HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-lt.tsv	1/28/2025	English (en)	Lithuanian (lt)

Volumes

Segments	SL tokens	SL characters	SL size
12,881,354	292M	1,532,375,361	1.43 GB

TL tokens	TL characters	TL size
249M	1,555,792,770	1.54 GB

Dataset top 10 domains

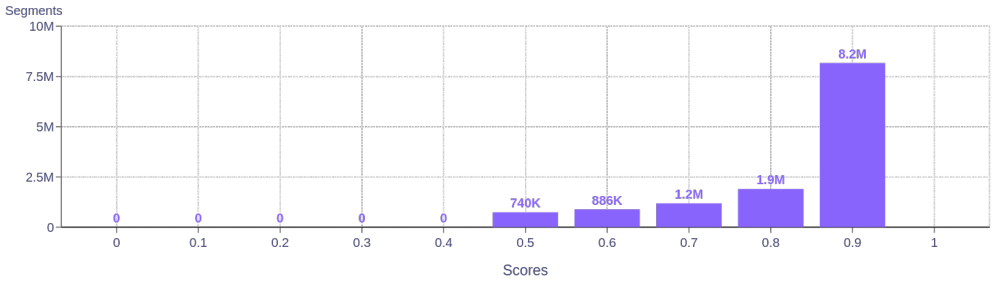
SL domain	Segments	TL domain	Segments
hotels.com	39.9%	hotels.com	15.4%
europa.eu	13.4%	europa.eu	10.4%
google.com	7.9%	agoda.com	4.7%
agoda.com	6.7%	google.com	3.8%
booking.com	5.3%	booking.com	2.9%
wikipedia.org	2.8%	wikipedia.org	2.4%
microsoft.com	2.2%	microsoft.com	1.4%
office.com	1.5%	office.com	1.3%
softoware.net	1.3%	studybible.info	1.1%
europages.co.uk	1.2%	europages.lt	1.1%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	123.7%	com	64.9%
eu	17.4%	lt	35.4%
lt	11.3%	eu	14.0%
org	9.9%	org	6.9%
net	4.8%	net	3.5%
co.uk	4.0%	info	2.5%
info	2.7%	com.br	0.5%
de	1.3%	de	0.4%
ie	1.1%	co.uk	0.4%
ca	0.8%	ru	0.3%

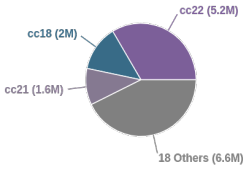
Translation likelihood

≥ 5 = 13M segments | **100.0%**
≥ 8 = 10M segments | **78.2%**
< 5 = 0 segments | **0.0%**



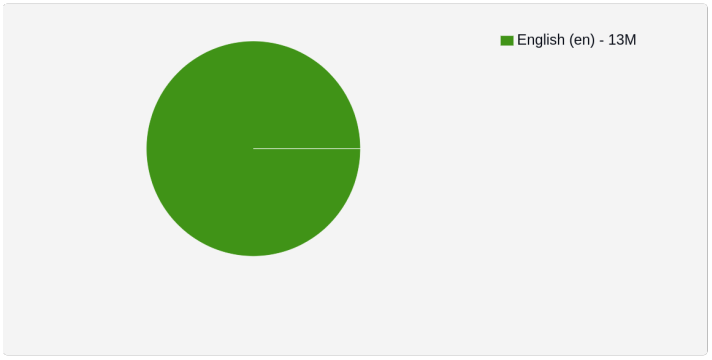
Collections

CC = 67.44%
IA = 32.56%

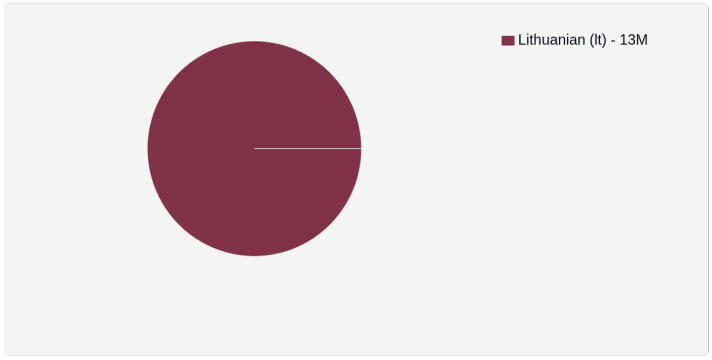


Language Distribution

Source

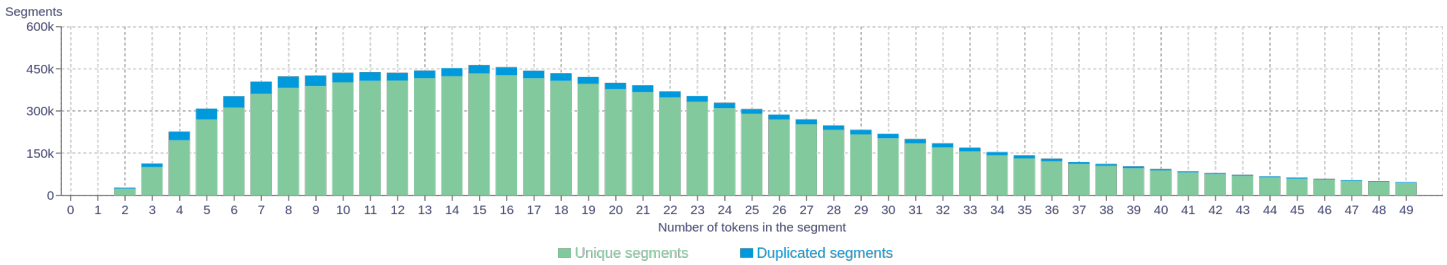


Target



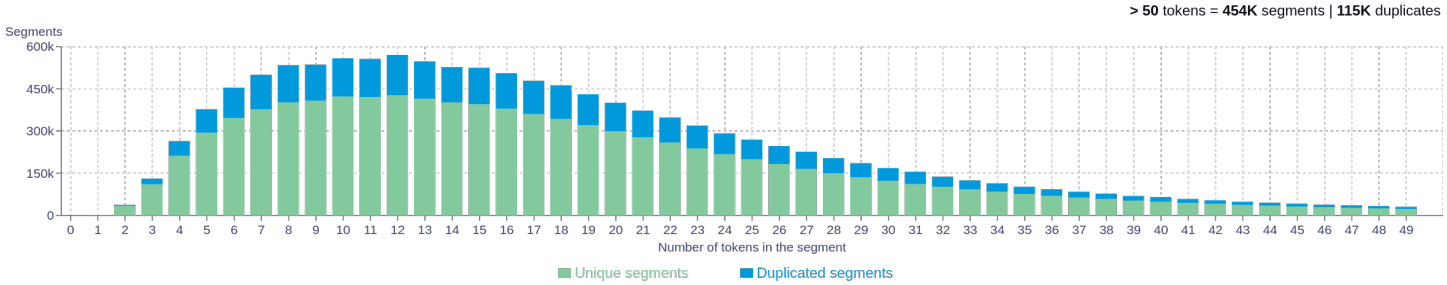
Source segment length distribution by token

<= 49 tokens = **11M** segments | **867K** duplicates
> 50 tokens = **782K** segments | **35K** duplicates

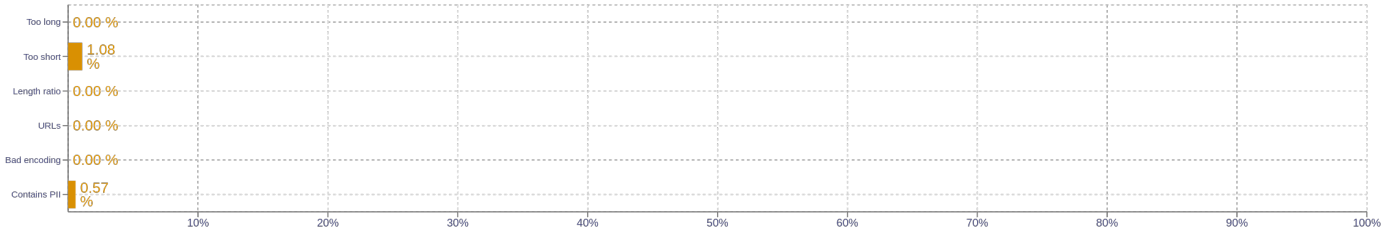


Target segment length distribution by token

<= 49 tokens = **9.4M** segments | **3.1M** duplicates
> 50 tokens = **454K** segments | **115K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	data | 1124855 · personal | 710061 · use | 592405 · information | 580026 · also | 522316
2	personal data | 567035 · privacy policy | 93659 · personal information | 91295 · member states | 82845 · data protection | 81592
3	processing of personal | 62073 · wi-fi in public | 42406 · non smoking rooms | 41570 · terms and conditions | 39089 · protected from spambots | 33551
4	processing of personal data | 61389 · processing of your personal | 51718 · wi-fi in public areas | 42388 · wi-fi in all rooms | 38696 · address is being protected | 33577
5	processing of your personal data | 47540 · free wi-fi in all rooms | 38682 · email address is being protected | 31237 · parliament and of the council | 27037 · people looked at this hotel | 22133

Target n-grams

Size	n-grams
1	yra | 2335340 · į | 1407631 · iš | 1143547 · jūsų | 1060417 · gali | 795428
2	gali būti | 270025 · asmens duomenų | 239844 · asmens duomenis | 230036 · jūsų asmens | 212560 · asmens duomenys | 126046
3	jūsų asmens duomenis | 105519 · jūsų asmens duomenų | 55565 · automobilių stovėjimo aikštelė | 54747 · asmens duomenų tvarkymo | 47801 · jūs turite teisę | 43404
4	nemokamas wifi visuose kambariuose | 38656 · europos parlamento ir tarybos | 30219 · km atstumu nuo šių | 29309 · atstumu nuo šių objektų | 29305 · el.pašto adresas yra apsaugotas | 24270
5	km atstumu nuo šių objektų | 29305 · šis el.pašto adresas yra apsaugotas | 23558 · šį viešbutį per paskutiniąją valandą | 21088 · peržiūrėjo šį viešbutį per paskutiniąją | 21088 · adresas yra apsaugotas nuo šiukšlių | 19137

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