HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
ell_Grek.jsonl.tsv	7/14/2025	Greek (el)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
70,328,862	1,847,935,923	608,316,724 (32.92 %)	50B	281,755,294,923	468.4 GB

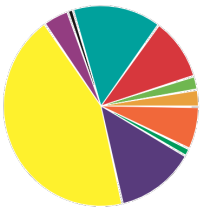
Top 10 domains

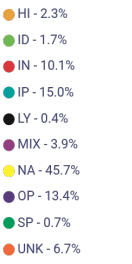
Domain	Docs	% of total
blogspot.gr	8.5M	12.15%
blogspot.com	6.6M	9.32%
wordpress.com	1.5M	2.18%
docplayer.gr	547K	0.78%
livester.gr	504K	0.72%
inewsgr.com	490K	0.70%
wikipedia.org	446K	0.63%
onsports.gr	314K	0.45%
blogspot.be	291K	0.41%

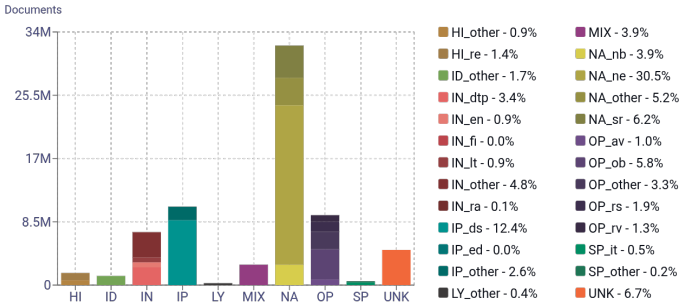
Top 10 TLDs

Domain	Docs	% of total
gr	47M	66.34%
com	16M	22.20%
org	1.2M	1.72%
net	828K	1.18%
eu	628K	0.89%
com.cy	619K	0.88%
com.gr	506K	0.72%
be	302K	0.43%
info	254K	0.36%

Register labels



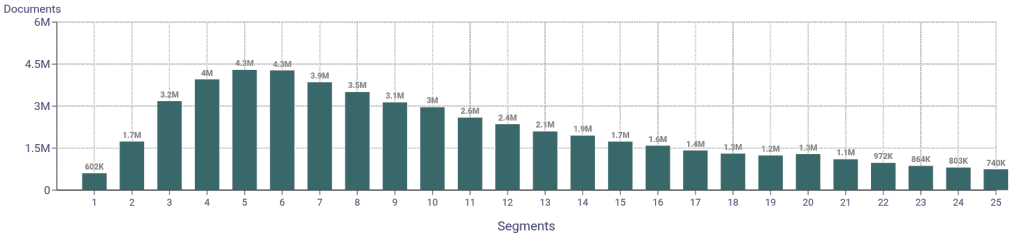




MT:2.0% | 1.4M Documents

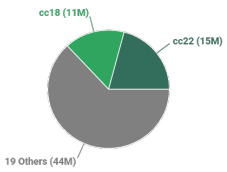
Documents size (in segments)

<= 25 segments **100%** (53M documents)
> 25 segments **0%** (0 documents)



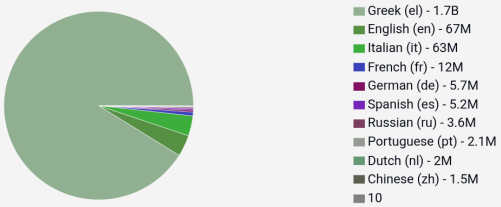
Document collections

CC = 54.93%
IA = 45.07%

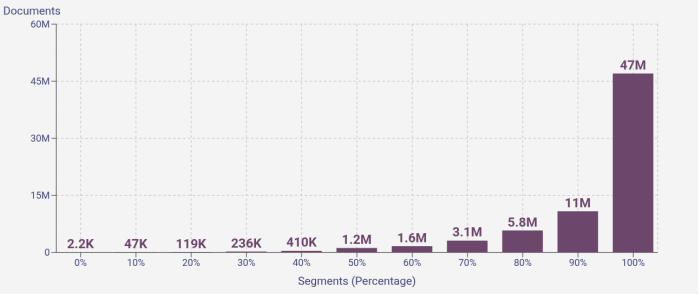


Language Distribution

Number of segments in the Greek (el) corpus

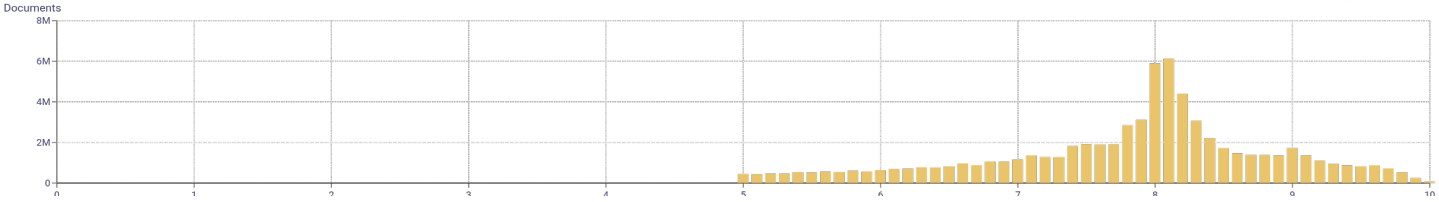


Percentage of segments in Greek (el) inside documents



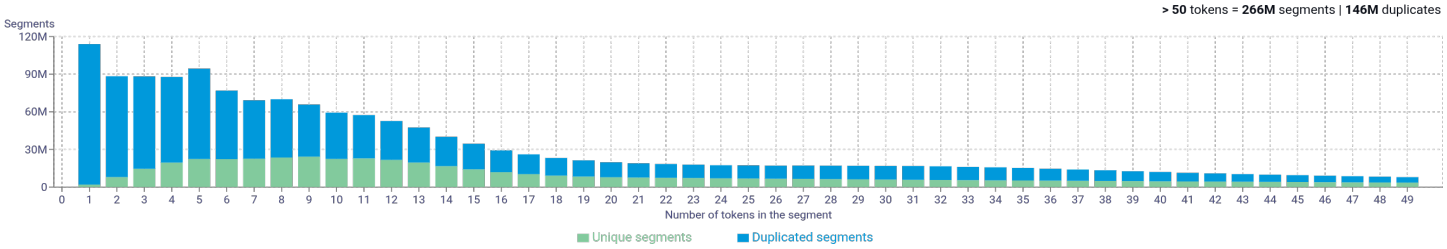
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (70M documents)

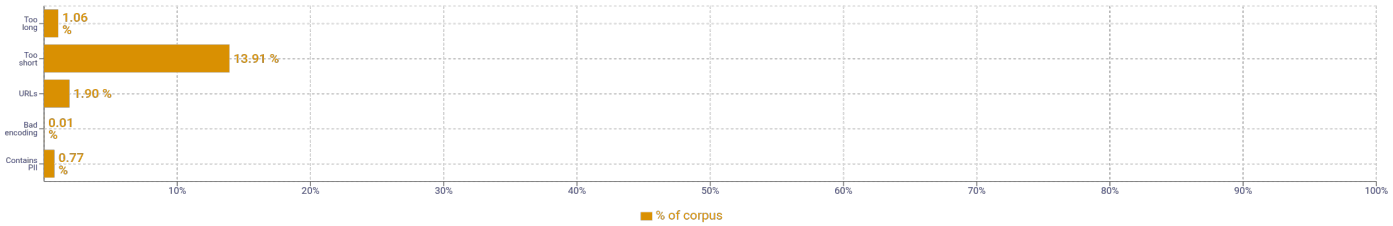


Segment length distribution by token

≤ 49 tokens = **1.6B** segments | **1.1B** duplicates
> 50 tokens = **266M** segments | **146M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	της | 709791617 · από | 500424729 · είναι | 349978341 · τους | 269685958 · τη | 237242716
2	από τη | 28406427 · από τους | 23028260 · πριν από | 15039016 · της χώρας | 10238213 · διαβάστε περισσότερα | 9908523
3	από την άλλη | 3109555 · ένας από τους | 2097234 · από την πλευρά | 2004434 · αυτή τη στιγμή | 1772930 · από την αρχή | 1704974
4	ένα από τα πιο | 937480 · άδειες χρήσης το παρόν | 474848 · αξίζει να σημειωθεί ότι | 452086 · αναρτητεα στο διαδικτυο ελληνικη | 443930 · α π ο σ | 422801
5	α π ο σ π | 418673 · άδειες χρήσης το παρόν εκπαιδευτικό | 399666 · α π ο φ α | 234451 · α σ μ α από | 204426 · google news και μάθετε πρώτοι | 185720

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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av