HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-mk.tsv	1/24/2025	English (en)	Macedonian (mk)

Volumes

Segments	SL tokens	SL characters	SL size
3,991,617	92M	479,196,200	459.09 MB

TL tokens	TL characters	TL size
90M	497,620,970	852.16 MB

Dataset top 10 domains

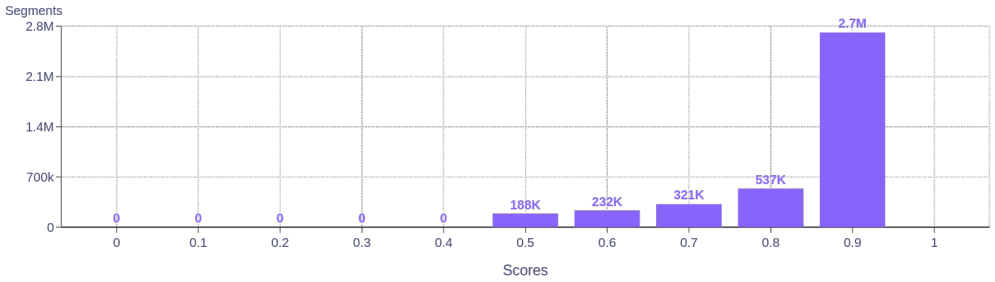
SL domain	Segments	TL domain	Segments
wikipedia.org	26.4%	wikipedia.org	21.0%
itsmygame.org	2.9%	skopelitissa.com	2.2%
masterstudies.com	2.7%	itsmygame.org	2.1%
skopelitissa.com	2.2%	voanews.com	1.8%
voanews.com	1.8%	masterstudies.mk	1.7%
educationbro.com	1.7%	vsaduidoma.com	1.7%
vsaduidoma.com	1.7%	rbth.com	1.3%
academiccourses.com	1.4%	skolarbete.nu	1.1%
rbth.com	1.3%	ofunnygames.com	1.0%
ofunnygames.com	1.2%	globalvoices.org	1.0%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	68.6%	com	47.5%
org	45.4%	org	34.4%
mk	7.5%	mk	15.8%
net	6.6%	net	5.1%
org.mk	2.6%	com.mk	4.3%
com.mk	2.4%	org.mk	2.7%
eu	2.1%	gov.mk	1.4%
co.uk	1.5%	edu.mk	1.3%
edu.mk	1.3%	eu	1.3%
gov.mk	1.3%	nu	1.1%

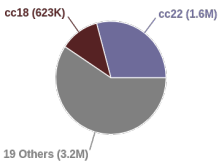
Translation likelihood

≥ 5 = 4M segments | **100.0%**
≥ 8 = 3.3M segments | **81.4%**
< 5 = 0 segments | **0.0%**



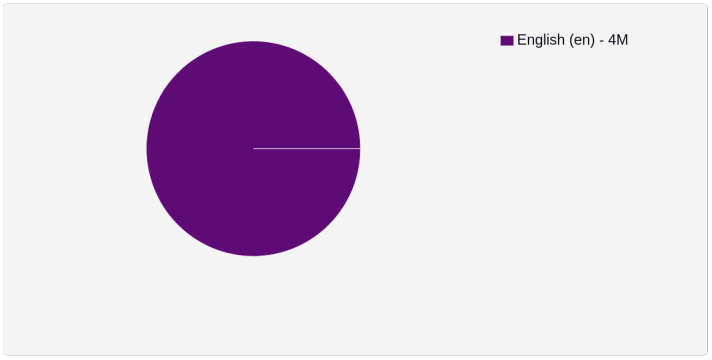
Collections

CC = 61.72%
IA = 38.28%

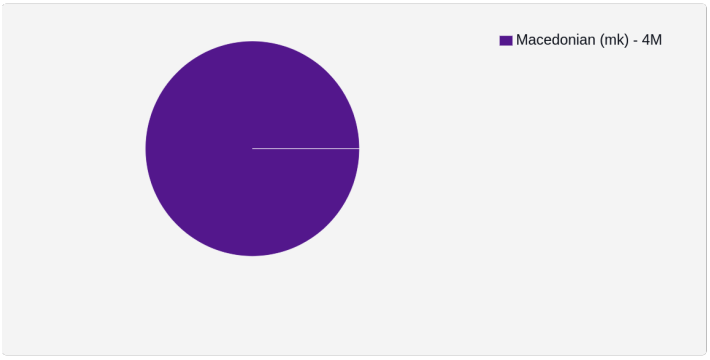


Language Distribution

Source

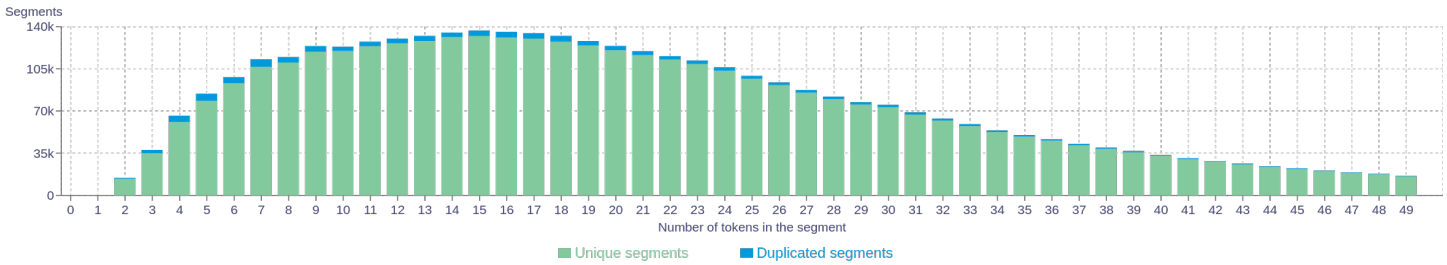


Target



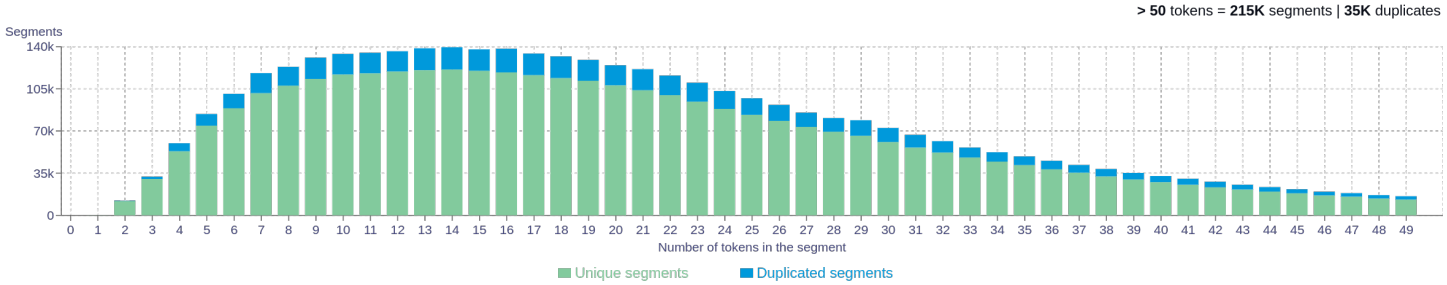
Source segment length distribution by token

<= 49 tokens = **3.6M** segments | **124K** duplicates
> 50 tokens = **231K** segments | **5.4K** duplicates

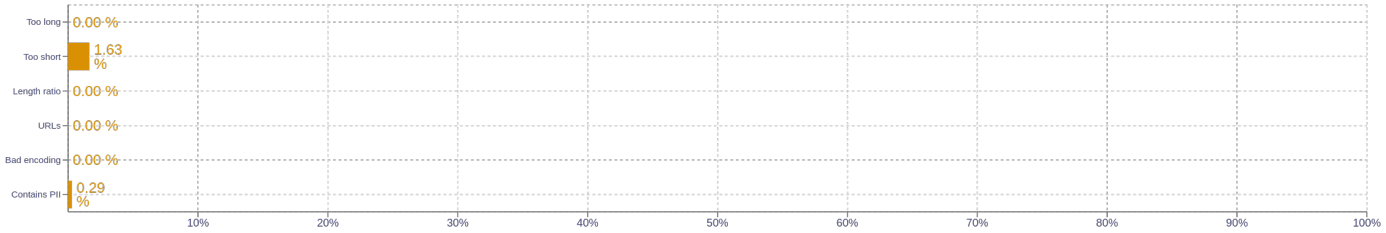


Target segment length distribution by token

<= 49 tokens = **3.3M** segments | **520K** duplicates
> 50 tokens = **215K** segments | **35K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	also | 188309 · one | 173113 · game | 135374 · time | 133126 · use | 122741
2	personal data | 41293 · united states | 20582 · personal information | 11744 · privacy policy | 11476 · email address | 11095
3	republic of macedonia | 15391 · like the game | 13713 · protected from spambots | 8743 · send the link | 8178 · copy and send | 8170
4	address is being protected | 8749 · link to a friend | 8168 · game with the world | 8168 · paste in the html | 5972 · code of your site | 5972
5	friend or all your friends | 8168 · copy and send the link | 8168 · email address is being protected | 8133 · paste in the html code | 5972 · html code of your site | 5972

Target n-grams

Size	n-grams
1	година | 208332 · време | 133335 · вашиот | 113716 · податоци | 105108 · овие | 96782
2	лични податоци | 35127 · уреди извор | 29094 · вашите лични | 21901 · личните податоци | 19451 · вашиот сајт | 19295
3	вашите лични податоци | 18766 · споделување на играта | 8170 · линк на пријател | 8170 · играта со светот | 8170 · заштитена од спамботови | 8037
4	пријател или сите ваши | 8170 · прати линк на пријател | 8170 · копирајте го и прати | 8170 · кодот на вашиот сајт | 5977 · ставете во html кодот | 5973
5	споделување на играта со светот | 8170 · пријател или сите ваши пријатели | 8170 · копирајте го и прати линк | 8170 · е-адреса е заштитена од спамботови | 8036 · кодот и ставете во html | 5973

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