HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
1,529,471	44M	233,258,710	223.4 MB

TL tokens	TL characters	TL size
40M	252,579,907	250.4 MB

Dataset top 10 domains

SL domain	Segments	TL domain	Segments
europa.eu	82.8%	europa.eu	64.9%
wikipedia.org	4.3%	wikipedia.org	3.4%
vsaduidoma.com	1.7%	vsaduidoma.com	1.7%
itsmygame.org	1.6%	itsmygame.org	1.5%
gov.mt	1.5%	wondershare.com	1.5%
wondershare.com	1.5%	gov.mt	1.4%
skolarbete.nu	1.3%	skolarbete.nu	1.3%
flashgames312.com	1.1%	flashgames312.com	1.0%
rxed.eu	0.9%	mintarticles.com	0.9%
mintarticles.com	0.9%	rxed.eu	0.8%

Dataset top 10 TLDs

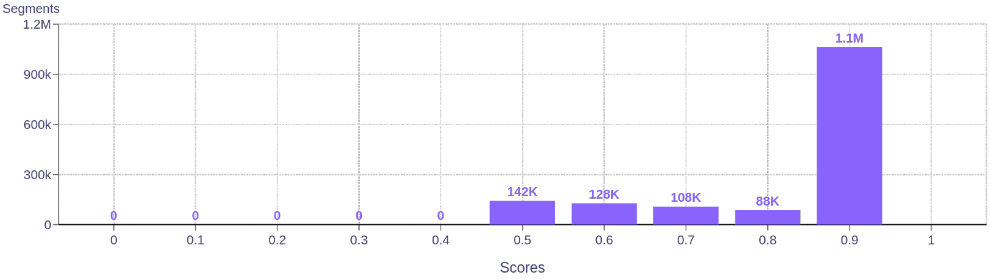
SL domain	Segments	TL domain	Segments
eu	86.0%	eu	67.6%
com	42.8%	com	33.5%
org	12.1%	org	10.1%
net	2.3%	mt	2.0%
mt	1.9%	net	1.6%
nu	1.4%	nu	1.4%
org.mt	1.0%	org.mt	1.1%
de	0.8%	com.mt	1.0%
com.mt	0.7%	de	0.7%
co.uk	0.6%	info	0.3%

Translation likelihood

≥ 5 = 1.5M segments | **100.0%**

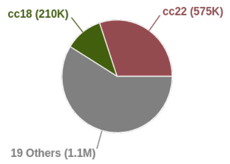
≥ 8 = 1.2M segments | **75.4%**

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



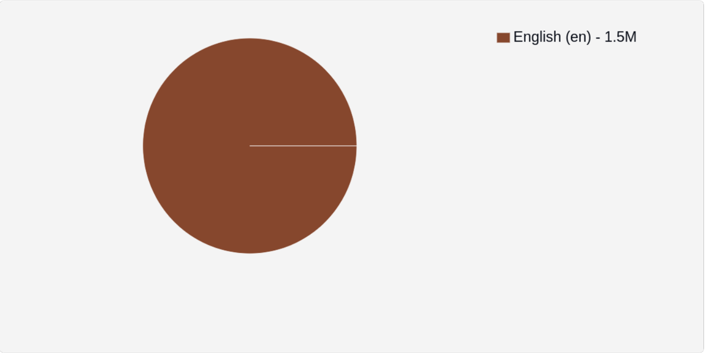
Collections

CC = 60.14%
IA = 39.86%

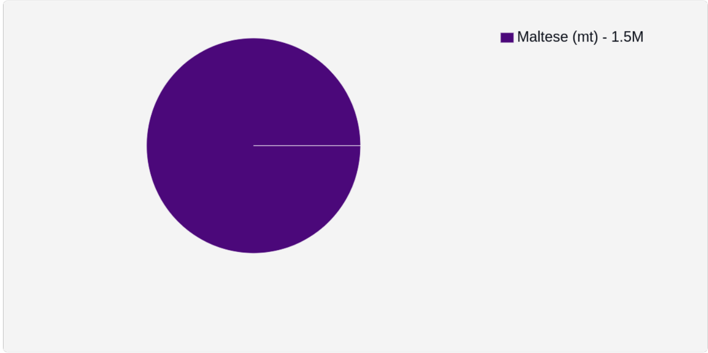


Language Distribution

Source



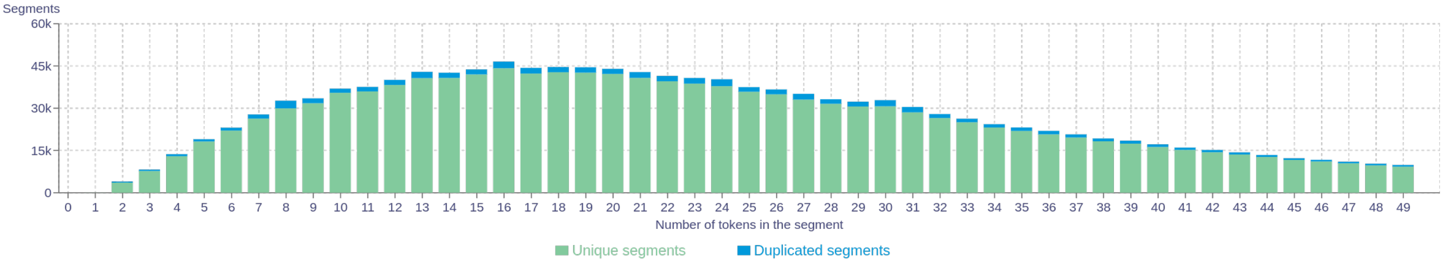
Target



Source segment length distribution by token

<= 49 tokens = **1.3M** segments | **67K** duplicates

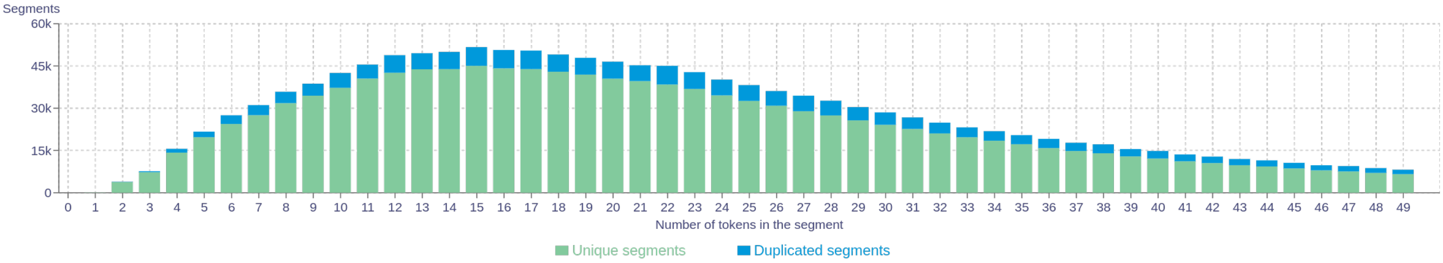
> 50 tokens = **183K** segments | **9K** duplicates



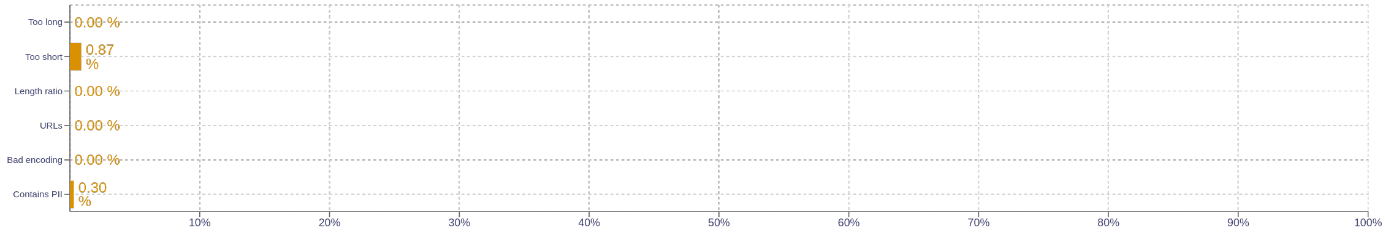
Target segment length distribution by token

<= 49 tokens = **1.2M** segments | **190K** duplicates

> 50 tokens = **143K** segments | **29K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	article | 141975 european | 114745 member | 101196 eu | 97364 may | 87957
2	member states | 55686 member state | 38327 european parliament | 29353 european union | 25396 personal data | 14073
3	accordance with article | 8686 like the game | 7068 court of justice | 6726 within the meaning | 6672 pursuant to article | 6146
4	referred to in article | 12636 referred to in paragraph | 6077 ec of the european | 4712 link to a friend | 4633 game with the world | 4633
5	parliament and of the council | 15027 ec of the european parliament | 4699 friend or all your friends | 4633 copy and send the link | 4633 within the meaning of article | 4043

Target n-grams

Size	n-grams
1	f | 288277 b | 269979 'mod | 86451 l-artikolu | 64897 skont | 64023
2	b 'mod | 86327 f 'dan | 40504 l-istati membri | 28745 'mod partikolari | 25378 tal-parlament ewropew | 21479
3	b 'mod partikolari | 25365 ewropew u tal-kunsill | 15667 f 'dan ir-rigward | 8356 f 'dan il-każ | 7611 l-link lil habib | 4633
4	tal-parlament ewropew u tal-kunsill | 15594 kopja u jibgħat l-link | 4633 jibgħat l-link lil habib | 4633 talba għal deċiżjoni preliminari | 2886 l-istati membri għandhom jiżguraw | 2514
5	l-link lil habib jew ħbieb | 4633 sib il-kliem ingliż li jibdew | 2236 suġġett talba għal deċiżjoni preliminari | 2111 kliem ingliż ġdid bl-istess pari | 2099 joħloq kliem ingliż ġdid bl-istess | 2099

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