HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-cy.tsv	1/23/2025	English (en)	Welsh (cy)

Volumes

Segments	SL tokens	SL characters	SL size
3,867,402	89M	466,816,217	447.37 MB

TL tokens	TL characters	TL size
98M	486,025,270	469.79 MB

Dataset top 10 domains

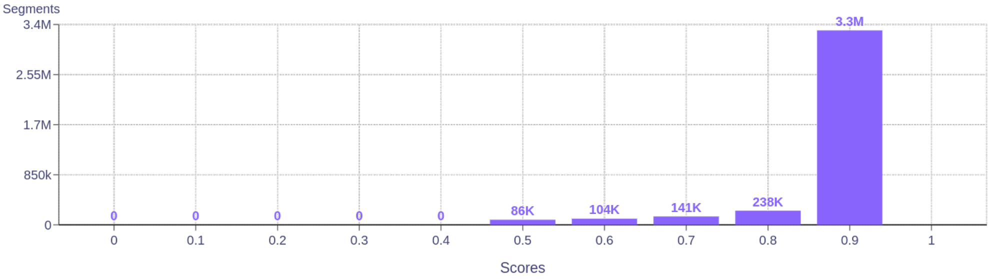
SL domain	Segments	TL domain	Segments
wikipedia.org	3.9%	wikipedia.org	3.4%
aber.ac.uk	3.5%	aber.ac.uk	3.2%
cardiff.ac.uk	3.4%	cardiff.ac.uk	3.1%
bangor.ac.uk	2.9%	bangor.ac.uk	2.2%
extendoffice.com	2.2%	extendoffice.com	1.9%
itsmygame.org	1.9%	swansea.ac.uk	1.6%
legislation.gov.uk	1.7%	moneyadviceservice.org.uk	1.5%
moneyadviceservice.org.uk	1.6%	itsmygame.org	1.5%
swansea.ac.uk	1.6%	legislation.gov.uk	1.4%
uwtsd.ac.uk	1.5%	uwtsd.ac.uk	1.3%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	49.0%	com	37.3%
ac.uk	18.9%	ac.uk	16.6%
org	16.9%	org	14.6%
gov.uk	13.1%	gov.uk	11.8%
org.uk	11.8%	org.uk	11.1%
wales	8.2%	cymru	8.2%
co.uk	8.2%	co.uk	6.8%
cymru	4.7%	wales	5.6%
net	3.6%	net	2.7%
police.uk	0.9%	police.uk	0.8%

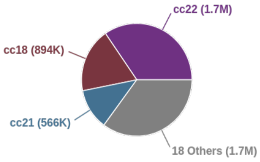
Translation likelihood

≥ 5 = 3.9M segments | **100.0%**
≥ 8 = 3.5M segments | **91.5%**
< 5 = 0 segments | **0.0%**



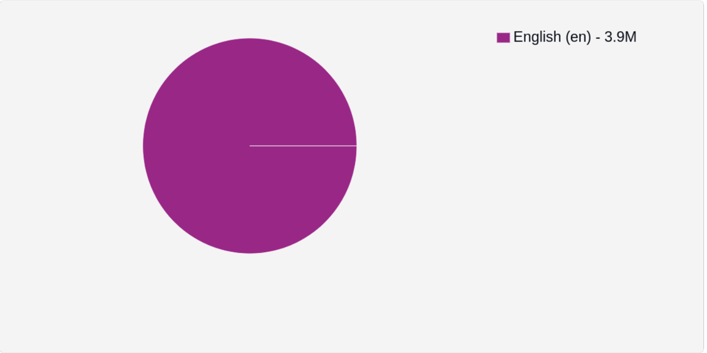
Collections

CC = 72.49%
IA = 27.51%

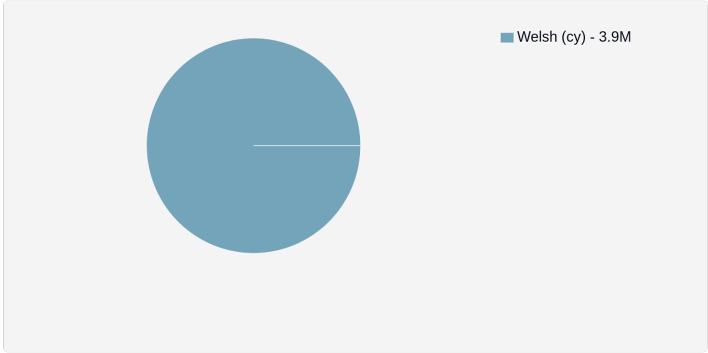


Language Distribution

Source

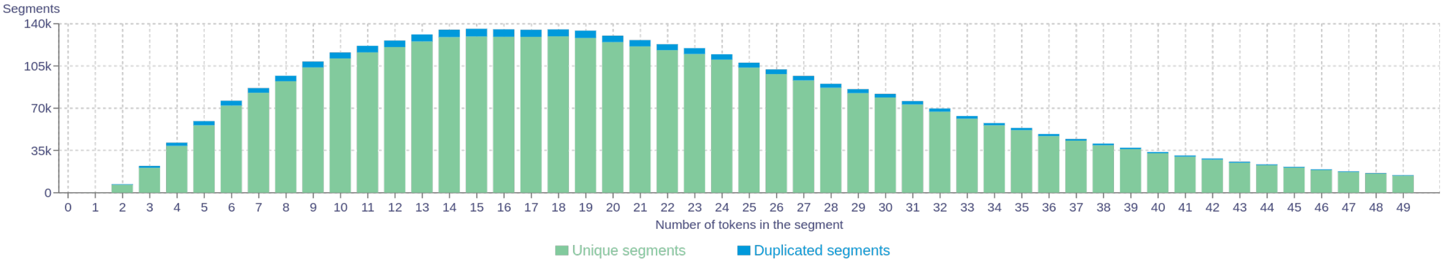


Target



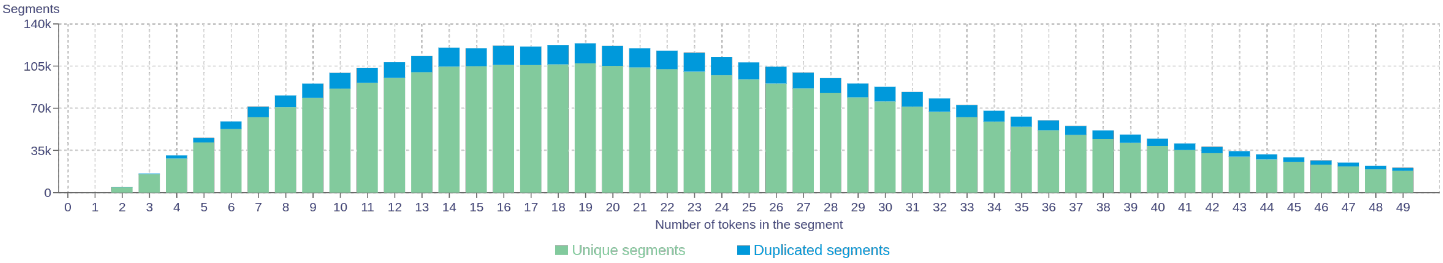
Source segment length distribution by token

<= 49 tokens = **3.6M** segments | **152K** duplicates
> 50 tokens = **163K** segments | **4K** duplicates

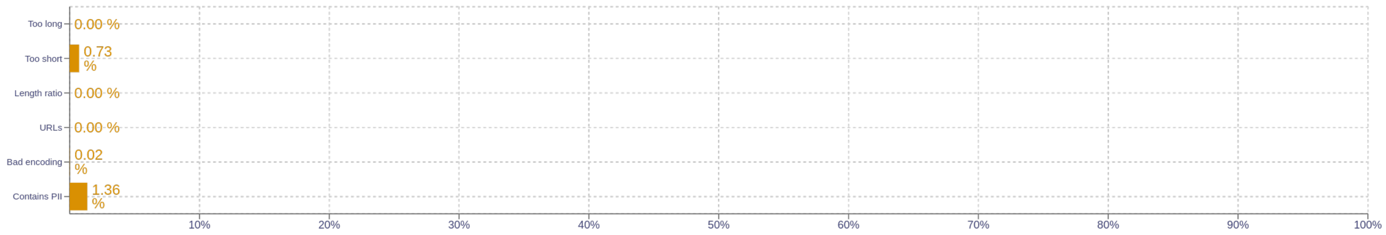


Target segment length distribution by token

<= 49 tokens = **3.2M** segments | **468K** duplicates
> 50 tokens = **247K** segments | **35K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	wales | 216543 also | 192585 information | 182355 use | 154190 people | 147039
2	welsh government | 31898 personal data | 31444 personal information | 23965 young people | 23413 please contact | 16105
3	like the game | 11512 terms and conditions | 9242 send the link | 7760 copy and send | 7757 share the game | 7756
4	link to a friend | 7755 game with the world | 7753 game with your best | 6063 paste in the html | 5860 code of your site | 5860
5	friend or all your friends | 7753 copy and send the link | 7753 game with your best friends | 6063 forget to rate this game | 6061 paste in the html code | 5860

Target n-grams

Size	n-grams
1	cymru | 239067 newydd | 148180 cynnwys | 135469 gwaith | 121327 gêm | 119877
2	data personol | 29069 llywodraeth cymru | 24239 gwybodaeth bersonol | 17361 gêm ar-lein | 11774 iechyd meddwl | 11617
3	rhagor o wybodaeth | 9155 anfon y ddolen | 7771 rhannu 'r gêm | 7755 copi ac anfon | 7755 ddolen i ffrind | 7754
4	ffrind neu eich ffrindiau | 7754 gêm hon gyda 'ch | 6096 gludo yn y cod | 5868 html ar eich safle | 5867 barn am y gêm | 4509
5	copi ac anfon y ddolen | 7754 anfon y ddolen i ffrind | 7754 rhannu gêm hon gyda 'ch | 6093 gêm hon gyda 'ch ffrindiau | 6077 gludo yn y cod html | 5867

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