

General overview

Corpus	Date	Language
hplt-v3-el_Grek	9/18/2025	Greek (el)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
87,390,814	1,873,537,442	956,631,122 (51.06 %)	51B	288,268,678,785	478.68 GB

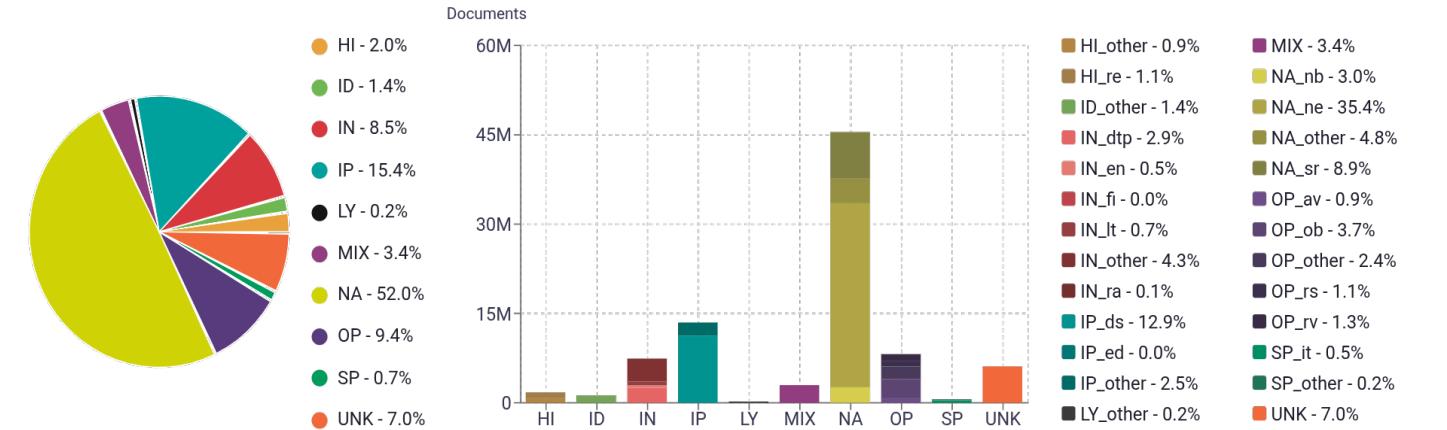
Top 10 domains

Domain	Docs	% of total
blogspot.com	5.5M	6.33%
blogspot.gr	2.5M	2.91%
wordpress.com	1.3M	1.52%
inewsgr.com	660K	0.75%
docplayer.gr	586K	0.67%
liverster.gr	479K	0.55%
onsports.gr	447K	0.51%
newsit.gr	403K	0.46%
sch.gr	355K	0.41%
gazzetta.gr	297K	0.34%

Top 10 TLDs

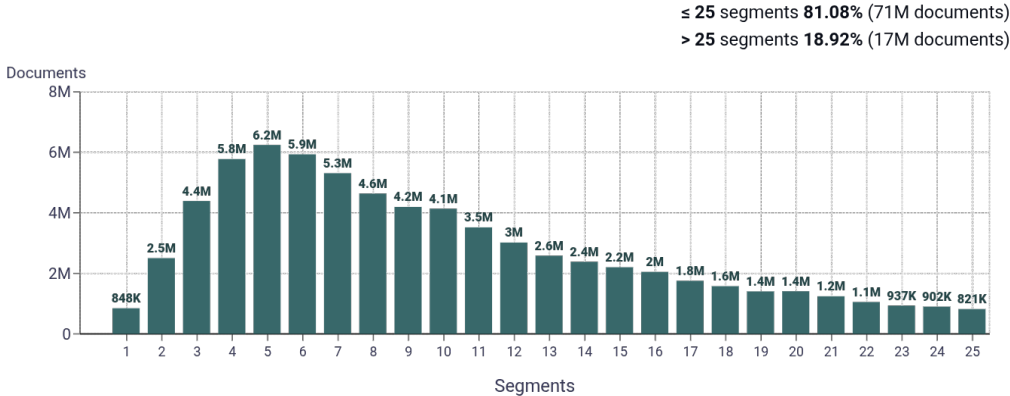
Domain	Docs	% of total
gr	61M	69.27%
com	18M	20.89%
org	1.3M	1.45%
net	1.2M	1.39%
com.cy	1.1M	1.29%
eu	845K	0.97%
com.gr	603K	0.69%
info	363K	0.42%
news	263K	0.30%
gov.gr	199K	0.23%

Register labels

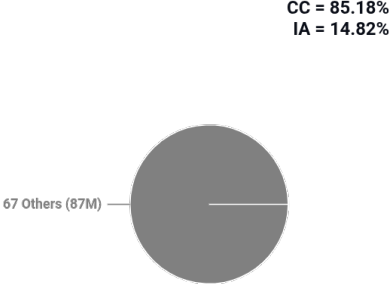


MT:3.2% | 2.8M Documents

Documents size (in segments) ⓘ

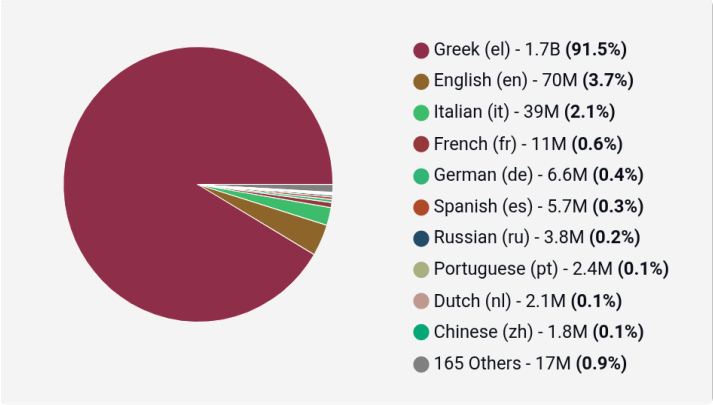


Document collections

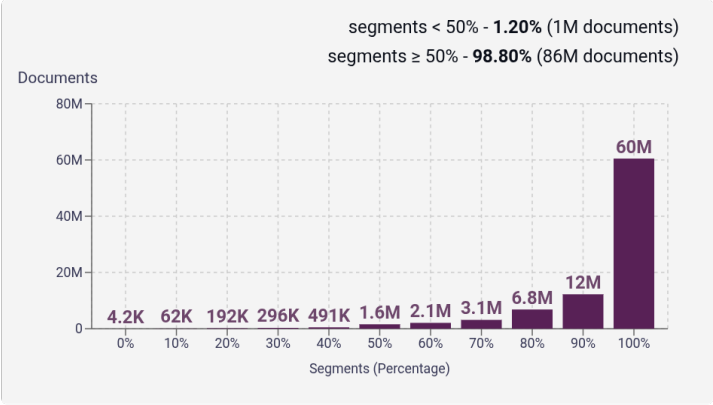


Language Distribution

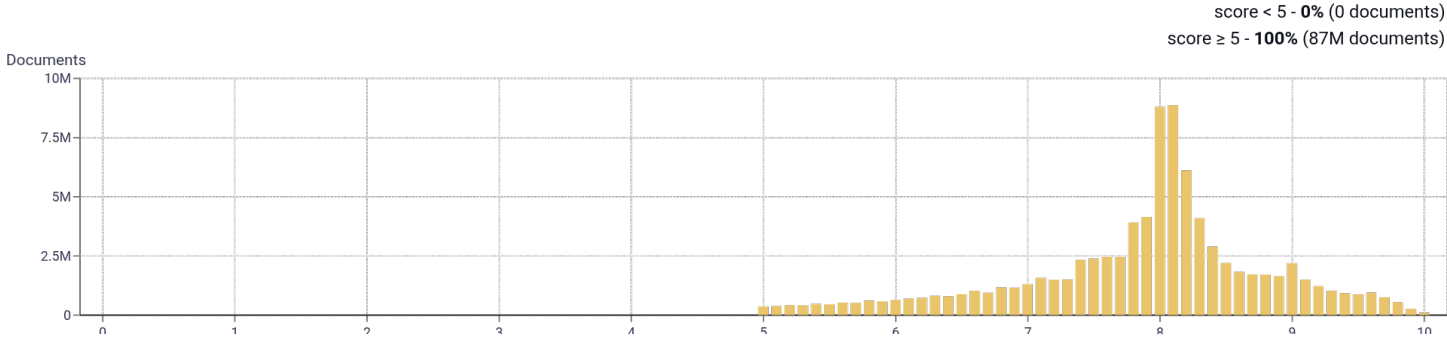
Number of segments in the Greek (el) corpus



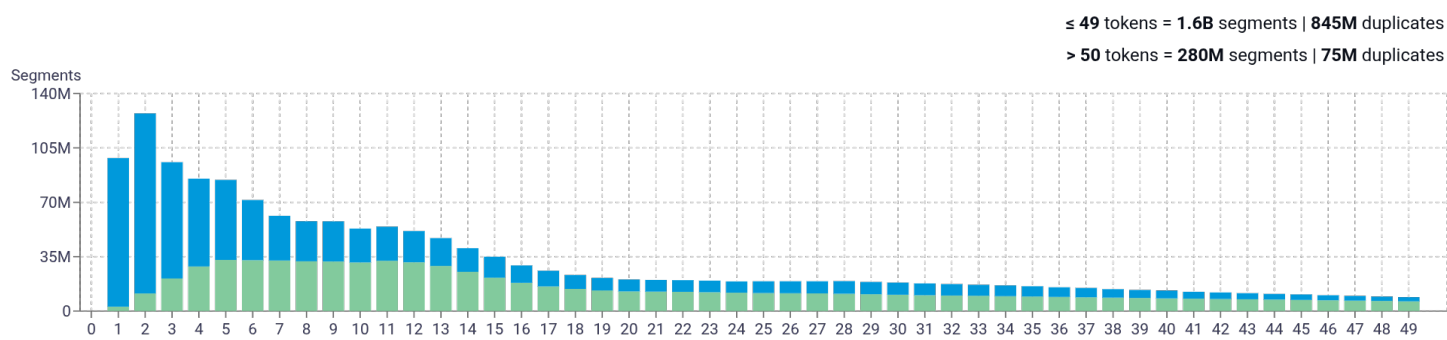
Percentage of segments in Greek (el) inside documents



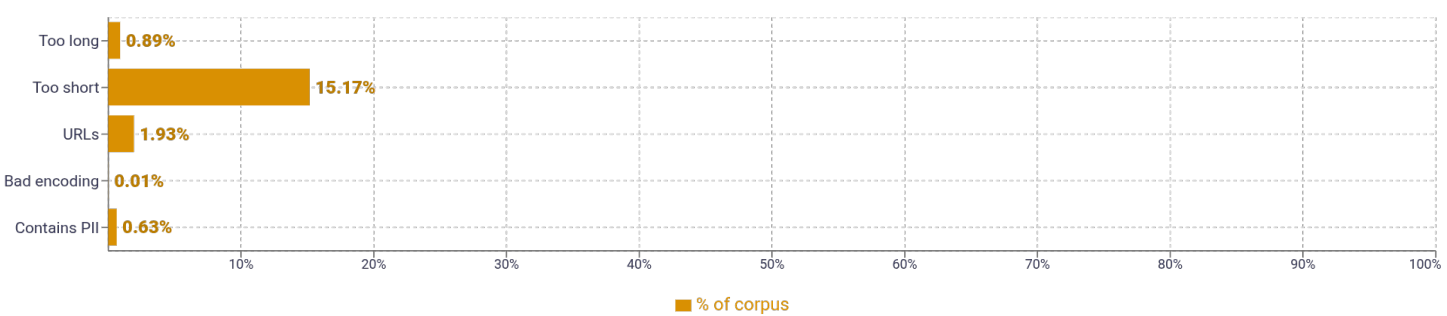
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	της   734,268,320    από   517,671,604    είναι   385,676,230    τους   269,884,253    τη   260,742,173	
2	από τη   30,721,350    διαβάστε περισσότερα   26,959,126    από τους   23,310,180    πριν από   13,082,562    είναι ένα   11,721,080	
3	τη διάρκεια της   5,674,492    μπορεί να είναι   4,541,165    πρέπει να είναι   3,373,865    από την άλλη   3,192,217 ειδήσεις σχετικές αναφορές   2,880,043	
4	ένα από τα πιο   1,316,218    από την άλλη πλευρά   721,598    τρόπο με τον οποίο   667,288    από την πρώτη στιγμή   634,594 από την πλευρά της   626,218	
5	ουδμία ευθύνη εκ του νόμου   543,503    περίπτωση που θεωρείτε πως θίγεστε   508,008    θεωρείτε πως θίγεστε από κάποιο   507,963 θίγεστε από κάποιο εξ αυτών   494,091    ευθύνη εκ του νόμου φέρει   483,125	

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 of 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	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				