HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
lit_Latn.jsonl.tsv	6/10/2025	Lithuanian (lt)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
13,338,275	322,101,556	137,631,836 (42.73 %)	8.1B	50,084,920,486	49.59 GB

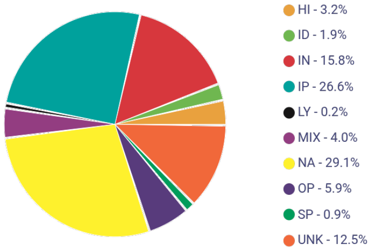
Top 10 domains

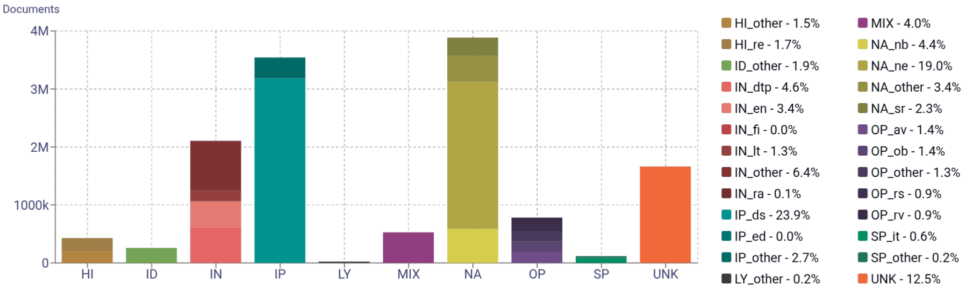
Domain	Docs	% of total
wikipedia.org	401K	3.00%
delfi.lt	253K	1.90%
15min.lt	241K	1.81%
lzinios.lt	231K	1.73%
diena.lt	207K	1.55%
mokslobaze.lt	142K	1.06%
lrt.lt	128K	0.96%
alkas.lt	125K	0.93%
blogspot.com	116K	0.87%
hotels.com	91K	0.68%

Top 10 TLDs

Domain	Docs	% of total
lt	11M	79.01%
com	1.4M	10.44%
org	534K	4.01%
eu	278K	2.09%
net	165K	1.24%
info	94K	0.70%
ru	22K	0.17%
co.uk	21K	0.16%
today	16K	0.12%
pl	14K	0.11%

Register labels

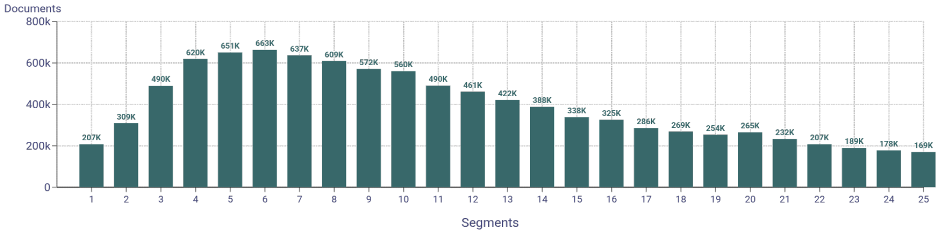




🤖 **MT**:9.4% | 1.3M Documents

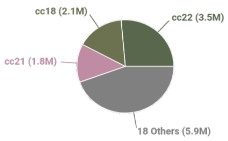
Documents size (in segments)

<= 25 segments **73.4%** (9.8M documents)
> 25 segments **26.6%** (3.5M documents)



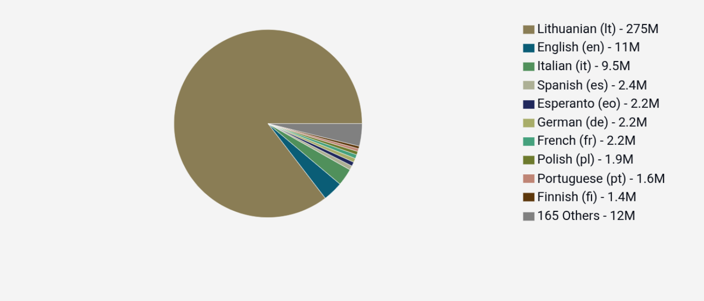
Documents by collection

CC = 66.62%
IA = 33.38%

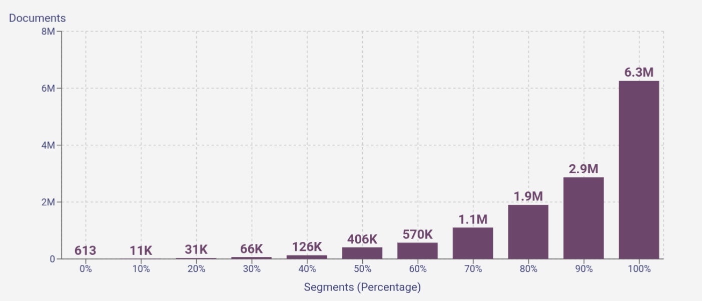


Language Distribution

Number of segments in the Lithuanian (lt) corpus

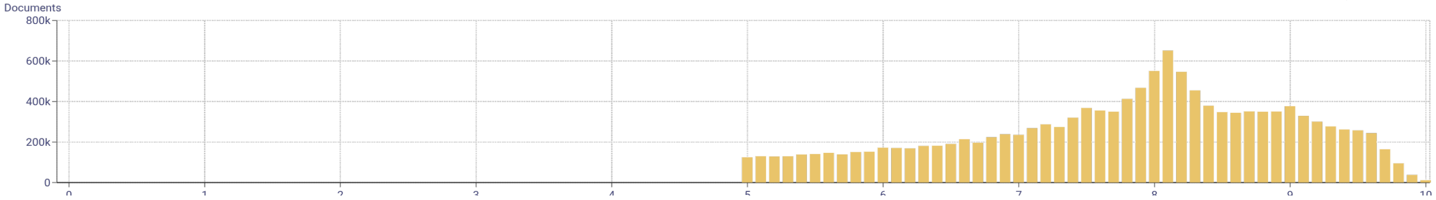


Percentage of segments in Lithuanian (lt) inside documents



Distribution of documents by document score

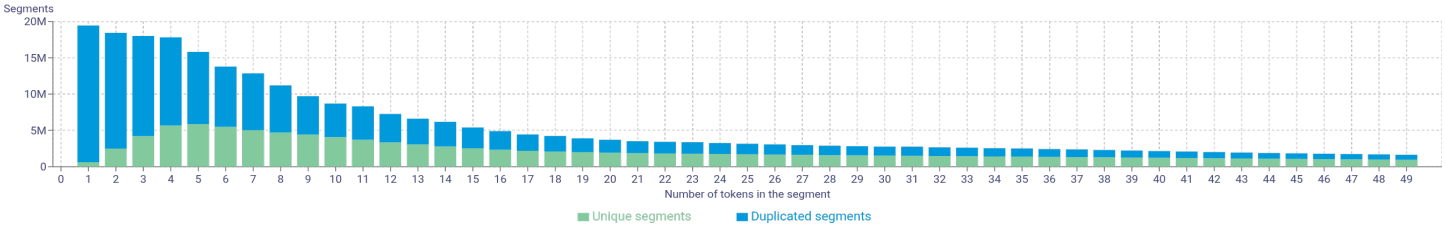
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (13M documents)



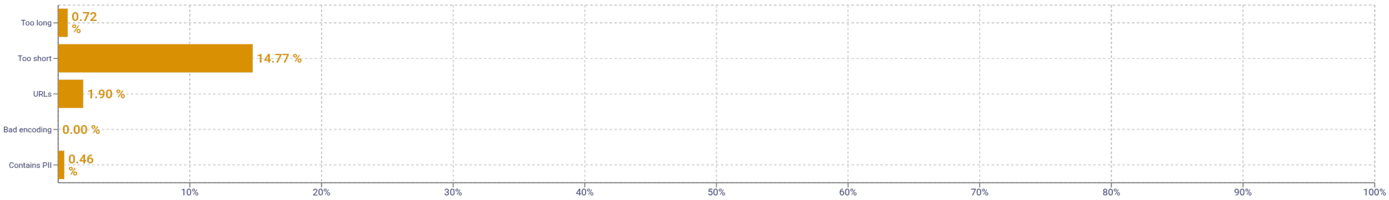
Segment length distribution by token

≤ 49 tokens = **109M** segments | **164M** duplicates

> 50 tokens = **49M** segments | **20M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	yra | 52706592 buvo | 21554146 gali | 19860198 m. | 14828048 lietuvos | 13546087
2	gali būti | 5800176 turi būti | 2411956 šiuo metu | 2050120 šiek tiek | 1681239 lietuvos respublikos | 1633727
3	kartus per dieną | 329067 tuo pačiu metu | 311230 akcijų pasirinkimo sandoriai | 280489 širdies ir kraujagyslių | 267274 automobilių stovėjimo aikštelė | 262772
4	šį viešbutį per paskutiniąją | 251796 viešbutį per paskutiniąją valandą | 251796 vilniaus visuomenės sveikatos biuras | 160519 socialinės apsaugos ir darbo | 125022 visuomenės sveikatos biuras kviečia | 106680
5	šį viešbutį per paskutiniąją valandą | 251796 peržiūrėjo šį viešbutį per paskutiniąją | 251796 visuomenės informavimo priemonėse bei interneto | 76992 informavimo priemonėse bei interneto tinklalapiuose | 76906 visuomenės sveikatos biuras vykdo programą | 75924

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