HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-cym_Latn	9/18/2025	Welsh (cy)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,081,979	21,098,406	14,720,494 (69.77 %)	651M	3,171,744,689	2.99 GB

Top 10 domains

Domain	Docs	% of total
wikipedia.org	59K	5.47%
llyw.cymru	27K	2.53%
bbc.co.uk	25K	2.34%
library.wales	20K	1.81%
bbc.com	18K	1.70%
360.cymru	18K	1.68%
testunau.org	15K	1.42%
s4c.cymru	14K	1.30%
cardiff.ac.uk	13K	1.22%
golwg360.cymru	13K	1.19%

Top 10 TLDs

Domain	Docs	% of total
com	316K	29.25%
cymru	185K	17.10%
org	149K	13.80%
co.uk	92K	8.55%
ac.uk	85K	7.87%
gov.uk	65K	6.05%
org.uk	65K	5.99%
wales	46K	4.28%
net	15K	1.40%
zone	6.4K	0.59%

Register labels

- HI - 3.3%
- ID - 0.2%
- IN - 26.9%
- IP - 15.5%
- LY - 0.2%
- MIX - 5.7%
- NA - 23.7%
- OP - 4.8%
- SP - 0.6%
- UNK - 19.0%

- HI_other - 2.8%
- HI_re - 0.6%
- ID_other - 0.2%
- IN_dtp - 10.1%
- IN_en - 5.7%
- IN_fi - 0.0%
- IN_lt - 0.8%
- IN_other - 10.3%
- IN_ra - 0.1%
- IP_ds - 11.8%
- IP_other - 3.7%
- LY_other - 0.2%
- MIX - 5.7%
- NA_nb - 3.4%
- NA_ne - 15.8%
- NA_other - 3.2%
- NA_sr - 1.3%
- OP_av - 0.4%
- OP_ob - 0.8%
- OP_other - 0.7%
- OP_rs - 2.3%
- OP_rv - 0.7%
- SP_it - 0.4%
- SP_other - 0.2%
- UNK - 19.0%
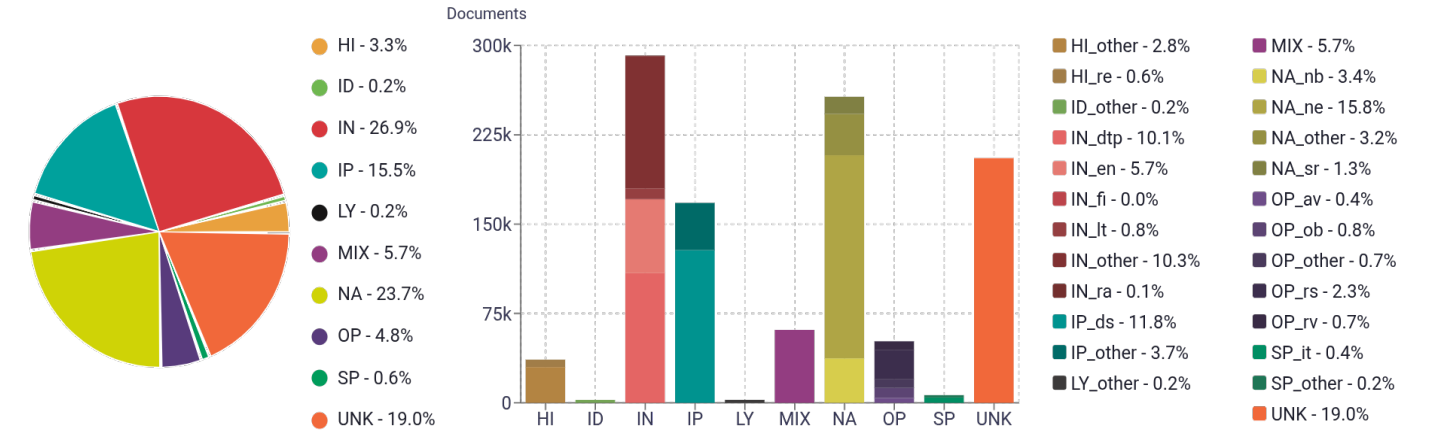
MT:15.4% | 167K Documents

Documents size (in segments) ⓘ

≤ 25 segments **79.95%** (865K documents)
> 25 segments **20.05%** (217K documents)

Document collections

CC = 93.43%
IA = 6.57%

67 Others (1.1M)
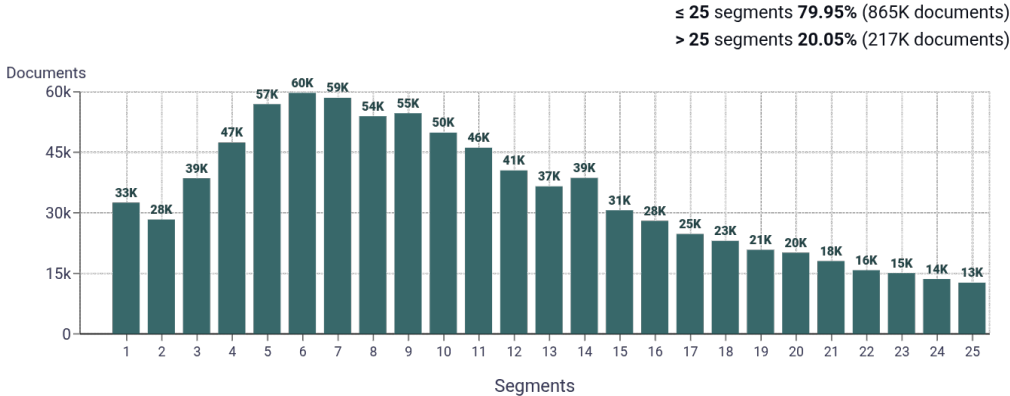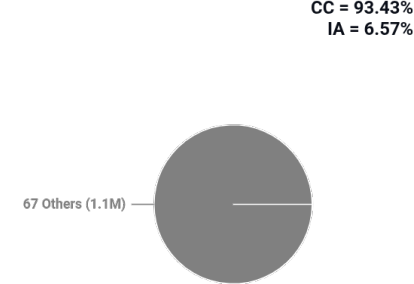
Language Distribution

Number of segments in the Welsh (cy) corpus

- Welsh (cy) - 17M **(80.1%)**
- English (en) - 1.4M **(6.5%)**
- Spanish (es) - 531K **(2.5%)**
- German (de) - 423K **(2.0%)**
- French (fr) - 298K **(1.4%)**
- Italian (it) - 271K **(1.3%)**
- Portuguese (pt) - 97K **(0.5%)**
- Dutch (nl) - 87K **(0.4%)**
- Lojban (jbo) - 72K **(0.3%)**
- Cornish (kw) - 70K **(0.3%)**
- 165 Others - 988K **(4.7%)**
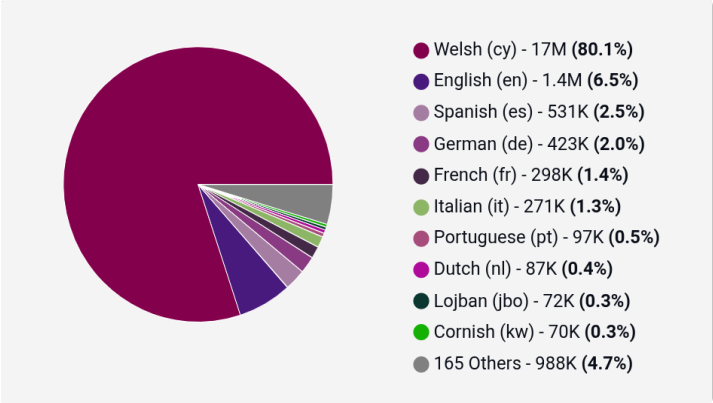
Percentage of segments in Welsh (cy) inside documents
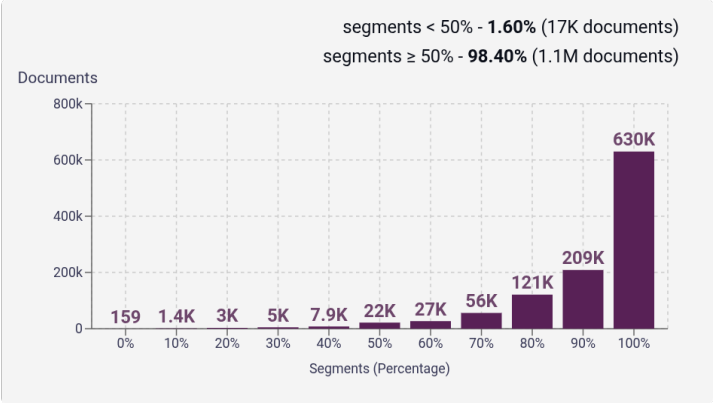
segments < 50% - **1.60%** (17K documents)
segments ≥ 50% - **98.40%** (1.1M documents)

Documents

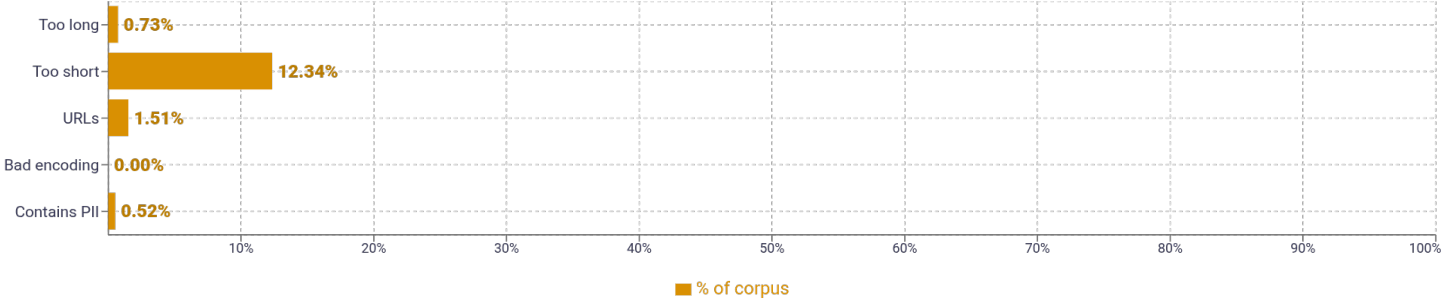|	
0%	159
10%	1.4K
20%	3K
30%	5K
40%	7.9K
50%	22K
60%	27K
70%	56K
80%	121K
90%	209K
100%	630K

Segments (Percentage)

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (1.1M documents)

Documents

Segment length distribution by token

≤ 49 tokens = **17M** segments | 5.9M duplicates
> 50 tokens = **3.7M** segments | 516K duplicates

Segments

Number of tokens in the segment

Segment noise distribution

Too long	**0.73%**
Too short	**12.34%**
URLs	**1.51%**
Bad encoding	**0.00%**
Contains PII	**0.52%**

■ % of corpus
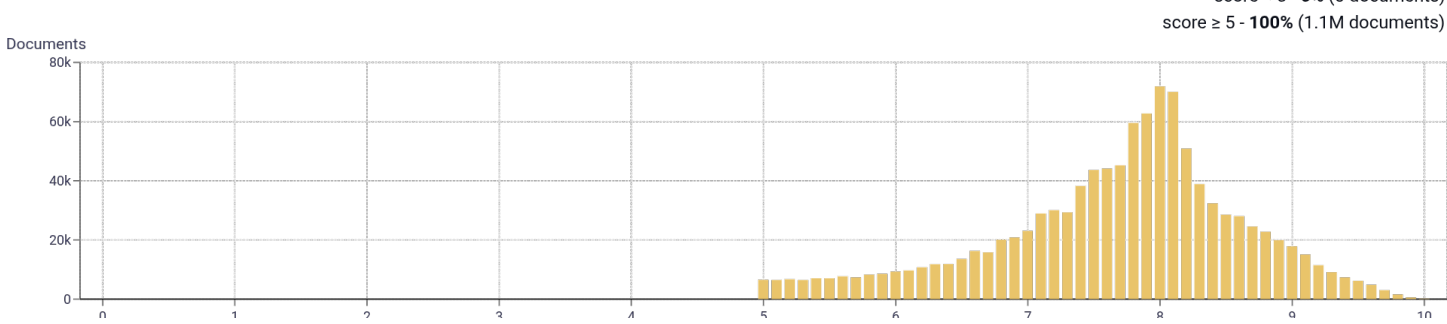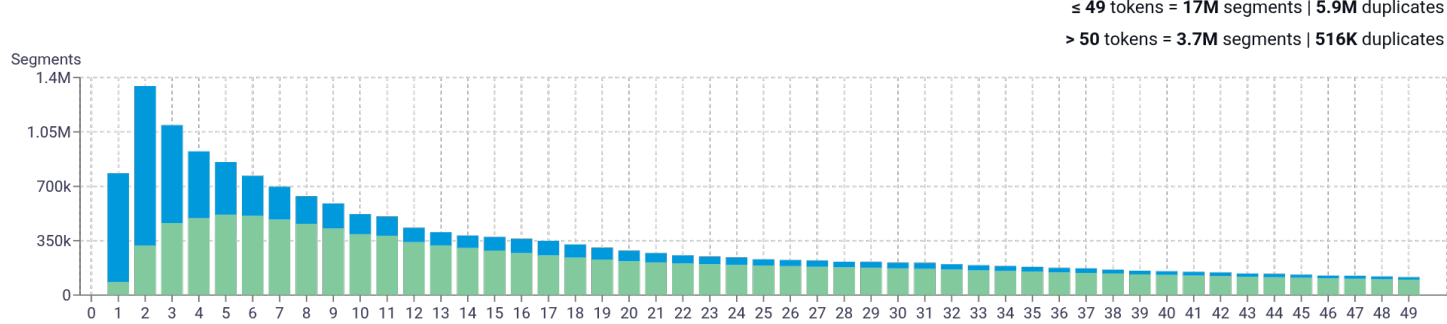
Frequent n-grams

SIZE	N-GRAMS	
1	cymru | 1,259,774 newydd | 938,436 gwaith | 677,413 amser | 656,153 cynnwys | 617,162	⧉
2	llywodraeth cymru | 124,047 golygu cod | 89,410 awyr agored | 58,998 unol daleithiau | 55,706 pobl ifanc | 47,850	⧉
3	cod y dudalen | 48,176 rhagor o wybodaeth | 41,413 cynnal a chadw | 25,907 ragor o wybodaeth | 25,653 gynnwys y canlynol | 25,228	⧉
4	golygu cod y dudalen | 48,172 poblogaidd y flwyddyn honno | 24,282 ganwyd y cyfarwyddwr ffilm | 21,854 actorion yn y ffilm | 16,970 newid yn yr hinsawdd | 16,561	⧉
5	ffilmiau gan gynnwys y canlynol | 24,303 ffilm fwyaf poblogaidd y flwyddyn | 24,289 sgwennwyd y sgript yn wreiddiol | 21,565 ellir dweud fod y gwaith | 8,453 gwaith wedi cyrraedd enwogrwydd byd-eang | 8,451	⧉

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

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	**How-to or instructions**	HI	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	**Informational persuasion**	IP	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	**Opinion**	OP
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	**Informational description**	IN	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				