HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-et.tsv	1/26/2025	English (en)	Estonian (et)

Volumes

Segments	SL tokens	SL characters	SL size
8,797,574	207M	1,090,375,785	1.02 GB

TL tokens	TL characters	TL size
158M	1,043,287,714	1.0 GB

Dataset top 10 domains

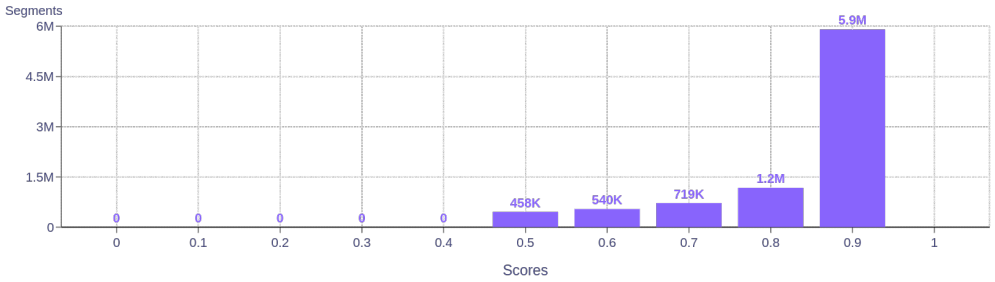
SL domain	Segments	TL domain	Segments
europa.eu	15.5%	europa.eu	12.5%
hotels.com	12.5%	hotels.com	6.0%
google.com	6.7%	wikipedia.org	3.8%
agoda.com	4.9%	agoda.com	3.6%
wikipedia.org	4.5%	google.com	3.2%
booking.com	3.4%	riigiteataja.ee	3.0%
microsoft.com	3.2%	microsoft.com	2.1%
riigiteataja.ee	2.9%	booking.com	2.1%
office.com	2.2%	office.com	2.0%
err.ee	1.0%	err.ee	1.1%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	87.4%	com	56.0%
ee	21.9%	ee	34.6%
eu	19.7%	eu	16.0%
org	10.8%	org	9.0%
net	3.5%	net	2.9%
info	2.4%	info	2.3%
co.uk	2.4%	fi	0.9%
de	1.2%	de	0.5%
fi	1.1%	ru	0.4%
ie	1.0%	lv	0.4%

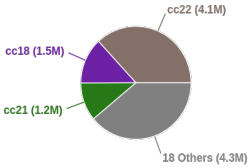
Translation likelihood

≥ 5 = 8.8M segments | **100.0%**
≥ 8 = 7.1M segments | **80.5%**
< 5 = 0 segments | **0.0%**



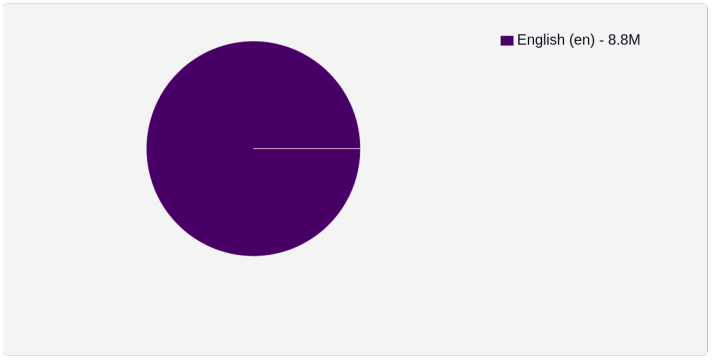
Collections

CC = 71.47%
IA = 28.53%

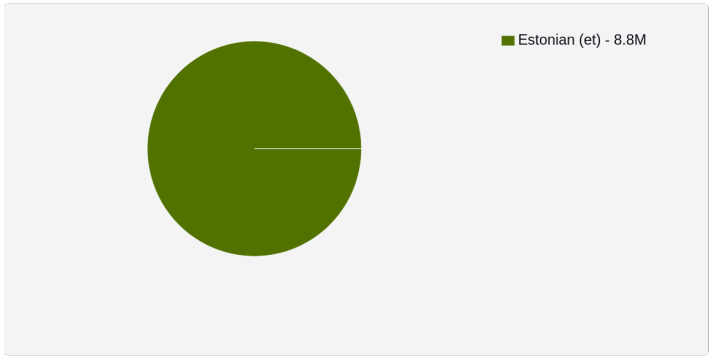


Language Distribution

Source

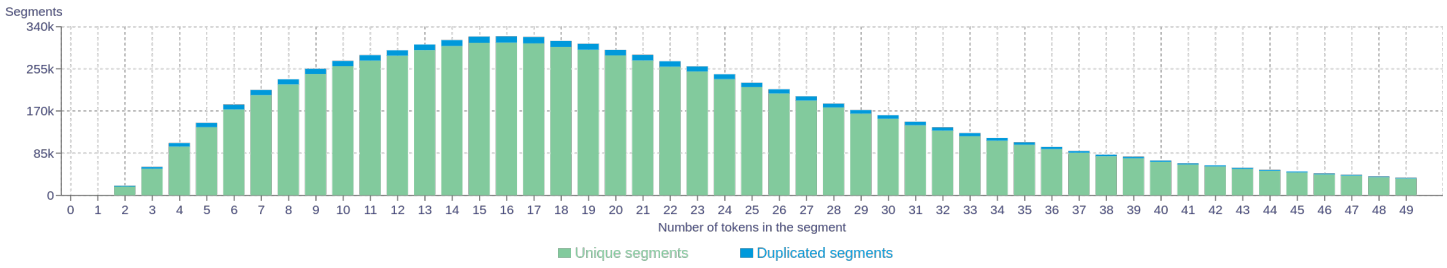


Target



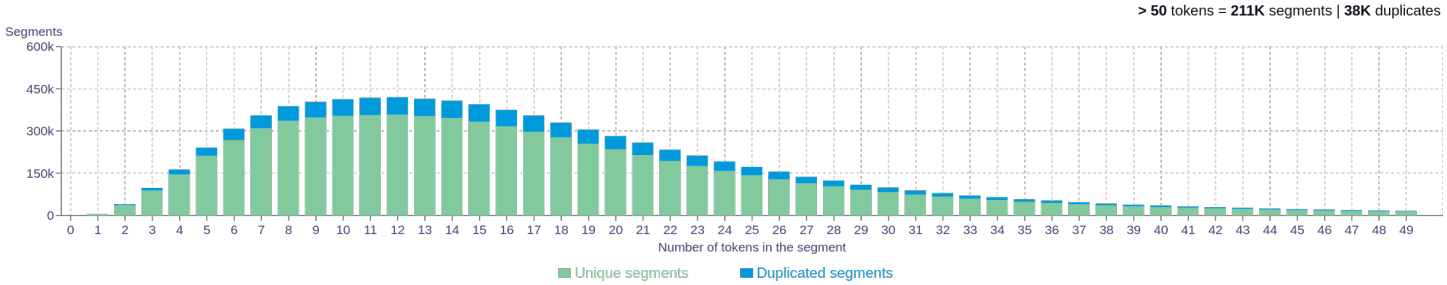
Source segment length distribution by token

<= 49 tokens = **7.9M** segments | **344K** duplicates
> 50 tokens = **547K** segments | **21K** duplicates

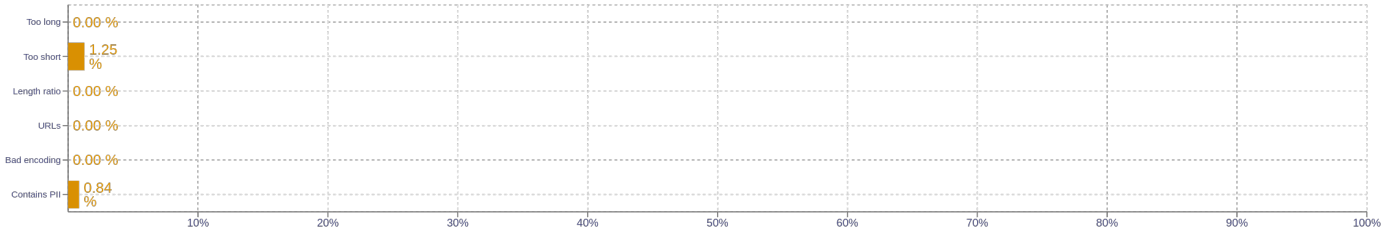


Target segment length distribution by token

<= 49 tokens = **7.3M** segments | **1.3M** duplicates
> 50 tokens = **211K** segments | **38K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	data | 500981 · also | 401557 · use | 362407 · one | 327699 · information | 323378
2	personal data | 209383 · member states | 65018 · european union | 45327 · member state | 44082 · data protection | 35424
3	processing of personal | 23251 · terms and conditions | 16851 · wi-fi in public | 15512 · choice for travelers | 11954 · right to object | 10908
4	processing of personal data | 23037 · processing of your personal | 17271 · wi-fi in all rooms | 16999 · wi-fi in public areas | 15501 · referred to in article | 13836
5	parliament and of the council | 19222 · free wi-fi in all rooms | 16988 · processing of your personal data | 16146 · people looked at this hotel | 7311 · hotel in the last hour | 7311

Target n-grams

Size	n-grams
1	või | 1169543 · ning | 741931 · ka | 567920 · teie | 432300 · meie | 368748
2	euroopa liidu | 45900 · võib olla | 40923 · teie isikuandmeid | 35941 · igal ajal | 32263 · ettevõttega ühendust | 31607
3	parlamendi ja nõukogu | 20766 · teil on õigus | 17484 · wifi-ühendus kõigis tubades | 17204 · tasuta wifi-ühendus kõigis | 17204 · iganes on sinu | 10752
4	euroopa parlamendi ja nõukogu | 20639 · tasuta wifi-ühendus kõigis tubades | 17204 · kuidas sinna kohale jõuda | 8168 · vaadata vabade kohtade olemasolu | 7717 · kohe vaadata vabade kohtade | 7717
5	kohe vaadata vabade kohtade olemasolu | 7717 · hindu ja kohe vaadata vabade | 7400 · vaadanud seda hotelli viimase tunni | 6623 · külastajat on vaadanud seda hotelli | 6623 · hotels.com abiga on lihtne leida | 6527

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