HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
2,697,582	60M	310,577,306	297.56 MB

TL tokens	TL characters	TL size
67M	352,141,678	357.62 MB

Dataset top 10 domains

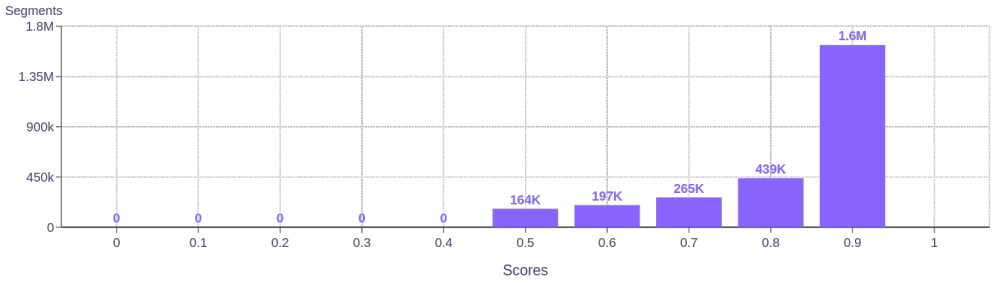
SL domain	Segments	TL domain	Segments
europa.eu	14.6%	europa.eu	9.0%
vsaduidoma.com	2.7%	vsaduidoma.com	2.7%
stealthsettings.com	2.4%	hookupdates.net	2.3%
hookupdates.net	2.3%	stealthsettings.com	2.0%
itsmygame.org	2.2%	wikipedia.org	2.0%
wikipedia.org	2.2%	datingmentor.org	1.9%
datingreviewer.net	1.9%	datingreviewer.net	1.9%
datingmentor.org	1.9%	besthookupwebsites.org	1.8%
besthookupwebsites.org	1.8%	itsmygame.org	1.8%
game-game.com	1.7%	game-game.com	1.7%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	70.0%	com	53.4%
ie	19.4%	ie	18.9%
eu	16.2%	org	12.7%
org	14.6%	net	11.9%
net	13.0%	eu	10.4%
nu	1.8%	nu	1.7%
co.uk	1.1%	gov.ie	1.0%
de	1.1%	de	0.9%
gov.ie	1.0%	ru	0.5%
ru	0.6%	co.uk	0.4%

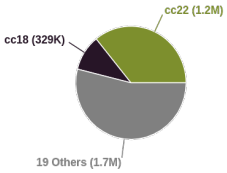
Translation likelihood

≥ 5 = 2.7M segments | **100.0%**
≥ 8 = 2.1M segments | **76.8%**
< 5 = 0 segments | **0.0%**



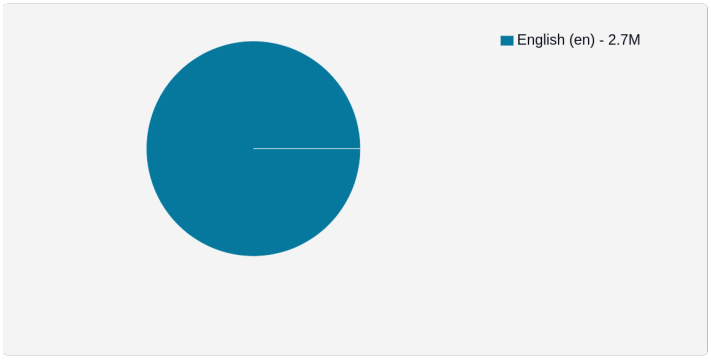
Collections

CC = 60.17%
IA = 39.83%

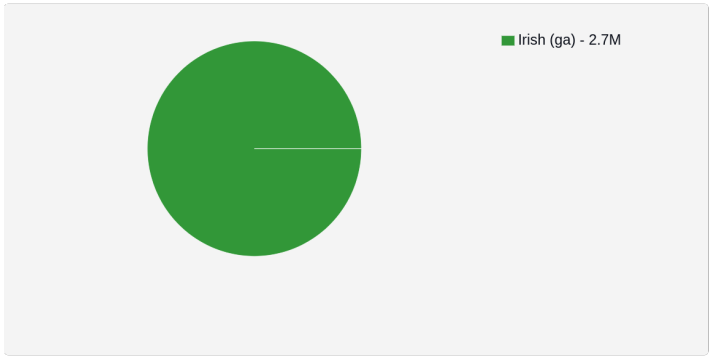


Language Distribution

Source

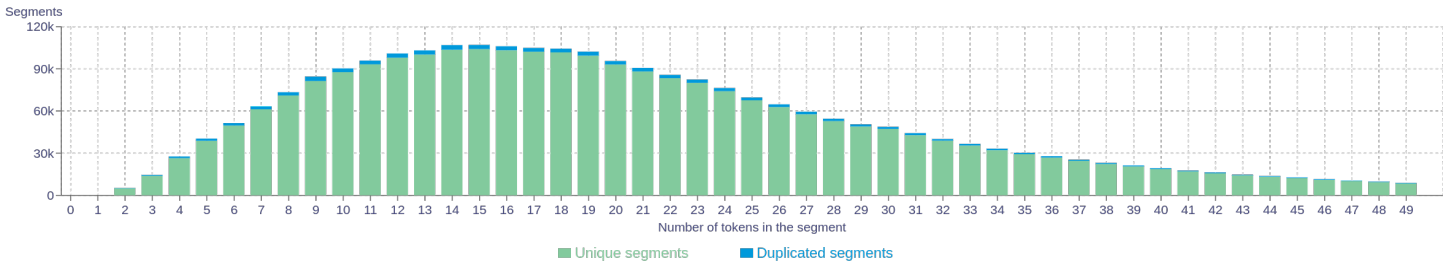


Target



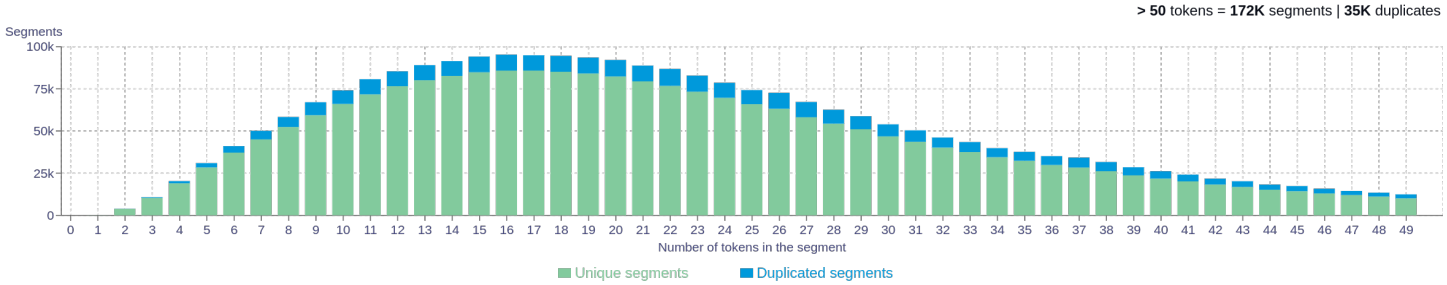
Source segment length distribution by token

<= 49 tokens = **2.5M** segments | **78K** duplicates
> 50 tokens = **120K** segments | **4.1K** duplicates

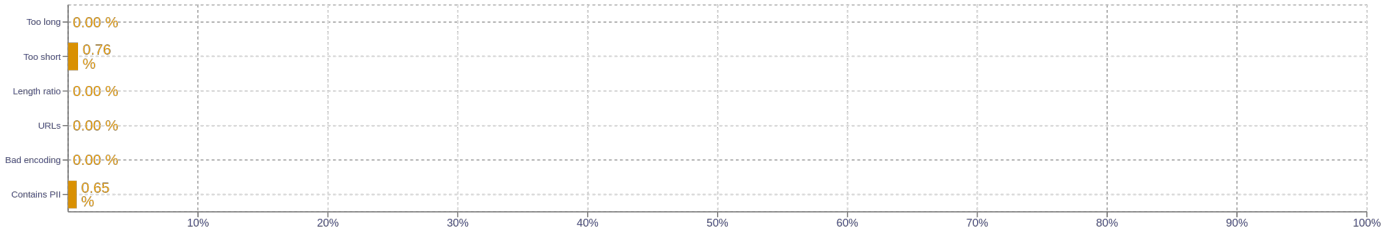


Target segment length distribution by token

<= 49 tokens = **2.2M** segments | **297K** duplicates
> 50 tokens = **172K** segments | **35K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	also | 127450 · one | 106209 · game | 105720 · use | 104056 · may | 95779
2	personal data | 19910 · european parliament | 15911 · member states | 12097 · personal information | 10272 · online game | 9704
3	like the game | 9523 · send the link | 6368 · share the game | 6362 · copy and send | 6359 · copy the code | 5888
4	link to a friend | 6360 · game with the world | 6358 · paste in the html | 5873 · code of your site | 5873 · referred to in article | 3456
5	parliament and of the council | 7904 · friend or all your friends | 6358 · copy and send the link | 6358 · copy the code and paste | 5874 · paste in the html code | 5873

Target n-grams

Size	n-grams
1	bhfuil | 350501 · féidir | 321757 · sin | 313927 · d | 263130 · atá | 246521
2	féidir leat | 120898 · níos mó | 87183 · suíomh gréasáin | 35043 · láithreán gréasáin | 29844 · mian leat | 29834
3	saor in aisce | 63643 · chuid is mó | 25520 · más mian leat | 20394 · maith leis sin | 16592 · fud an domhain | 14108
4	heorpa agus ón gcomhairle | 8154 · roinnt ar an cluiche | 6361 · sheoladh ar an nasc | 6360 · cluiche leis an domhan | 6360 · chóipeáil agus a sheoladh | 6359
5	cara nó gach do chairde | 6359 · más mian leat an cluiche | 6332 · cód html de do shuíomh | 5880 · cód agus greamaigh an cód | 5880 · chóipeáil an cód agus greamaigh | 5880

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