HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
3,188,231	73M	386,240,961	369.86 MB

TL tokens	TL characters	TL size
62M	387,260,038	426.3 MB

Dataset top 10 domains

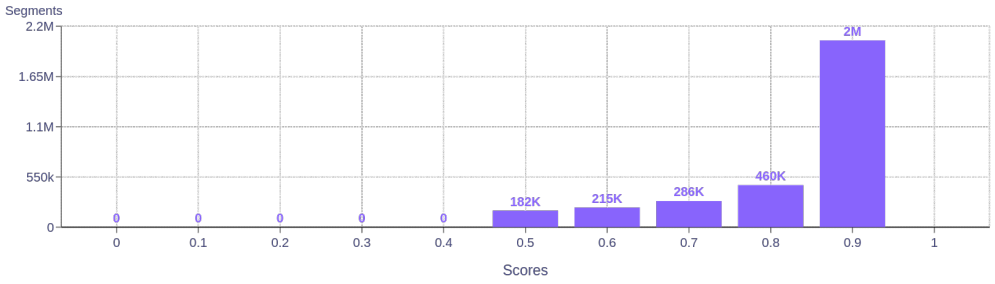
SL domain	Segments	TL domain	Segments
wikipedia.org	3.7%	wikipedia.org	3.5%
softoware.net	3.4%	softoware.net	2.8%
report.az	2.3%	report.az	2.3%
trend.az	2.1%	trend.az	1.9%
educationbro.com	1.9%	president.az	1.6%
itsmygame.org	1.9%	vsaduidoma.com	1.6%
president.az	1.8%	dualjuridik.org	1.5%
vsaduidoma.com	1.6%	itsmygame.org	1.3%
dualjuridik.org	1.6%	jw.org	1.3%
jw.org	1.3%	aliyev-heritage.org	1.3%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	55.7%	com	42.7%
az	22.0%	az	26.6%
org	21.9%	org	19.7%
net	13.4%	net	10.7%
gov.az	4.8%	gov.az	4.9%
co.uk	1.5%	edu.az	1.3%
edu.az	1.3%	info	1.1%
info	1.2%	co.uk	1.1%
eu	1.1%	eu	0.8%
co	0.7%	ru	0.6%

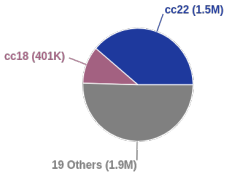
Translation likelihood

≥ 5 = 3.2M segments | **100.0%**
≥ 8 = 2.5M segments | **78.6%**
< 5 = 0 segments | **0.0%**



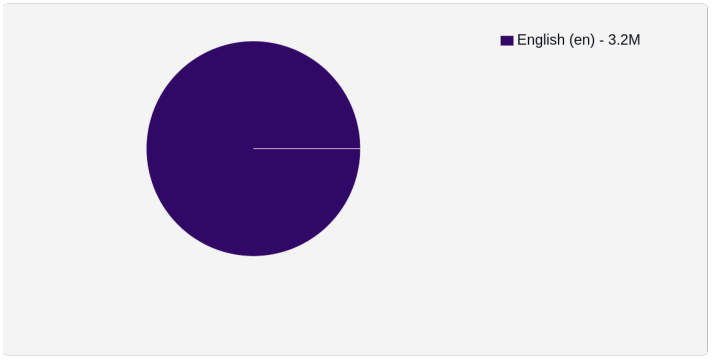
Collections

CC = 67.18%
IA = 32.82%

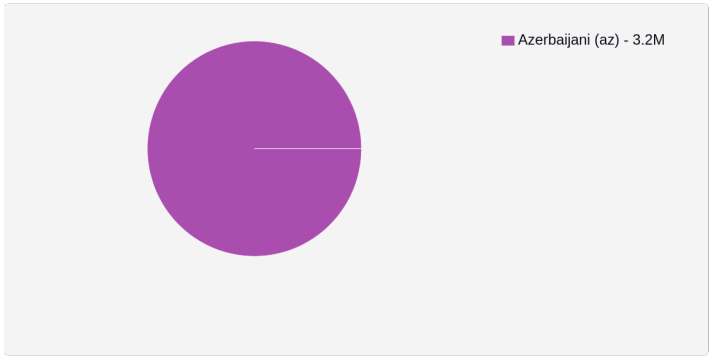


Language Distribution

Source

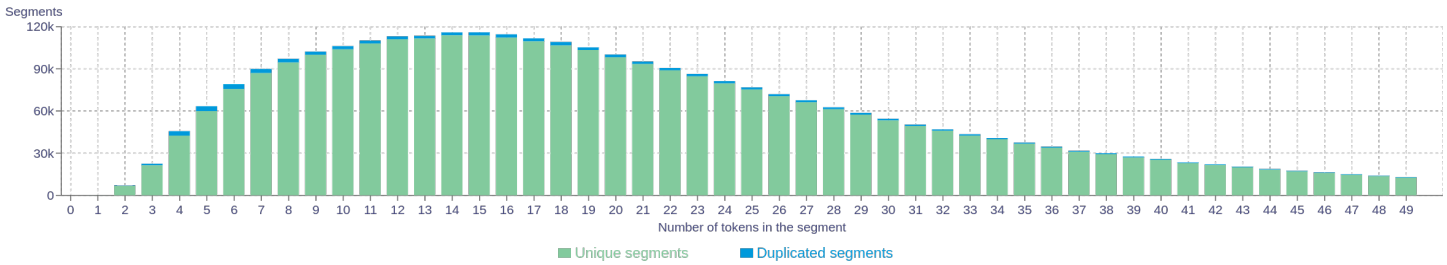


Target



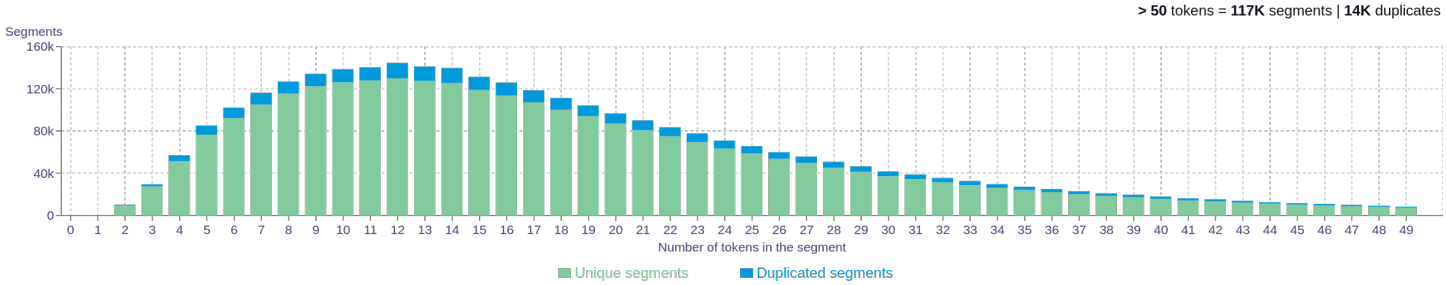
Source segment length distribution by token

<= 49 tokens = **2.9M** segments | **68K** duplicates
> 50 tokens = **190K** segments | **3.6K** duplicates

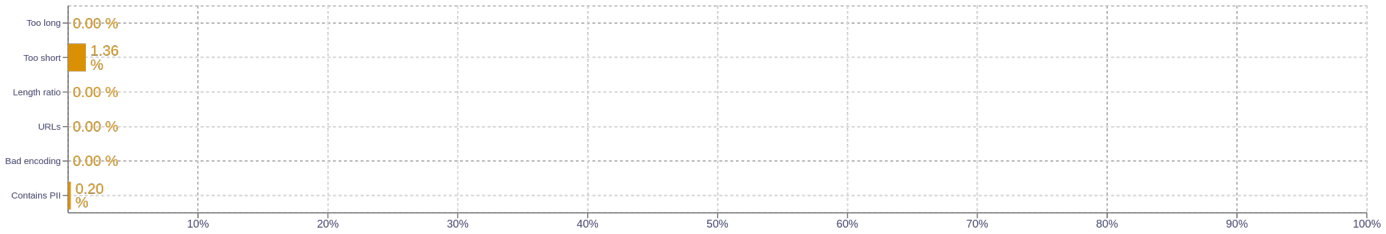


Target segment length distribution by token

<= 49 tokens = **2.8M** segments | **304K** duplicates
> 50 tokens = **117K** segments | **14K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	azerbaijan | 239055 · also | 127582 · one | 119111 · time | 97843 · new | 95567
2	ilham aliyev | 22311 · heydar aliyev | 15729 · united states | 12811 · armed forces | 11305 · personal data | 11260
3	republic of azerbaijan | 42376 · president ilham aliyev | 10577 · azerbaijan ilham aliyev | 6968 · like the game | 4653 · around the world | 4521
4	president of the republic | 13213 · republic of azerbaijan ilham | 5241 · games like the game | 4418 · ministry of foreign affairs | 3146 · archived from the original | 3006
5	republic of azerbaijan ilham aliyev | 5223 · technical characteristics of the game | 2655 · general staff of armed forces | 2114 · armed forces of ukraine says | 2068 · foreign affairs of the republic | 2048

Target n-grams

Size	n-grams
1	azərbaycan | 201024 · böyük | 103937 · tərəfindən | 98791 · yeni | 97624 · digər | 94771
2	azərbaycan respublikasının | 33025 · edə bilərsiniz | 28257 · azərbaycan respublikası | 21615 · eyni zamanda | 19684 · imkan verir | 15814
3	azərbaycan respublikasının prezidenti | 9585 · xarici işlər naziri | 6818 · prezidenti ilham əliyev | 6509 · prezident ilham əliyev | 5764 · respublikasının prezidenti ilham | 5509
4	azərbaycan respublikasının prezidenti ilham | 5486 · respublikasının prezidenti ilham əliyev | 3926 · azərbaycan prezidenti ilham əliyev | 2393 · pulsuz mp3 mahnıları dinləyə | 2176 · mp3 mahnıları dinləyə bilərsiniz | 2176
5	azərbaycan respublikasının prezidenti ilham əliyev | 3910 · pulsuz mp3 mahnıları dinləyə bilərsiniz | 2176 · musiqi pulsuz mp3 mahnıları dinləyə | 2168 · xarici işlər naziri elmar məmmədyarov | 1615 · ukrayna silahlı qüvvələrinin baş qərargahının | 1388

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