HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
1,159,869	28M	146,414,761	140.2 MB

TL tokens	TL characters	TL size
23M	157,542,059	151.6 MB

Dataset top 10 domains

SL domain	Segments	TL domain	Segments
educationbro.com	7.4%	lacroixx.com	4.6%
lacroixx.com	4.7%	wikipedia.org	3.6%
wikipedia.org	3.7%	sdelalremont.ru	3.3%
sdelalremont.ru	3.3%	educationbro.com	3.2%
game-game.com	3.1%	game-game.uz	3.1%
yellowpages.uz	2.2%	yellowpages.uz	1.8%
asia-news.com	1.4%	stopzavisimosti.ru	1.4%
stopzavisimosti.ru	1.4%	asia-news.com	1.4%
samdizajner.ru	1.3%	sovetisosveta.ru	1.3%
sovetisosveta.ru	1.3%	samdizajner.ru	1.3%

Dataset top 10 TLDs

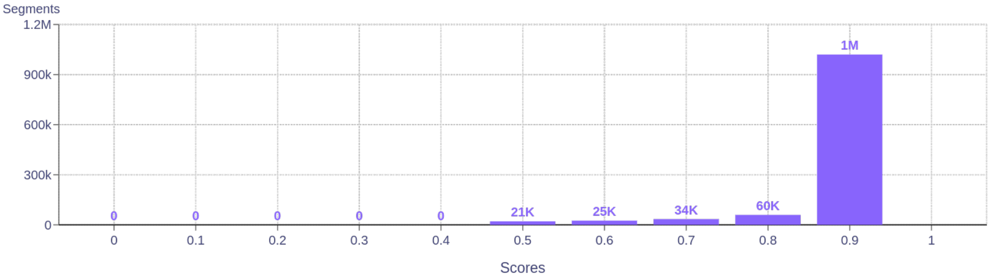
SL domain	Segments	TL domain	Segments
com	65.3%	com	48.7%
uz	15.5%	uz	20.4%
org	12.9%	org	10.9%
net	9.4%	net	8.7%
ru	7.7%	ru	7.7%
eu	1.2%	top	1.0%
top	1.1%	today	0.8%
today	0.9%	info	0.8%
info	0.8%	eu	0.7%
name	0.7%	name	0.7%

Translation likelihood

≥ 5 = 1.2M segments | **100.0%**

≥ 8 = 1.1M segments | **93.1%**

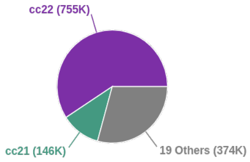
< 5 = 0 segments | **0.0%**



Collections

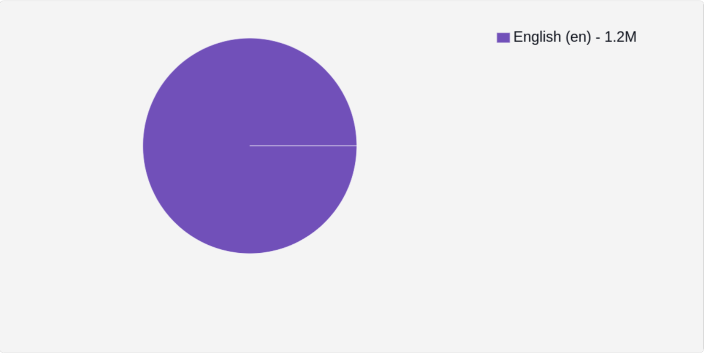
CC = 79.96%

IA = 20.04%

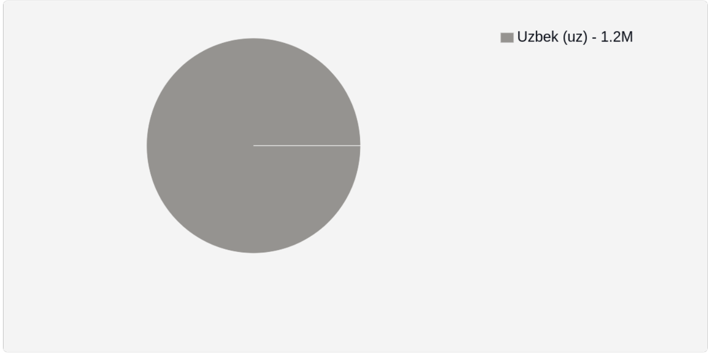


Language Distribution

Source



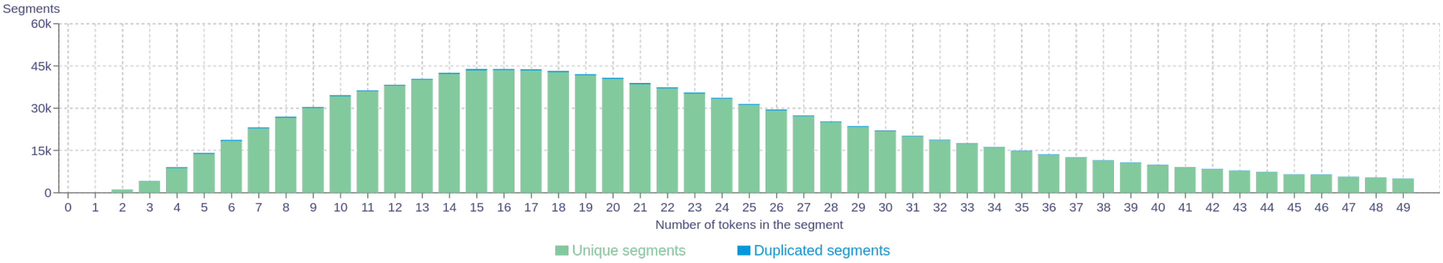
Target



Source segment length distribution by token

<= 49 tokens = **1.1M** segments | **13K** duplicates

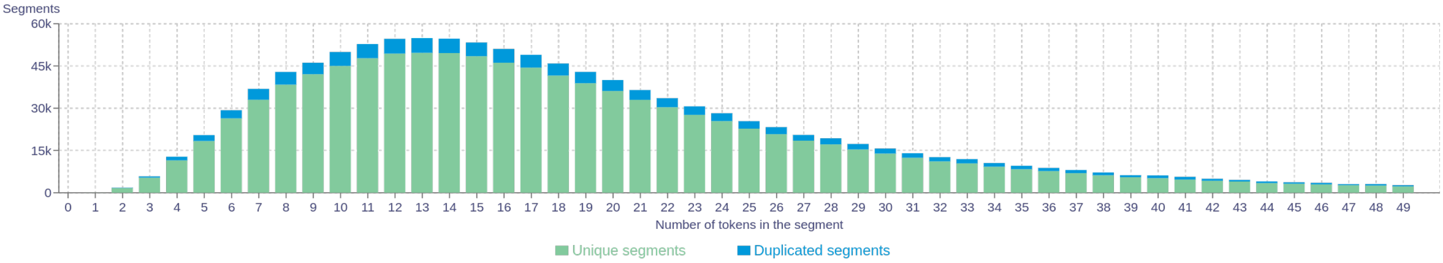
> 50 tokens = **69K** segments | **982** duplicates



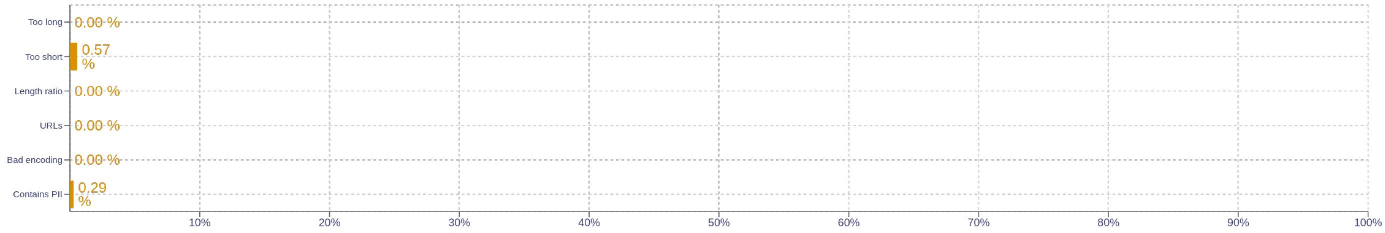
Target segment length distribution by token

<= 49 tokens = **1M** segments | **114K** duplicates

> 50 tokens = **34K** segments | **5.8K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	also | 54375 one | 43774 time | 43686 uzbekistan | 41444 use | 41121
2	personal data | 6197 united states | 4858 personal information | 4144 make sure | 3084 privacy policy | 2706
3	republic of uzbekistan | 15524 around the world | 1813 cabinet of ministers | 1785 terms and conditions | 1225 manufacturers and suppliers | 1165
4	president of the republic | 3963 archived from the original | 1711 one of the best | 1374 one of the leading | 1262 ministers of the republic | 973
5	republic of uzbekistan shavkat mirziyoyev | 792 oliy majlis of the republic | 728 names coloring page printable game | 627 pictures of the names coloring | 617 list of companies with contact | 590

Target n-grams

Size	n-grams
1	bo'ladi | 45644 o'z | 39370 bo'lishi | 39303 yuqori | 37919 bo'lsa | 32512
2	o‘zbekiston respublikasi | 8531 amalga oshirish | 7447 amalga oshiriladi | 6133 ishonch hosil | 5366 elektron pochta | 4502
3	ishonch hosil qiling | 2138 o‘zbekiston respublikasi prezidenti | 1772 asl nusxadan arxivlandi | 1698 o‘zbekiston respublikasi oliy | 1005 o'z ichiga olishi | 970
4	respublikasi prezidenti shavkat mirziyoyev | 675 chiqaruvchilari va etkazib beruvchilardan | 671 o‘zbekiston respublikasi prezidenti shavkat | 640 respublikasi prezidenti islom karimov | 586 chiqaruvchilar etkazib beruvchi ulgurji | 578
5	armatura ishlab chiqaruvchilar etkazib beruvchi | 576 shift balandligi armatura ishlab chiqaruvchilar | 569 balandligi armatura ishlab chiqaruvchilar etkazib | 569 o'g'il bolalar uchun barcha ismlar | 546 o‘zbekiston respublikasi prezidenti shavkat mirziyoyev | 478

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