HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-is.tsv	1/22/2025	English (en)	Icelandic (is)

Volumes

Segments	SL tokens	SL characters	SL size
2,694,541	59M	295,187,731	282.92 MB

TL tokens	TL characters	TL size
54M	289,932,912	305.16 MB

Dataset top 10 domains

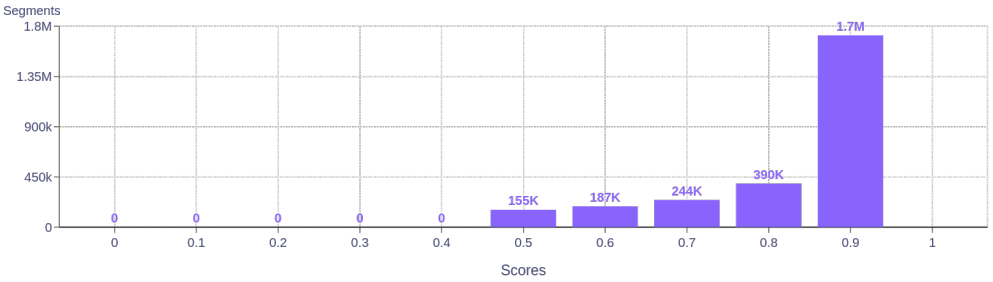
SL domain	Segments	TL domain	Segments
hotels.com	78.7%	hotels.com	36.7%
biblegateway.com	10.9%	biblegateway.com	9.9%
booking.com	7.0%	booking.com	4.8%
wikipedia.org	3.8%	wikipedia.org	3.7%
vsaduidoma.com	2.6%	vsaduidoma.com	2.6%
eso.org	2.6%	collectiveray.com	2.3%
europa.eu	2.1%	eso.org	1.9%
lds.org	2.1%	jw.org	1.8%
collectiveray.com	2.1%	lds.org	1.7%
jw.org	1.9%	orangesmile.com	1.4%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	160.8%	com	97.7%
org	20.5%	org	17.2%
is	9.0%	is	15.1%
net	5.2%	net	3.5%
eu	4.5%	eu	3.1%
nu	1.4%	nu	1.4%
co.uk	1.0%	de	0.8%
de	0.9%	info	0.7%
info	0.8%	top	0.6%
dk	0.6%	dk	0.6%

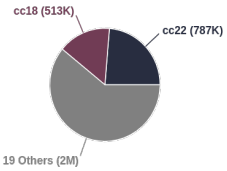
Translation likelihood

≥ 5 = 2.7M segments | **100.0%**
≥ 8 = 2.1M segments | **78.3%**
< 5 = 0 segments | **0.0%**



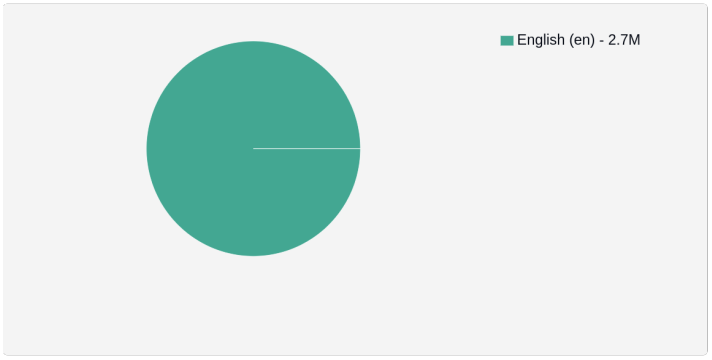
Collections

CC = 62.22%
IA = 37.78%

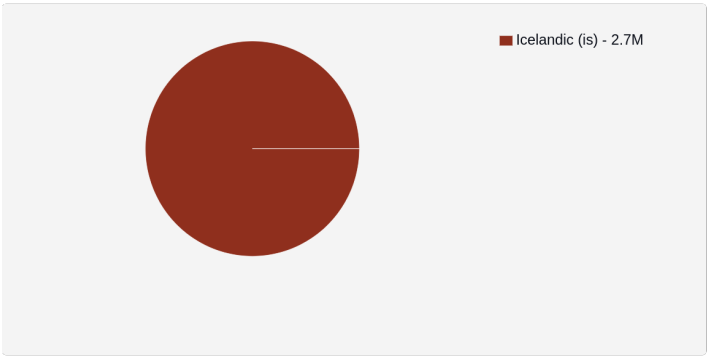


Language Distribution

Source

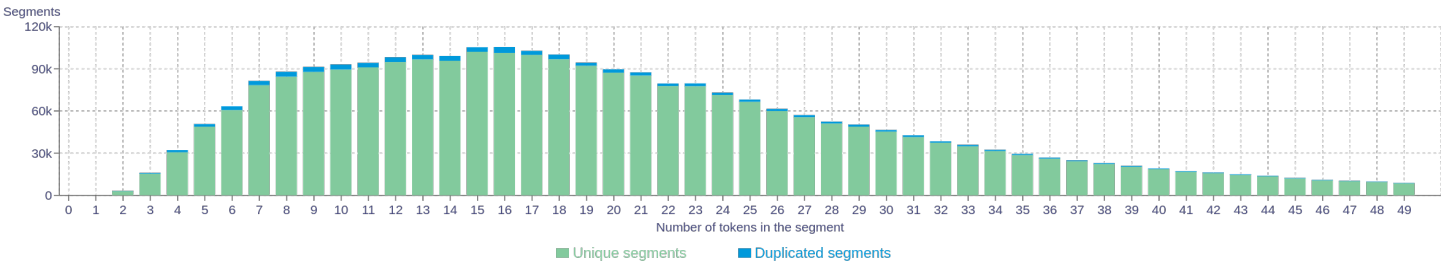


Target



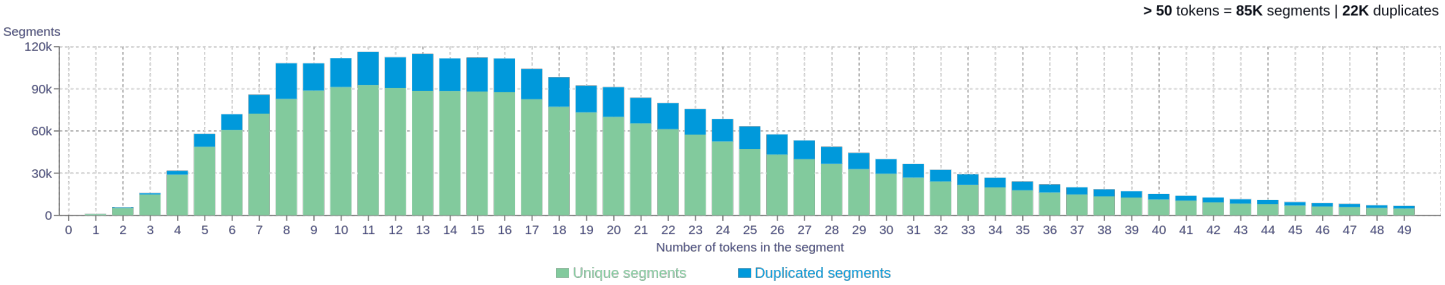
Source segment length distribution by token

<= 49 tokens = **2.5M** segments | **80K** duplicates
> 50 tokens = **120K** segments | **3.1K** duplicates

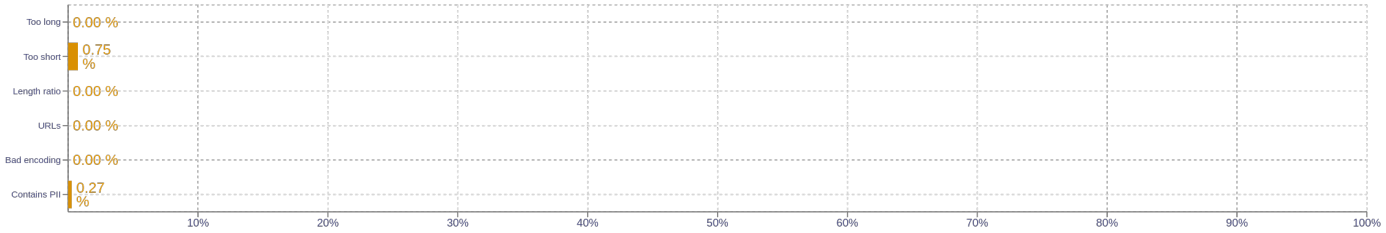


Target segment length distribution by token

<= 49 tokens = **2M** segments | **565K** duplicates
> 50 tokens = **85K** segments | **22K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	hotels | 143231 · also | 126529 · one | 113504 · hotel | 107412 · people | 99859
2	last hour | 27915 · people looked | 27872 · hotels within | 25971 · personal data | 19165 · hotels near | 14427
3	search for hotels | 6989 · like the game | 6245 · guest reviews yet | 5480 · choice for travelers | 5084 · last minute deals | 4781
4	looked at this hotel | 27849 · hotel in the last | 27849 · hotels on a map | 20321 · km from the city | 17892 · great choice for travelers | 5079
5	people looked at this hotel | 27849 · hotel in the last hour | 27849 · km from the city center | 10433 · km from the city centre | 7445 · great choice for travelers interested | 5079

Target n-grams

Size	n-grams
1	var | 161657 · getur | 122212 · hafa | 118042 · hótel | 113560 · einnig | 99824
2	síðustu klukkustund | 27836 · einstaklingar skoðuðu | 27828 · km fjarlægð | 14402 · sjá hótel | 12034 · getur verið | 10452
3	skoðuðu þetta hótel | 27828 · hótel á síðustu | 27828 · km frá miðbænum | 23270 · leitaðu að hótelum | 15926 · hótel í nágrenninu | 11841
4	hótel á síðustu klukkustund | 27828 · einstaklingar skoðuðu þetta hótel | 27828 · sjá hótel í nágrenninu | 11838 · frábært val fyrir ferðalanga | 7199 · býður upp á úrval | 6399
5	skoðuðu þetta hótel á síðustu | 27828 · hótel í nágrenninu á korti | 11838 · fara á kort sem sýnir | 6924 · hotels.com býður upp á úrval | 6190 · samgöngurþegar flogið er á staðinn | 5370

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