HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-ca.tsv	1/27/2025	English (en)	Catalan (ca)

Volumes

Segments	SL tokens	SL characters	SL size
13,080,859	319M	1,675,733,471	1.57 GB

TL tokens	TL characters	TL size
359M	1,785,980,705	1.71 GB

Dataset top 10 domains

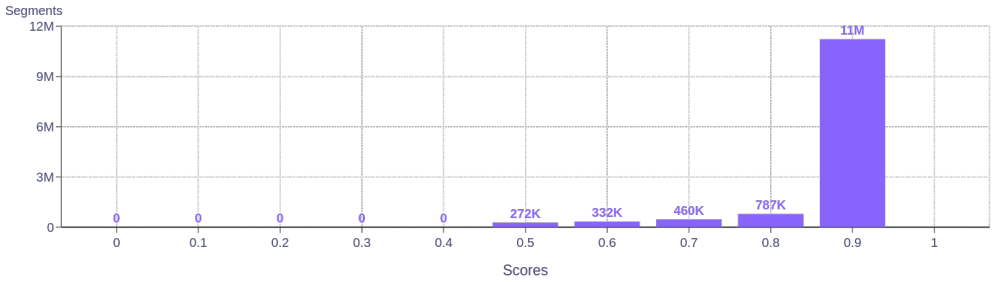
SL domain	Segments	TL domain	Segments
wikipedia.org	22.4%	wikipedia.org	18.7%
google.com	9.2%	google.com	4.4%
agoda.com	5.1%	agoda.com	3.7%
booking.com	3.0%	booking.com	1.9%
marxists.org	1.8%	marxists.org	1.7%
soin-et-nature.com	1.1%	soin-et-nature.com	1.0%
airwise.com	0.7%	destinia.cat	0.8%
upc.edu	0.7%	destinia.ad	0.7%
lucasfox.com	0.7%	upc.edu	0.7%
vsaduidoma.com	0.7%	vsaduidoma.com	0.7%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	91.2%	com	65.5%
org	39.4%	org	31.1%
cat	9.1%	cat	15.7%
es	8.5%	es	8.5%
net	5.0%	net	4.3%
edu	2.0%	edu	1.9%
co.uk	1.7%	ad	1.2%
eu	1.6%	eu	1.1%
info	0.8%	info	0.6%
de	0.6%	gob.es	0.3%

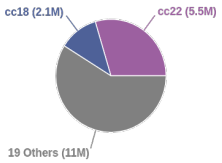
Translation likelihood

≥ 5 = 13M segments | **100.0%**
≥ 8 = 12M segments | **91.9%**
< 5 = 0 segments | **0.0%**



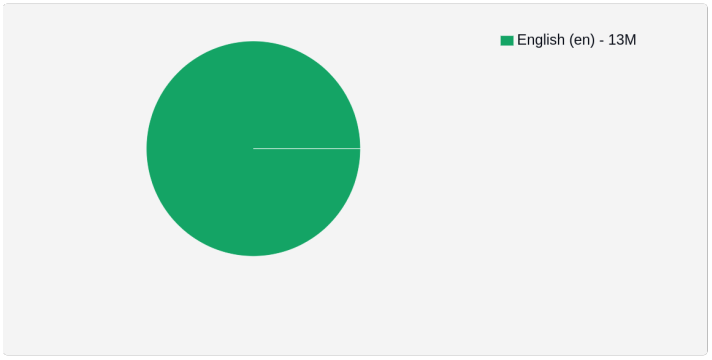
Collections

CC = 65.05%
IA = 34.95%

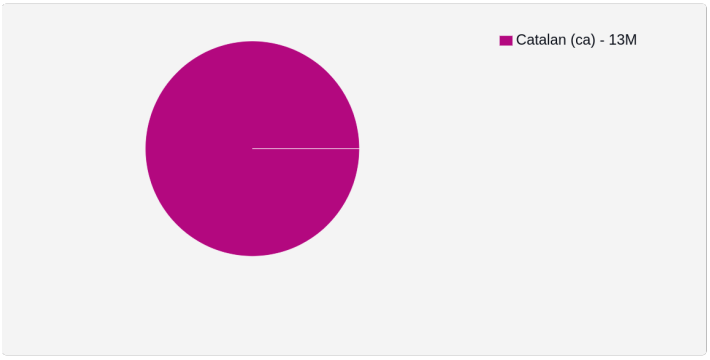


Language Distribution

Source

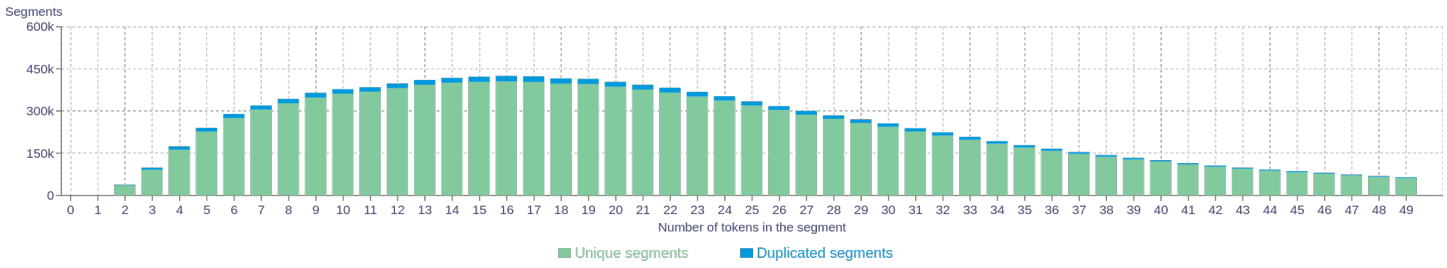


Target



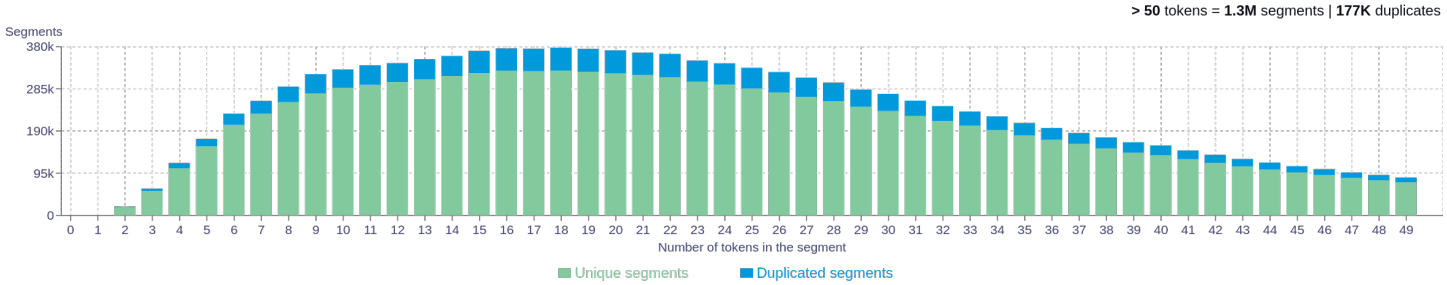
Source segment length distribution by token

<= 49 tokens = **12M** segments | **548K** duplicates
> 50 tokens = **918K** segments | **35K** duplicates

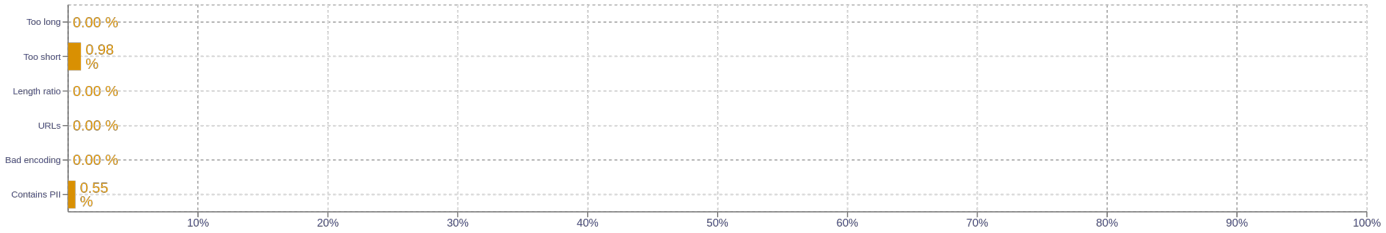


Target segment length distribution by token

<= 49 tokens = **10M** segments | **1.6M** duplicates
> 50 tokens = **1.3M** segments | **177K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	also | 629866 · one | 576132 · use | 465340 · new | 433494 · time | 412425
2	personal data | 106794 · united states | 49450 · third parties | 43581 · de la | 42194 · free wi-fi | 38370
3	reserves the right | 23663 · wi-fi in public | 21978 · around the world | 17710 · choice for travelers | 17175 · terms and conditions | 16788
4	wi-fi in public areas | 21964 · wi-fi in all rooms | 21704 · use of the website | 13755 · one of the best | 12379 · address is being protected | 10535
5	free wi-fi in all rooms | 21685 · great choice for travelers interested | 9099 · one of the most important | 8393 · email address is being protected | 7676 · neighborhood is a great choice | 7614

Target n-grams

Size	n-grams
1	pot | 516049 · web | 481565 · lloc | 462284 · dades | 422758 · informació | 398015
2	lloc web | 176237 · dades personals | 89910 · pàgina web | 82136 · estats units | 69010 · correu electrònic | 64805
3	modifica el codi | 64734 · protecció de dades | 42772 · dur a terme | 23608 · reserva el dret | 23165 · dades de caràcter | 22042
4	centre de la ciutat | 22877 · dades de caràcter personal | 21800 · zones públiques amb wi-fi | 19957 · fantàstica per als viatgers | 13121 · viatgers que els interessa | 12745
5	gratuït en totes les habitacions | 22082 · opció fantàstica per als viatgers | 13121 · barri és una opció fantàstica | 10706 · protecció de dades de caràcter | 10457 · tractament de les seves dades | 8596

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