HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-lit_Latn	9/18/2025	Lithuanian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
20,406,756	510,993,236	273,216,536 (53.47 %)	13B	80,228,908,679	79.39 GB

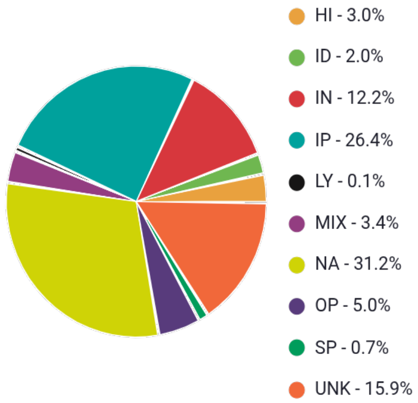
Top 10 domains

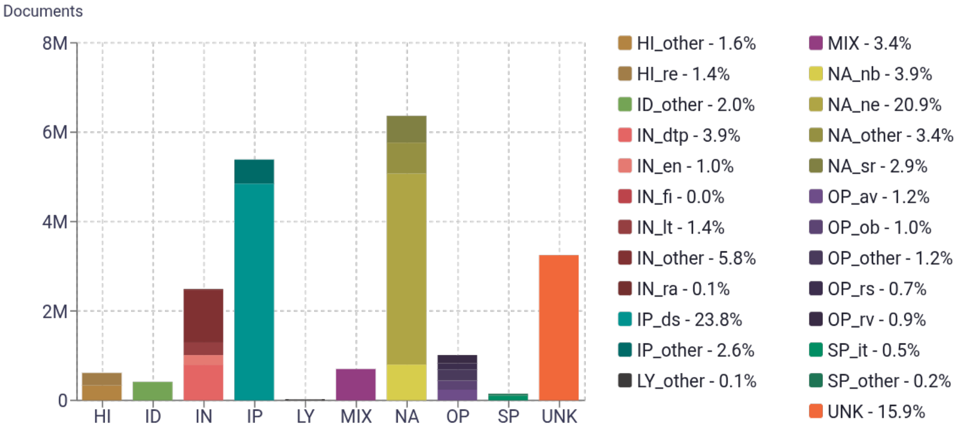
Domain	Docs	% of total
delfi.lt	674K	3.30%
15min.lt	361K	1.77%
diena.lt	227K	1.11%
lrt.lt	225K	1.10%
tv3.lt	191K	0.93%
lzinios.lt	163K	0.80%
lrytas.lt	155K	0.76%
blogspot.com	126K	0.62%
wikipedia.org	119K	0.59%
vz.lt	108K	0.53%

Top 10 TLDs

Domain	Docs	% of total
lt	16M	78.51%
com	2.4M	11.85%
eu	473K	2.32%
org	358K	1.75%
net	240K	1.18%
info	161K	0.79%
news	57K	0.28%
pt	42K	0.21%
today	40K	0.20%
ru	38K	0.18%

Register labels

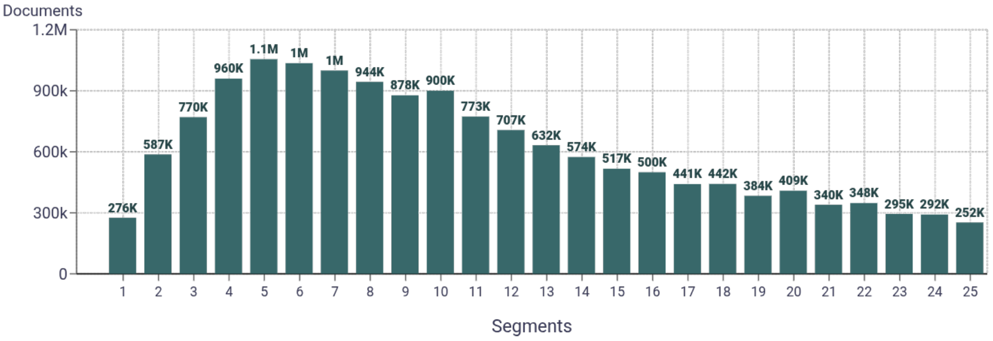




MT:13.0% | 2.6M Documents

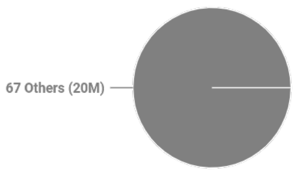
Documents size (in segments)

≤ 25 segments **75.03%** (15M documents)
> 25 segments **24.97%** (5.1M documents)



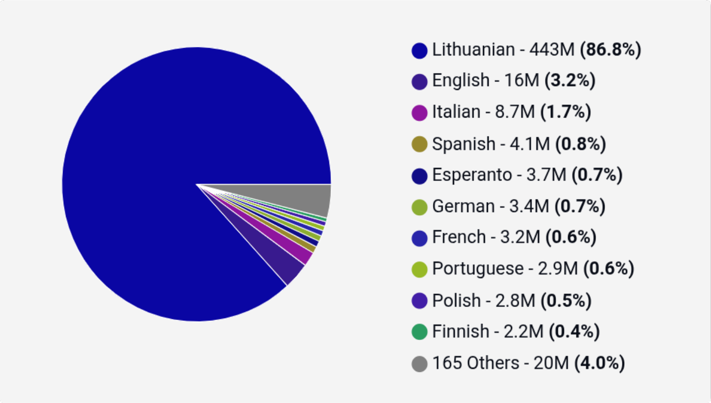
Document collections

CC = 89.58%
IA = 10.42%



Language Distribution

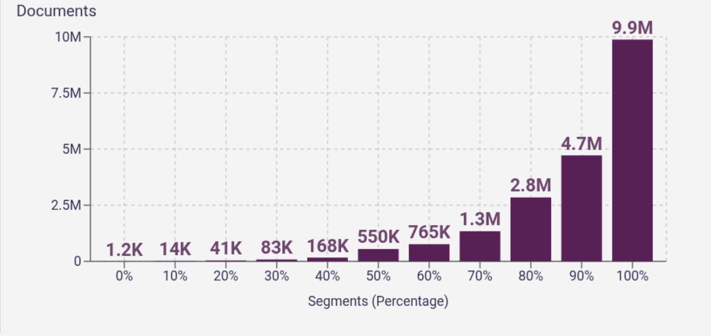
Number of segments in the Lithuanian corpus



Percentage of segments in Lithuanian inside documents

segments < 50% - **1.50%** (307K documents)

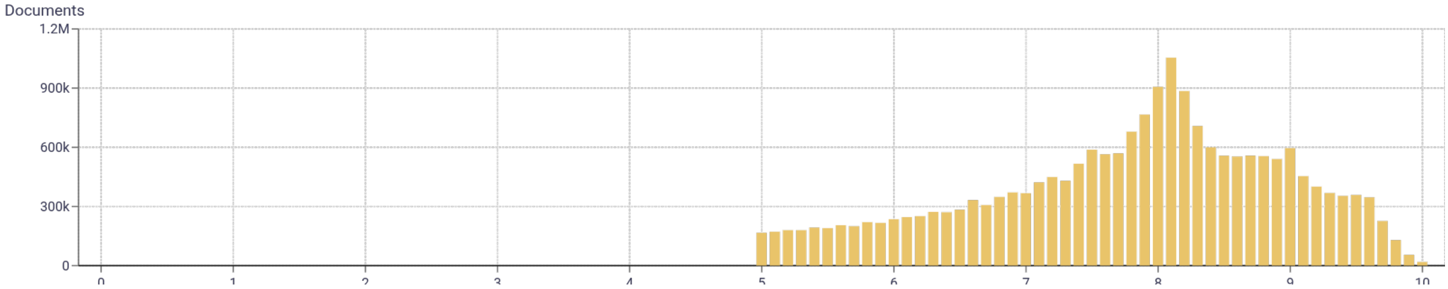
segments ≥ 50% - **98.50%** (20M documents)



Distribution of documents by document score

score < 5 - **0%** (0 documents)

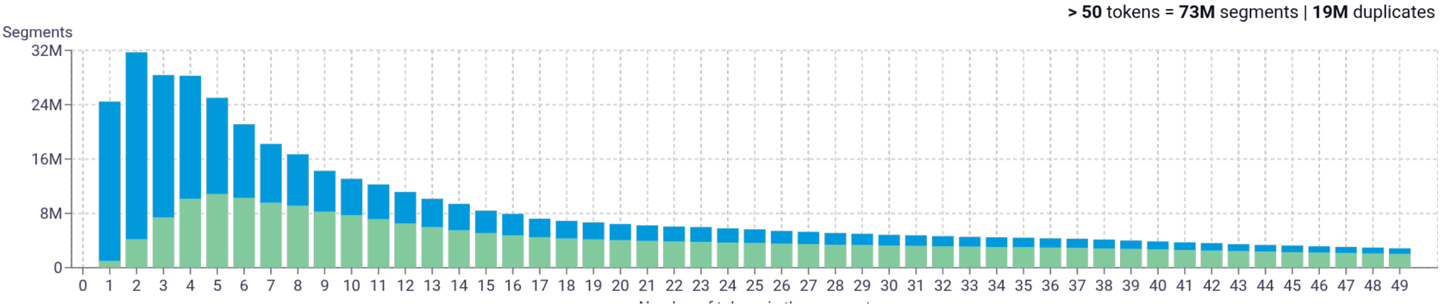
score ≥ 5 - **100%** (20M documents)



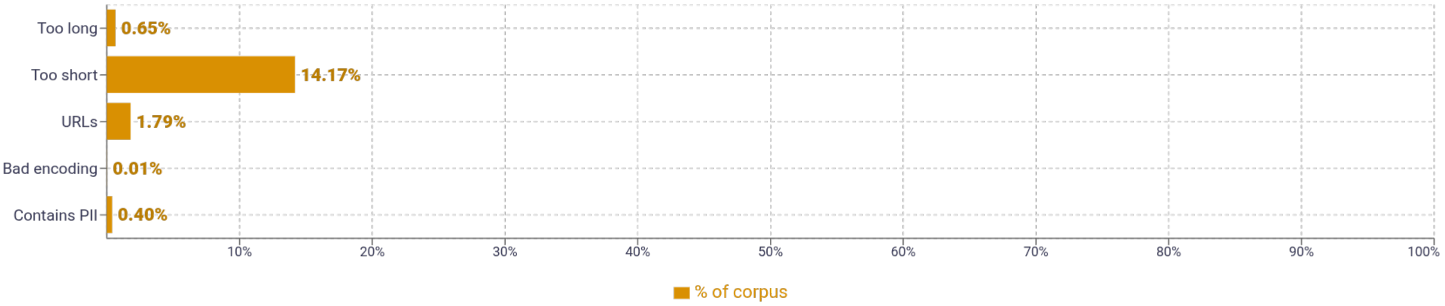
Segment length distribution by token

≤ 49 tokens = **438M** segments | **220M** duplicates

> 50 tokens = **73M** segments | **19M** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	yra | 102,855,558 gali | 38,199,783 buvo | 31,351,329 būti | 24,889,289 m. | 19,934,910
2	gali būti | 11,778,151 turi būti | 4,289,850 šiuo metu | 3,319,473 šiek tiek | 3,248,173 yra labai | 2,902,737
3	kartus per dieną | 962,830 tuo pačiu metu | 795,972 širdies ir kraujagyslių | 510,760 akcijų pasirinkimo sandoriai | 504,755 gali būti naudojamas | 407,102
4	brand new ir aukštos | 208,180 new ir aukštos kokybės | 199,284 socialinės apsaugos ir darbo | 177,900 tris kartus per dieną | 167,488 teismo civilinių bylų skyriaus | 153,339
5	brand new ir aukštos kokybės | 199,151 teismo civilinių bylų skyriaus teisėjų | 115,376 d. nutartis civilinėje byloje nr. | 84,372 aukščiausiojo teismo civilinių bylų skyriaus | 79,477 visuomenės informavimo priemonėse bei interneto | 79,344

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.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	**How-to or instructions**	**HI**	Description of a thing or person	dtp
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				