HPLT Analytics report

General overview

Corpus	Date	SL	TL
hplt-v2-en-lv.tsv	1/27/2025	English (en)	Latvian (lv)

Volumes

Segments	SL tokens	SL characters	SL size
11,294,618	255M	1,339,163,933	1.25 GB

TL tokens	TL characters	TL size
221M	1,338,919,670	1.36 GB

Dataset top 10 domains

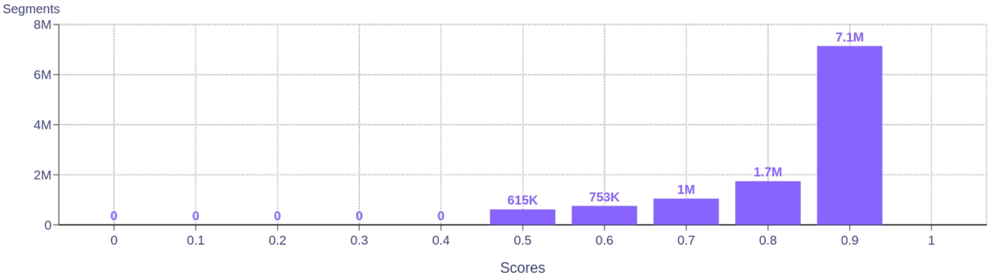
SL domain	Segments	TL domain	Segments
hotels.com	40.9%	hotels.com	15.9%
europa.eu	16.2%	europa.eu	12.7%
google.com	9.7%	agoda.com	4.9%
booking.com	6.9%	google.com	4.5%
agoda.com	6.7%	booking.com	3.4%
microsoft.com	2.8%	wikipedia.org	2.2%
wikipedia.org	2.4%	microsoft.com	1.7%
office.com	1.8%	office.com	1.6%
airwise.com	1.2%	likumi.lv	1.0%
orangesmile.com	1.0%	coolmom.info	0.8%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	128.1%	com	70.7%
eu	19.9%	lv	28.7%
lv	11.3%	eu	16.0%
org	8.7%	org	6.5%
net	3.8%	net	2.8%
co.uk	2.7%	info	2.3%
info	2.6%	gov.lv	1.0%
de	1.2%	lt	0.5%
ie	1.0%	ru	0.5%
gov.lv	0.9%	ee	0.4%

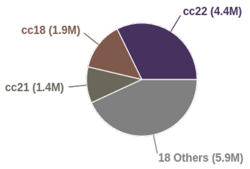
Translation likelihood

≥ 5 = 11M segments | **100.0%**
≥ 8 = 8.9M segments | **78.6%**
< 5 = 0 segments | **0.0%**



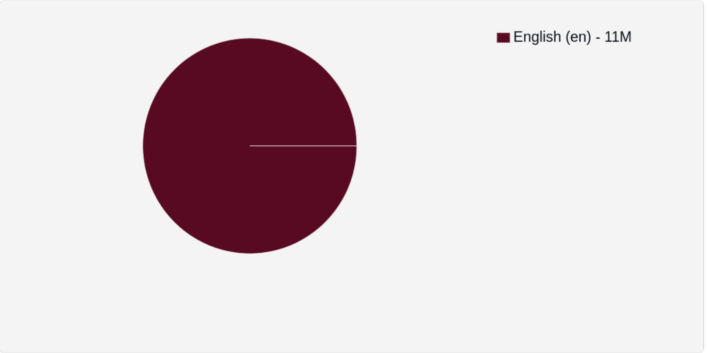
Collections

CC = 67.58%
IA = 32.42%

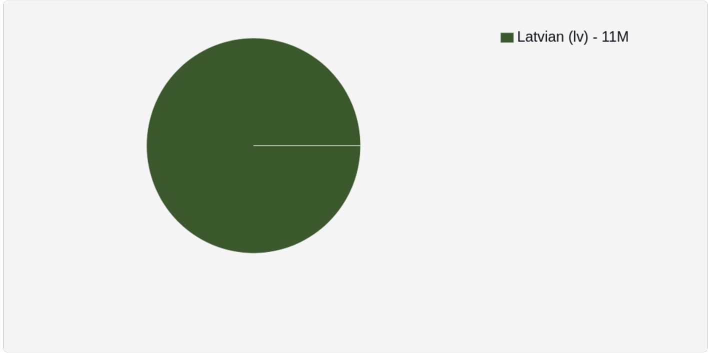


Language Distribution

Source

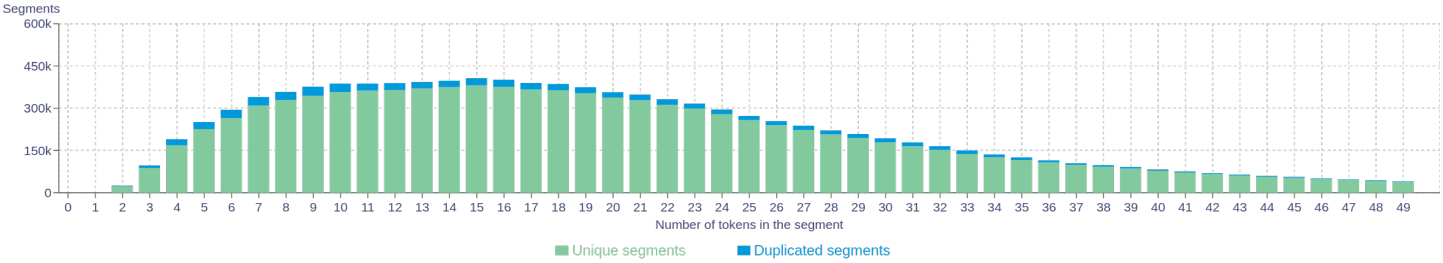


Target



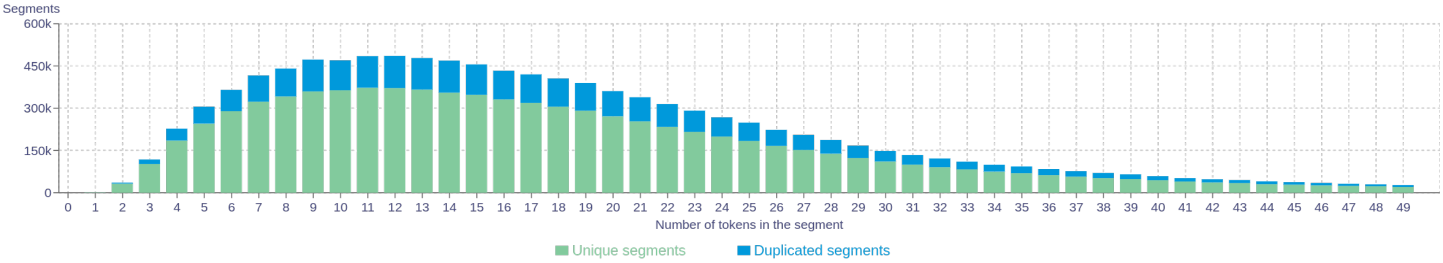
Source segment length distribution by token

<= 49 tokens = **9.9M** segments | **702K** duplicates
> 50 tokens = **665K** segments | **28K** duplicates

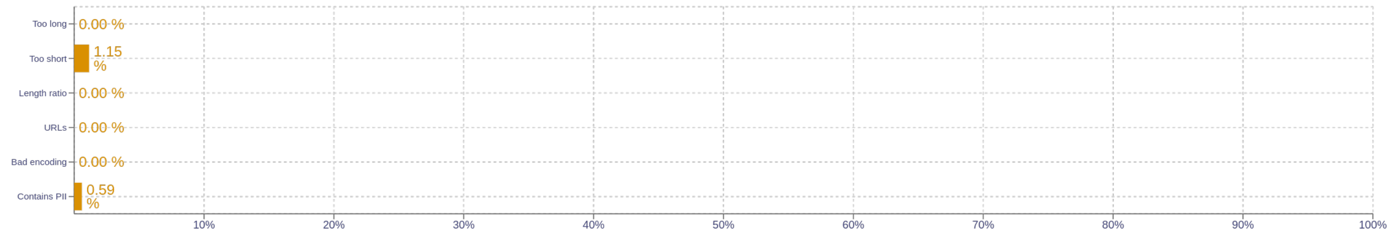


Target segment length distribution by token

<= 49 tokens = **8.3M** segments | **2.6M** duplicates
> 50 tokens = **407K** segments | **103K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	data | 817777 use | 499163 personal | 489837 information | 480818 also | 456057
2	personal data | 384024 member states | 88255 public areas | 71878 personal information | 63787 privacy policy | 61502
3	processing of personal | 49875 wi-fi in public | 46291 non smoking rooms | 39596 terms and conditions | 31853 24-hour front desk | 26569
4	processing of personal data | 49400 wi-fi in public areas | 46276 wi-fi in all rooms | 42348 processing of your personal | 33286 address is being protected | 25796
5	free wi-fi in all rooms | 42332 parliament and of the council | 30920 processing of your personal data | 30720 email address is being protected | 24377 people looked at this hotel | 15678

Target n-grams

Size	n-grams
1	kas | 1589926 jūsu | 793388 to | 771683 jūs | 748841 jums | 645928
2	personas datu | 202714 personas datus | 128843 jūsu personas | 128039 datu apstrādi | 70689 kas atrodas | 63828
3	jūsu personas datus | 58482 personas datu apstrādi | 53659 jums ir tiesības | 50808 bezmaksas wi-fi visos | 41956 wi-fi visos numuros | 41955
4	bezmaksas wi-fi visos numuros | 41955 eiropas parlamenta un padomes | 33524 papildu ziņas par naktsmītni | 28025 e-pasta adrese ir aizsargāta | 23829 aizsargāta no mēstuļu robotiem | 23007
5	šī e-pasta adrese ir aizsargāta | 23485 adrese ir aizsargāta no mēstuļu | 23007 km attālumā no apskates vietas | 20865 pēdējās stundas laikā šo viesnīcu | 15170 laikā šo viesnīcu ir skatījušas | 15170

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