HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
198,265	4.7M	24,245,327	23.22 MB

TL tokens	TL characters	TL size
4.8M	25,528,845	24.93 MB

Dataset top 10 domains

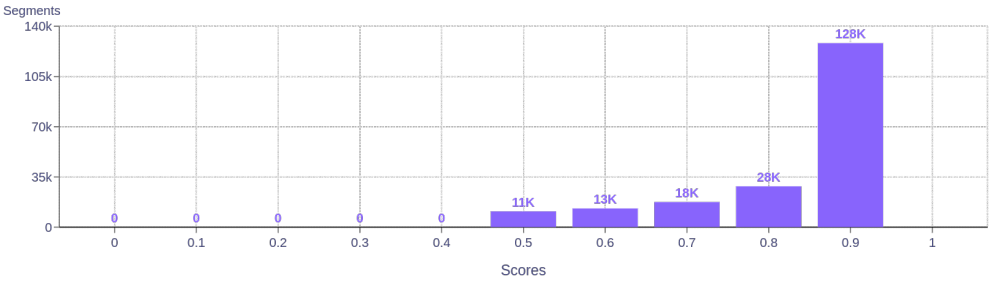
SL domain	Segments	TL domain	Segments
wikipedia.org	40.1%	wikipedia.org	35.3%
itsmygame.org	3.4%	itsmygame.org	2.6%
vsaduidoma.com	2.6%	vsaduidoma.com	2.5%
skolarbete.nu	1.8%	skolarbete.nu	1.8%
flashgames312.com	1.3%	mintarticles.com	1.3%
schools-wikipedia.org	1.3%	flashgames312.com	1.2%
mintarticles.com	1.3%	turismo.gal	0.8%
educationbro.com	1.2%	soft-free-download.com	0.7%
todojuegosgratis.es	0.9%	zientzia.eus	0.7%
wikimedia.org	0.9%	todojuegosgratis.es	0.7%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	65.6%	org	47.2%
org	57.5%	com	46.7%
es	9.1%	es	10.5%
net	4.5%	gal	4.6%
eu	2.8%	net	3.3%
gal	2.6%	eu	2.4%
nu	2.0%	nu	1.9%
gob.es	1.4%	gob.es	1.4%
co.uk	1.1%	eus	0.9%
eus	1.0%	ru	0.8%

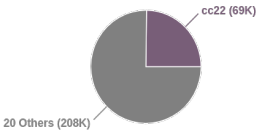
Translation likelihood

≥ 5 = 198K segments | **100.0%**
≥ 8 = 157K segments | **79.0%**
< 5 = 0 segments | **0.0%**



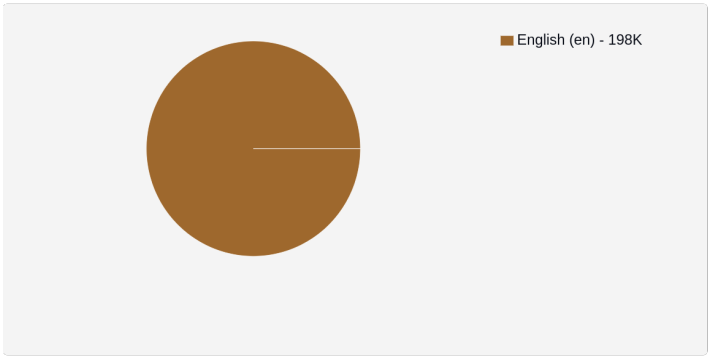
Collections

CC = 59.12%
IA = 40.88%

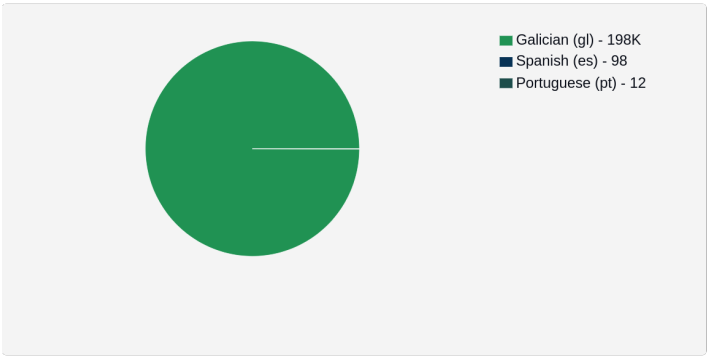


Language Distribution

Source

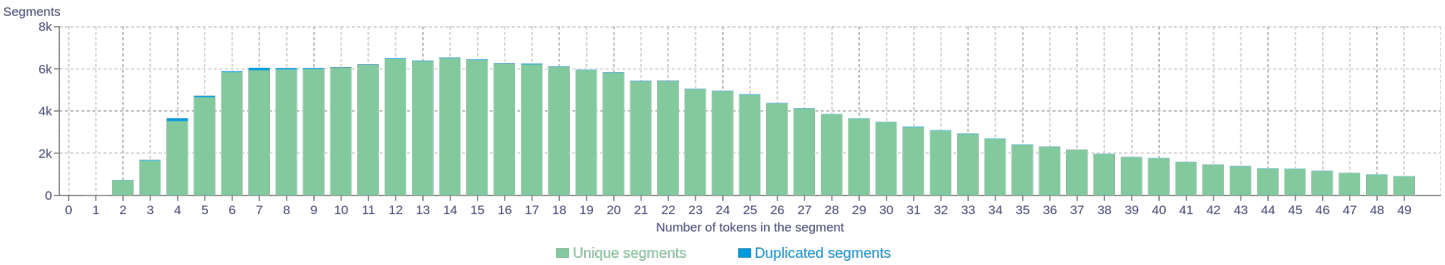


Target



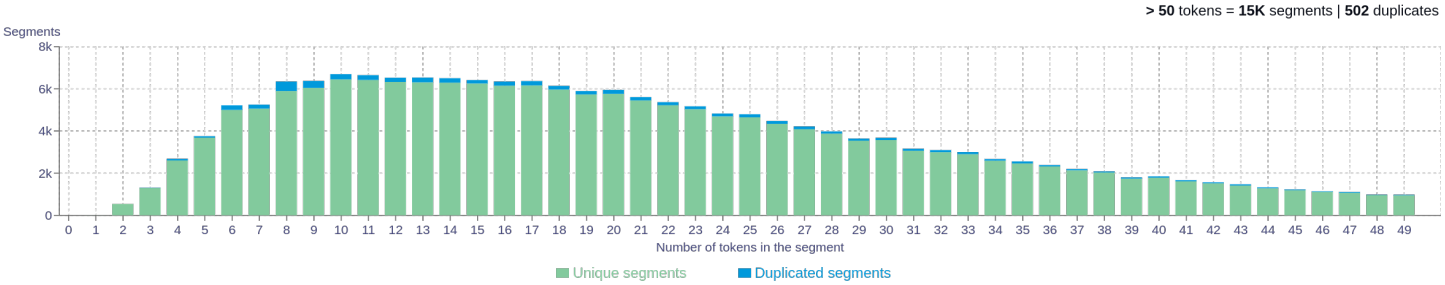
Source segment length distribution by token

<= 49 tokens = **183K** segments | **1.1K** duplicates
> 50 tokens = **14K** segments | **43** duplicates

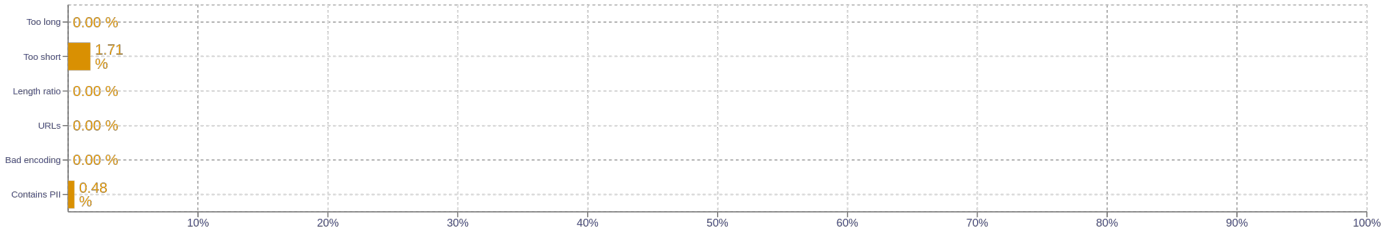


Target segment length distribution by token

<= 49 tokens = **178K** segments | **6K** duplicates
> 50 tokens = **15K** segments | **502** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	data | 8629 · also | 8523 · one | 8479 · use | 8066 · time | 6493
2	personal data | 3014 · united states | 1009 · data protection | 850 · third parties | 800 · privacy policy | 596
3	protected from spambots | 364 · protection of personal | 361 · play the game | 322 · processing of personal | 317 · terms and conditions | 310
4	address is being protected | 364 · use of the website | 358 · protection of personal data | 348 · processing of personal data | 314 · paste in the html | 303
5	email address is being protected | 340 · paste in the html code | 303 · html code of your site | 303 · copy the code and paste | 303 · technical characteristics of the game | 259

Target n-grams

Size	n-grams
1	datos | 8974 · web | 8966 · editar | 8815 · información | 6280 · entre | 5993
2	sitio web | 4137 · datos persoais | 2318 · estados unidos | 1543 · páxina web | 1473 · correo electrónico | 1040
3	editar a fonte | 3953 · protección de datos | 1308 · datos de carácter | 662 · política de privacidade | 557 · enderezo de correo | 446
4	datos de carácter persoal | 650 · tratamento dos seus datos | 445 · robots de correo lixo | 306 · pega o código html | 303 · html do seu sitio | 303
5	protección de datos de carácter | 361 · protexido dos robots de correo | 306 · correo está a ser protexido | 306 · código html do seu sitio | 303 · código e pega o código | 303

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