HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-be.tsv	1/23/2025	English (en)	Belarusian (be)

Volumes

Segments	SL tokens	SL characters	SL size
3,140,958	67M	347,852,394	333.32 MB

TL tokens	TL characters	TL size
61M	350,897,169	603.84 MB

Dataset top 10 domains

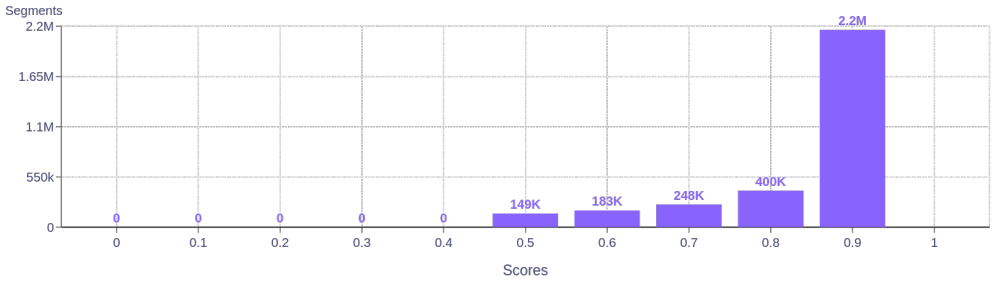
SL domain	Segments	TL domain	Segments
google.com	29.1%	google.com	12.6%
wikipedia.org	9.7%	wikipedia.org	8.6%
spring96.org	4.2%	spring96.org	4.2%
w3eacademy.com	3.2%	w3eacademy.com	3.1%
masterstudies.com	3.0%	sdelalremont.ru	2.1%
dreambook.in.ua	2.4%	dreambook.in.ua	1.9%
itsmygame.org	2.2%	itsmygame.org	1.9%
bachelorstudies.com	2.2%	tostpost.com	1.5%
sdelalremont.ru	2.1%	masterstudies.by	1.5%
academiccourses.com	2.1%	ofunnygames.com	1.4%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	103.7%	com	63.0%
org	26.4%	org	22.6%
ru	6.5%	by	12.6%
by	5.9%	ru	7.9%
net	4.7%	net	3.5%
eu	3.8%	eu	3.2%
in.ua	2.5%	in.ua	2.0%
gov.by	1.7%	gov.by	1.8%
nu	1.5%	nu	1.5%
info	1.3%	info	1.1%

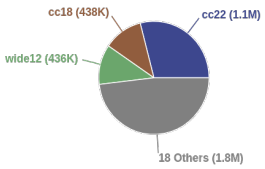
Translation likelihood

≥ 5 = 3.1M segments | **100.0%**
≥ 8 = 2.6M segments | **81.5%**
< 5 = 0 segments | **0.0%**



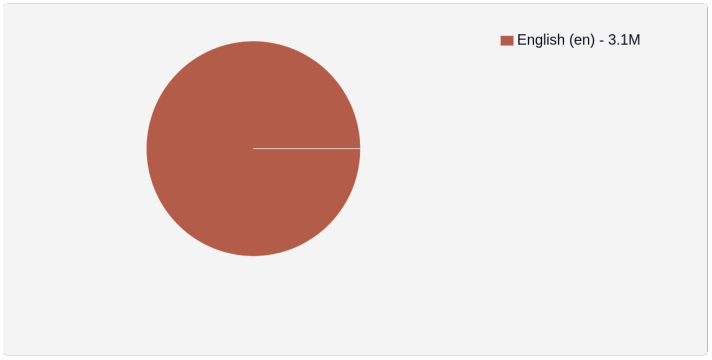
Collections

CC = 58.18%
IA = 41.82%

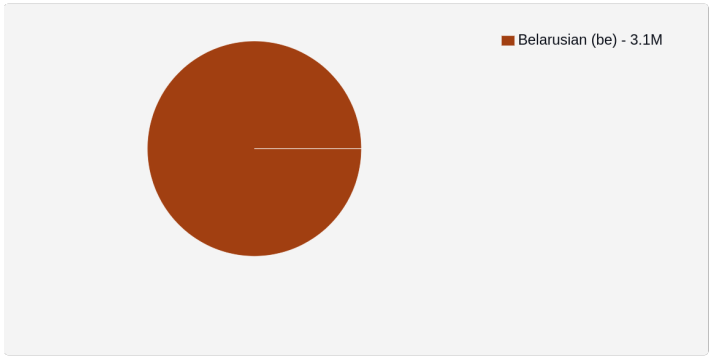


Language Distribution

Source

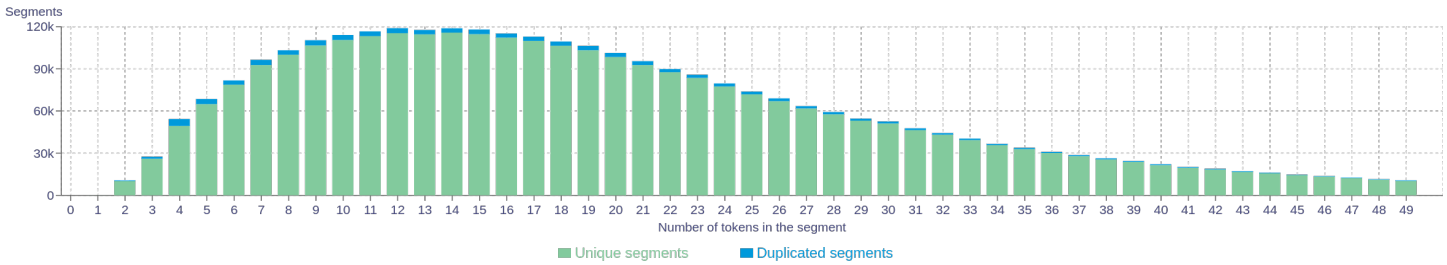


Target



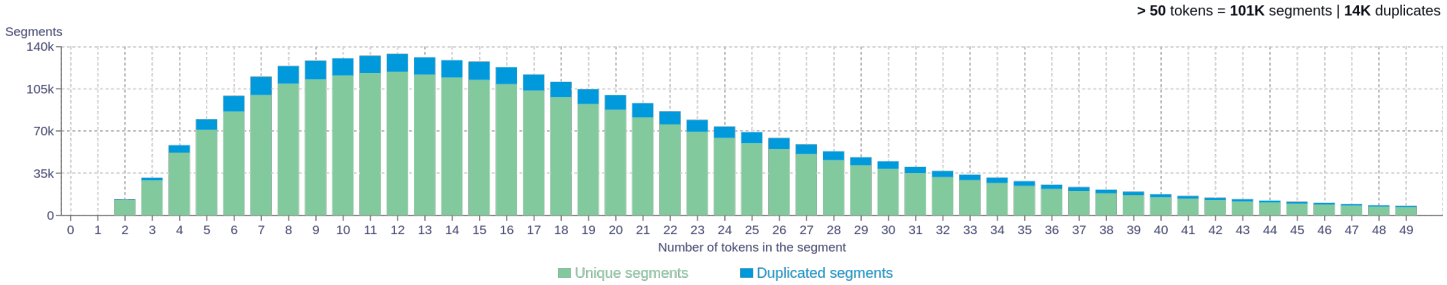
Source segment length distribution by token

<= 49 tokens = **2.9M** segments | **92K** duplicates
> 50 tokens = **144K** segments | **3.5K** duplicates

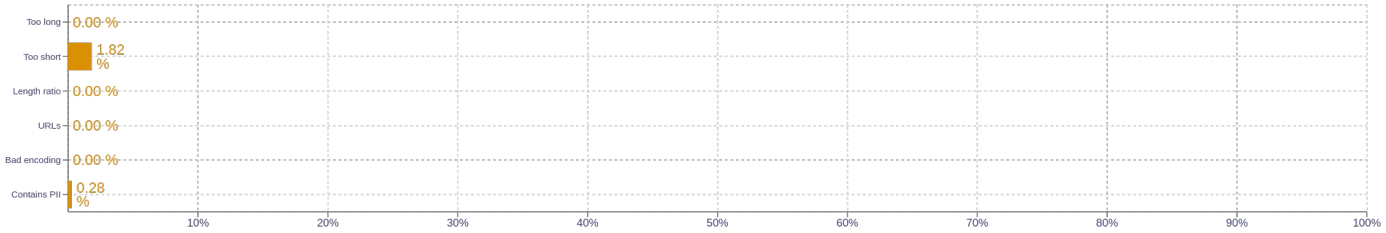


Target segment length distribution by token

<= 49 tokens = **2.7M** segments | **368K** duplicates
> 50 tokens = **101K** segments | **14K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	game | 141741 · one | 125526 · also | 124939 · time | 96951 · people | 85693
2	human rights | 25211 · online game | 10811 · united states | 10071 · personal data | 9435 · best friends | 7839
3	republic of belarus | 18559 · like the game | 8836 · forget to rate | 7705 · share this game | 7630 · rate this game | 7621
4	game with your best | 7604 · play online flash game | 6427 · link to a friend | 5218 · game with the world | 5215 · paste in the html | 5096
5	forget to rate this game | 7608 · game with your best friends | 7604 · friend or all your friends | 5215 · copy and send the link | 5215 · paste in the html code | 5096

Target n-grams

Size	n-grams
1	ў | 972655 · да | 303926 · як | 277276 · гэта | 268772 · калі | 246403
2	можа быць | 31292 · такім чынам | 23018 · рэспублікі беларусь | 20335 · акрамя таго | 16891 · тым ліку | 16560
3	са сваімі лепшымі | 7665 · падзяліцца гэтай гульнёй | 7662 · сваімі лепшымі сябрамі | 7659 · гульнёй са сваімі | 7656 · гэтай гульнёй са | 7654
4	са сваімі лепшымі сябрамі | 7659 · гэтай гульнёй са сваімі | 7653 · падзяліцца гэтай гульнёй са | 7652 · гульнёй са сваімі лепшымі | 7636 · адпраўце спасылку свайму сябру | 5217
5	падзяліцца гэтай гульнёй са сваімі | 7651 · гэтай гульнёй са сваімі лепшымі | 7636 · гульнёй са сваімі лепшымі сябрамі | 7636 · сябру або ўсім сваім сябрам | 5216 · спасылку свайму сябру або ўсім | 5216

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