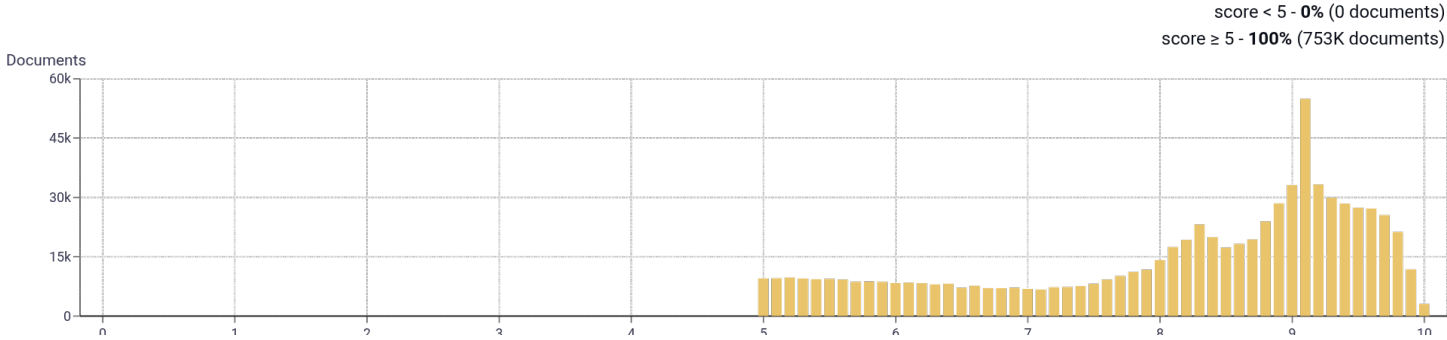
Number of segments in the Maltese (mt) corpus



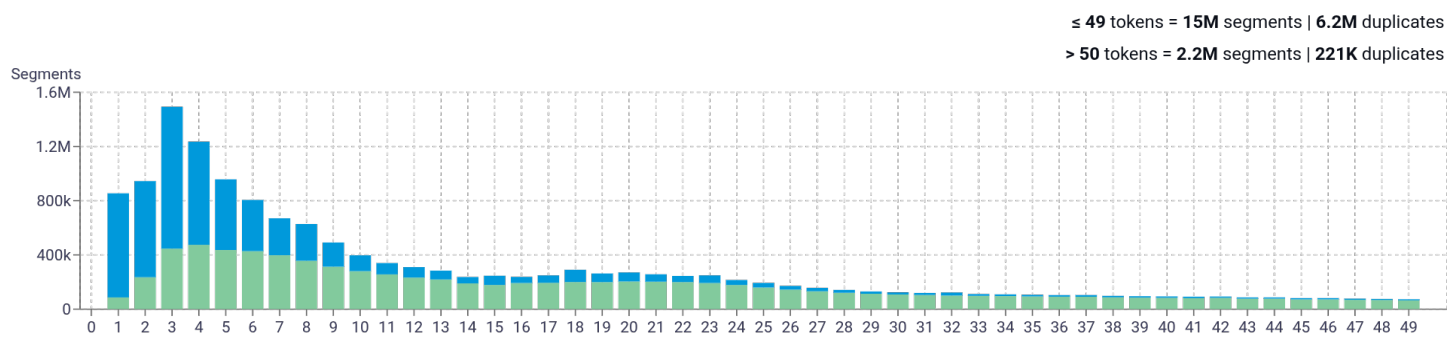
Percentage of segments in Maltese (mt) inside documents



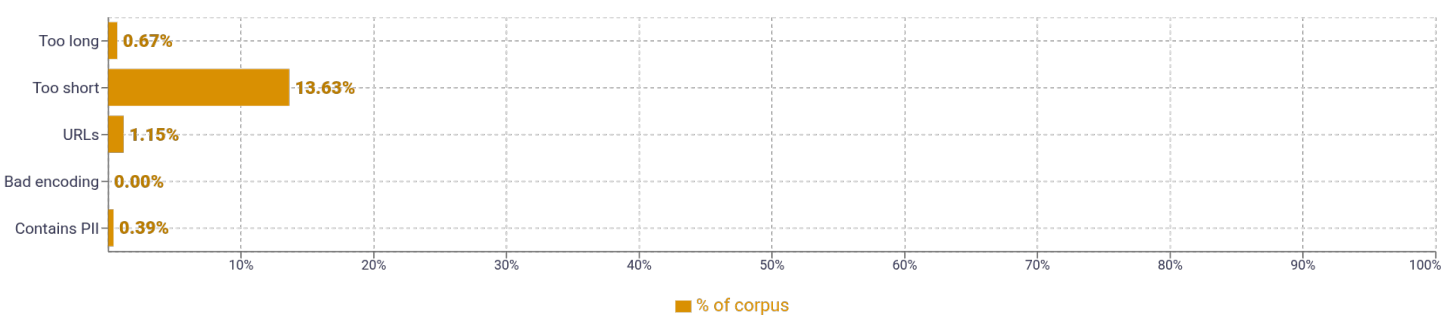
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	f 2,312,812 b 2,265,802 ġewwa 1,355,706 skont 1,118,624 reviews 911,951	📄
2	rating medju 850,356 skont dan-numru 850,353 b 'mod 395,688 dar ġewwa 298,959 f 'dan 186,567	📄
3	unità tal-kiri ġewwa 163,625 postijiet fejn toqgħod 98,367 mfahħra ħafna għall-post 69,257 sib u bbukkja 68,405 b 'mod partikolari 66,746	📄
4	ikseb l-isparzu li għandek 65,348 għall-vaganzi għal kull stil 65,348 sib u bbukkja postijiet 50,123 postijiet fejn toqgħod b 34,132 postijiet uniċi fejn toqgħod 31,992	📄
5	kirjiet għall-vaganzi għal kull stil 65,348 mfahħra ħafna għall-post fejn jinsabu 37,103 sib u bbukkja postijiet uniċi 32,978 uniċi fejn toqgħod fuq airbnb 31,991 bbukkja postijiet uniċi fejn toqgħod 31,991	📄

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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>, , , 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				