HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
10,336,528	244M	1,285,859,216	1.2 GB

TL tokens	TL characters	TL size
222M	1,253,947,936	1.2 GB

Dataset top 10 domains

SL domain	Segments	TL domain	Segments
europa.eu	16.3%	europa.eu	13.1%
google.com	7.7%	wikipedia.org	5.9%
wikipedia.org	7.2%	google.com	3.7%
agoda.com	3.9%	agoda.com	2.8%
office.com	2.5%	office.com	2.4%
booking.com	2.5%	booking.com	1.5%
microsoft.com	2.0%	gear4music.si	1.5%
jw.org	0.9%	microsoft.com	1.4%
coolmom.info	0.8%	jw.org	0.8%
gear4music.com	0.7%	coolmom.info	0.8%

Dataset top 10 TLDs

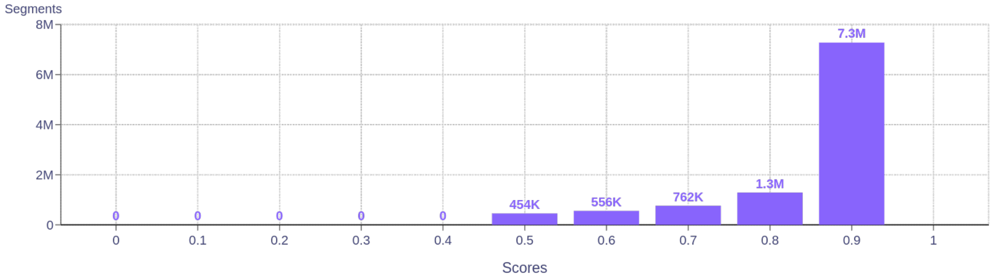
SL domain	Segments	TL domain	Segments
com	79.6%	com	52.2%
eu	21.4%	si	31.0%
org	15.4%	eu	17.0%
si	11.7%	org	12.0%
net	4.6%	net	3.7%
co.uk	3.0%	info	2.1%
de	2.3%	hr	0.9%
info	2.2%	de	0.9%
ie	1.6%	at	0.5%
hr	1.0%	ws	0.4%

Translation likelihood

≥ 5 = 10M segments | **100.0%**

≥ 8 = 8.6M segments | **82.9%**

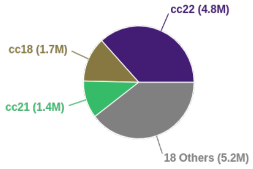
< 5 = 0 segments | **0.0%**



Collections

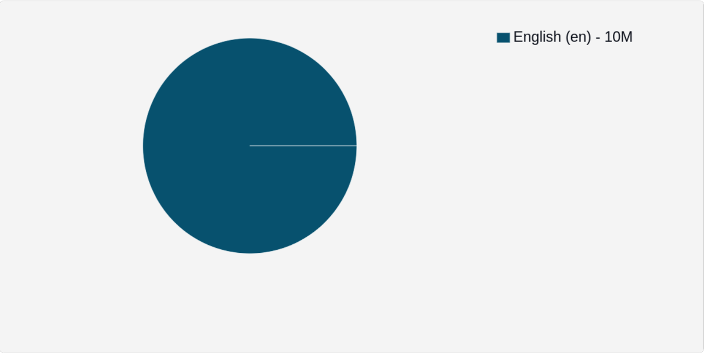
CC = 70.06%

IA = 29.94%

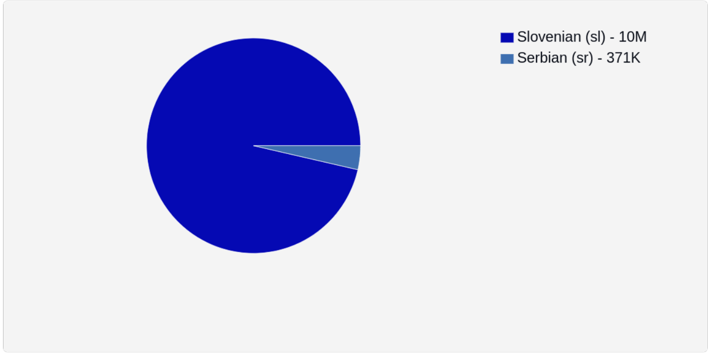


Language Distribution

Source



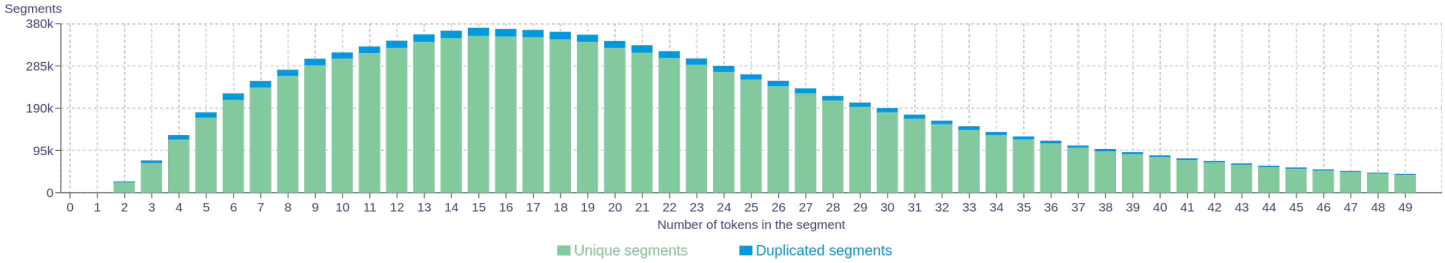
Target



Source segment length distribution by token

<= 49 tokens = **9.2M** segments | **478K** duplicates

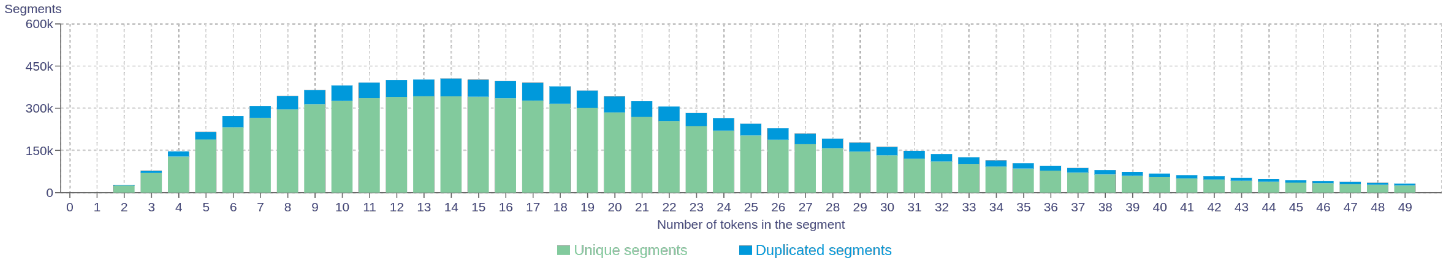
> 50 tokens = **643K** segments | **28K** duplicates



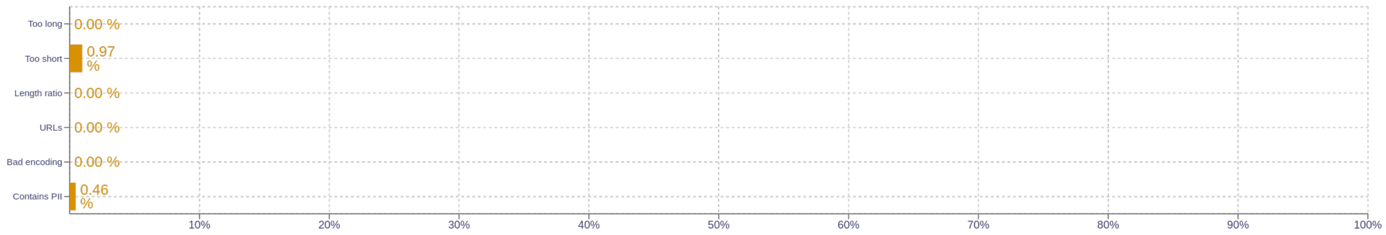
Target segment length distribution by token

<= 49 tokens = **8.3M** segments | **1.6M** duplicates

> 50 tokens = **484K** segments | **104K** duplicates



Segment pair noise distribution



Source n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>data | 789303
also | 510214
use | 479423
personal | 420456
information | 395624</td></tr>
<tr><td>2</td><td>personal data | 339249
member states | 86607
data protection | 60015
member state | 53802
european union | 48065</td></tr>
<tr><td>3</td><td>processing of personal | 43647
right to object | 21043
personal data concerning | 19494
terms and conditions | 16300
wi-fi in public | 16102</td></tr>
<tr><td>4</td><td>processing of personal data | 43303
processing of your personal | 28977
use of the website | 20053
referred to in article | 18534
wi-fi in all rooms | 16129</td></tr>
<tr><td>5</td><td>processing of your personal data | 27457
parliament and of the council | 25606
free wi-fi in all rooms | 16118
right to lodge a complaint | 7481
ec of the european parliament | 6935</td></tr>
</table>

Target n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>lahko | 1223037
podatkov | 514816
več | 416456
strani | 354017
osebnih | 256641</td></tr>
<tr><td>2</td><td>osebnih podatkov | 244755
osebne podatke | 98960
spletne strani | 97440
osebni podatki | 83651
spletnega mesta | 62598</td></tr>
<tr><td>3</td><td>nanašajo osebni podatki | 32998
parlamenta in sveta | 27277
skladu s členom | 27178
sveta z dne | 19126
obdelavi osebnih podatkov | 17563</td></tr>
<tr><td>4</td><td>evropskega parlamenta in sveta | 27164
wi-fi na javnih mestih | 15644
obdelavo vaših osebnih podatkov | 12031
osebne podatke v zvezi | 9450
osebnih podatkov v zvezi | 7221</td></tr>
<tr><td>5</td><td>parlamenta in sveta z dne | 16539
izvajanje ali obrambo pravnih zahtevkov | 6954
predel je odličen za popotnike | 6750
posameznikov pri obdelavi osebnih podatkov | 6474
varstvu posameznikov pri obdelavi osebnih | 6151</td></tr>
</table>

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