HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
1,491,873	36M	186,821,902	178.85 MB

TL tokens	TL characters	TL size
29M	194,307,903	185.81 MB

Dataset top 10 domains

SL domain	Segments	TL domain	Segments
wikipedia.org	16.6%	wikipedia.org	15.2%
vsaduidoma.com	3.4%	vsaduidoma.com	3.3%
sacred-texts.com	2.1%	sacred-texts.com	2.3%
libreoffice.org	2.1%	amightywind.com	1.5%
amightywind.com	1.6%	lifebogger.com	1.3%
lifebogger.com	1.2%	libreoffice.org	1.1%
forvo.com	1.1%	astelus.com	1.0%
flashgames312.com	1.1%	flashgames312.com	1.0%
astelus.com	1.0%	zientzia.eus	1.0%
zientzia.eus	1.0%	itsmygame.org	0.9%

Dataset top 10 TLDs

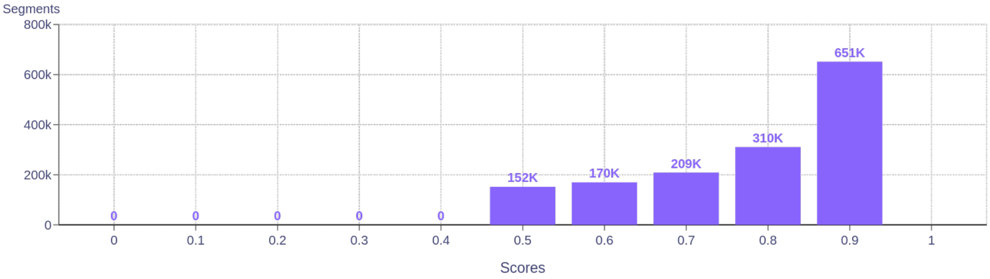
SL domain	Segments	TL domain	Segments
com	69.2%	com	53.1%
org	32.1%	org	26.3%
eus	11.9%	eus	15.9%
es	11.5%	es	11.4%
net	5.4%	net	3.4%
eu	3.5%	eu	2.9%
info	1.3%	gob.es	1.2%
gob.es	1.3%	info	1.0%
fr	0.9%	com.br	0.6%
co.uk	0.9%	fr	0.6%

Translation likelihood

≥ 5 = 1.5M segments | **100.0%**

≥ 8 = 962K segments | **64.5%**

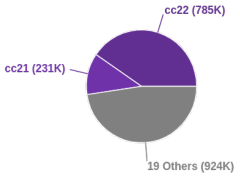
< 5 = 0 segments | **0.0%**



Collections

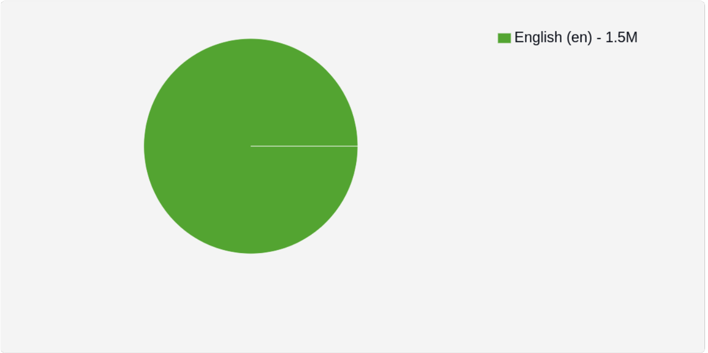
CC = 73.20%

IA = 26.80%

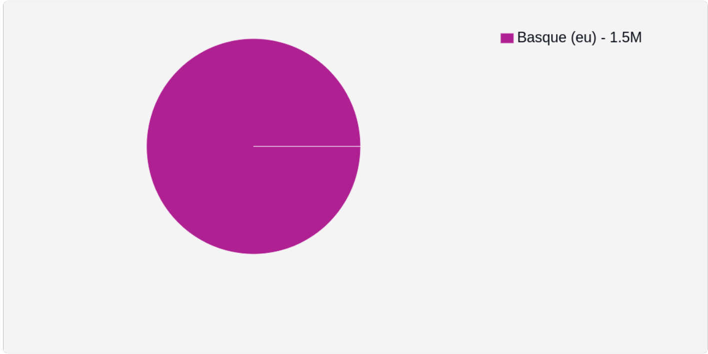


Language Distribution

Source



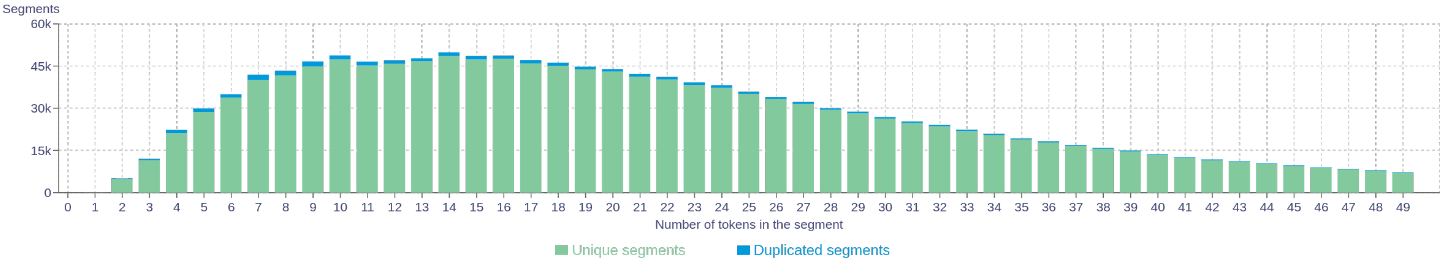
Target



Source segment length distribution by token

<= 49 tokens = **1.3M** segments | **36K** duplicates

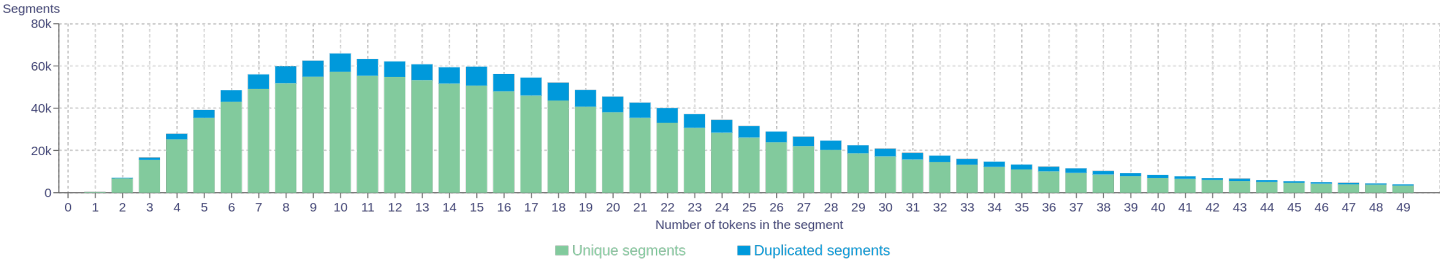
> 50 tokens = **108K** segments | **2K** duplicates



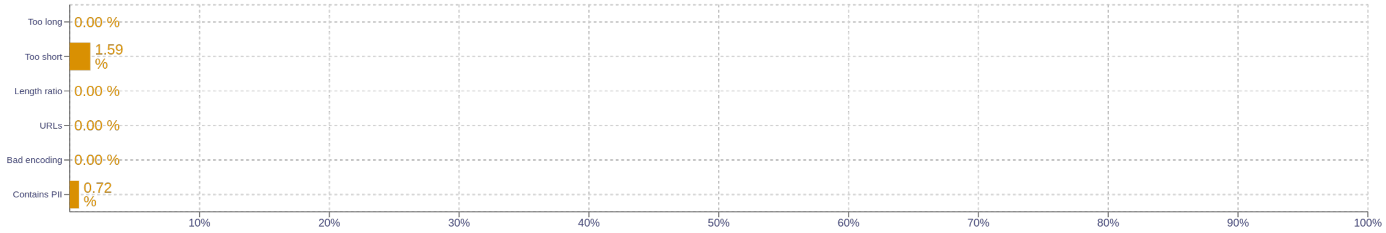
Target segment length distribution by token

<= 49 tokens = **1.2M** segments | **208K** duplicates

> 50 tokens = **53K** segments | **7.3K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	data | 74725 use | 63377 also | 63094 one | 61832 information | 58259
2	personal data | 24556 basque country | 11085 third parties | 7146 data protection | 6672 united states | 6476
3	play the game | 3013 protected from spambots | 2952 protection of personal | 2754 reserves the right | 2323 terms and conditions | 2032
4	address is being protected | 2959 protection of personal data | 2694 use of the website | 2521 processing of personal data | 1947 university of the basque | 1562
5	email address is being protected | 2868 university of the basque country | 1541 processing of your personal data | 1303 need javascript enabled to view | 1112 information society and electronic commerce | 1028

Target n-grams

Size	n-grams
1	behar | 77928 izango | 61214 egiten | 54935 duen | 53646 nahi | 48401
2	ahal izango | 17933 datu pertsonalak | 15612 iturburu kodea | 11163 aldatu iturburu | 11103 aukera ematen | 10001
3	aldatu iturburu kodea | 11103 ameriketako estatu batuetako | 2093 buruzko informazio gehiago | 1783 aukera ematen diote | 1594 loratagia eta txabola | 1513
4	elektroniko hau spambot-etatik babestuta | 1327 helbide elektroniko hau spambot-etatik | 1304 ditugu zure datu pertsonalak | 1029 jar zaitez gurekin harremanetan | 985 gizartearen eta merkataritza elektronikoaren | 916
5	helbide elektroniko hau spambot-etatik babestuta | 1304 gizartearen eta merkataritza elektronikoaren zerbitzuei | 819 helbide elektroniko honek spam bot-en | 739 elektroniko honek spam bot-en kontrako | 580 datu pertsonalak babesteari buruzko abenduaren | 567

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