HPLT Analytics report

General overview

Corpus	Analytics date	Language
HPLT-v2-ell_Grek.tsv	10/1/2024	Greek (el)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
70,328,890	1,849,478,014			468.4 GB	281,755,294,923

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
tripadvisor.com.gr	212K	0.30

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
nl	220K	0.31

Documents size (in segments)

<= 25 segments **76.07%** (53M documents)
> 25 segments **23.93%** (17M documents)
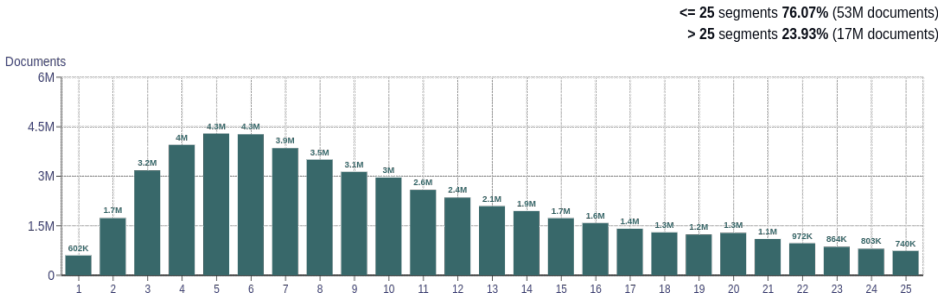
Documents by collection

cc18 (11M)
cc22 (15M)
19 Others (44M)
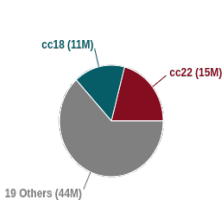
Language Distribution

Number of segments

- Greek (el) - 1.7B
- English (en) - 69M
- Italian (it) - 63M
- French (fr) - 12M
- German (de) - 5.7M
- Spanish (es) - 5.2M
- Russian (ru) - 3.6M
- Portuguese (pt) - 2.1M
- Dutch (nl) - 2M
- Chinese (zh) - 1.5M
- 165 Others - 16M
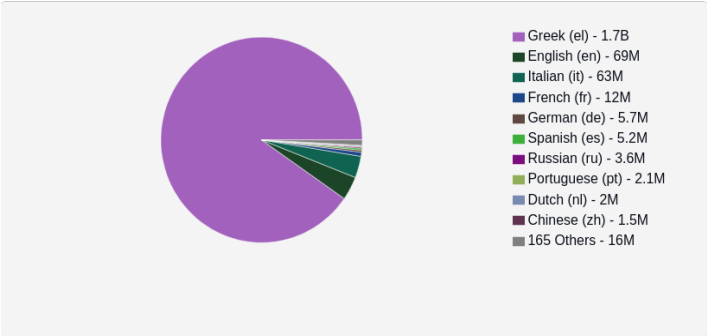
Percentage of segments in Greek (el) inside documents
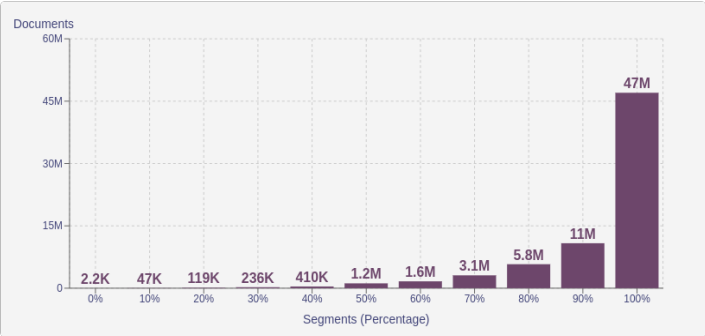
Segments (Percentage)	Documents
0%	2.2K
10%	47K
20%	119K
30%	236K
40%	410K
50%	1.2M
60%	1.6M
70%	3.1M
80%	5.8M
90%	11M
100%	47M

Distribution of documents by document score

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

Segment noise distribution

- Too long: 0.00 %
- Too short: 5.75 %
- URLs: 1.84 %
- Bad encoding: 0.01 %
- Contains PII: 0.77 %
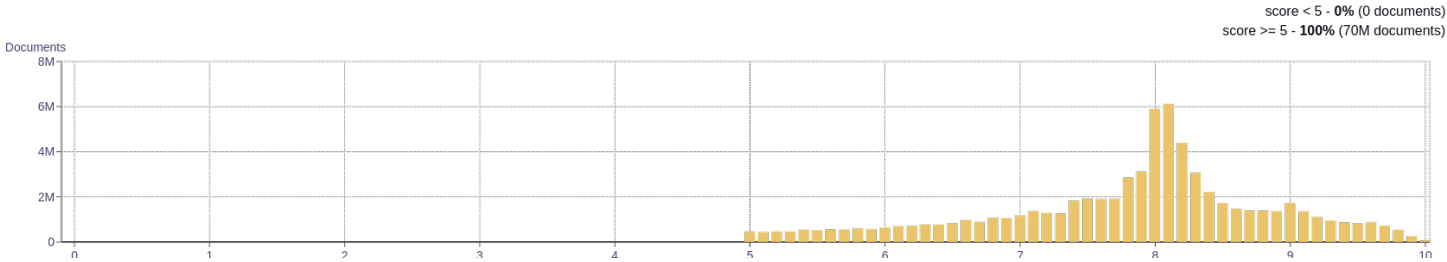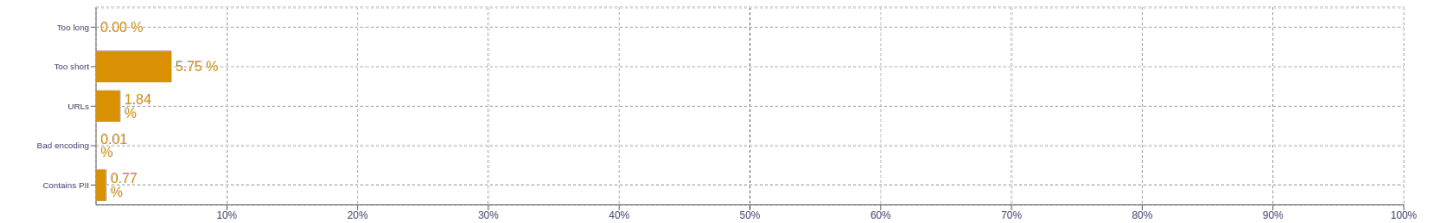
About HPLT Analytics

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