HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
1,943,935	45M	243,384,581	233.06 MB

TL tokens	TL characters	TL size
37M	241,361,208	420.98 MB

Dataset top 10 domains

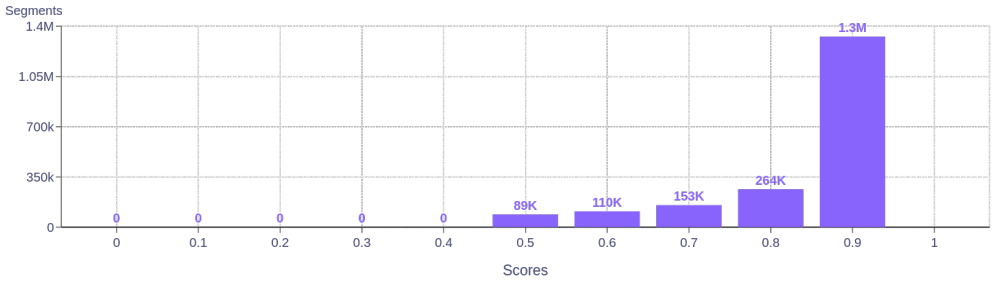
SL domain	Segments	TL domain	Segments
zan.kz	2.9%	strategy2050.kz	2.9%
vsaduidoma.com	2.8%	zan.kz	2.8%
educationbro.com	2.7%	vsaduidoma.com	2.8%
strategy2050.kz	2.7%	egov.kz	2.5%
egov.kz	2.6%	office.com	2.2%
office.com	2.6%	akorda.kz	1.9%
akorda.kz	2.0%	game-game.kz	1.8%
game-game.com	1.8%	forumdaily.com	1.3%
forumdaily.com	1.4%	educationbro.com	1.2%
wikipedia.org	1.4%	wikipedia.org	1.2%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	50.6%	kz	42.8%
kz	37.7%	com	38.1%
org	7.7%	org	6.8%
net	5.5%	net	5.1%
ru	4.6%	ru	4.6%
edu.kz	2.9%	edu.kz	2.8%
gov.kz	1.3%	gov.kz	1.3%
eu	1.2%	info	0.9%
info	1.0%	name	0.8%
name	0.8%	eu	0.8%

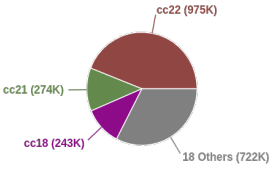
Translation likelihood

≥ 5 = 1.9M segments | **100.0%**
≥ 8 = 1.6M segments | **81.9%**
< 5 = 0 segments | **0.0%**



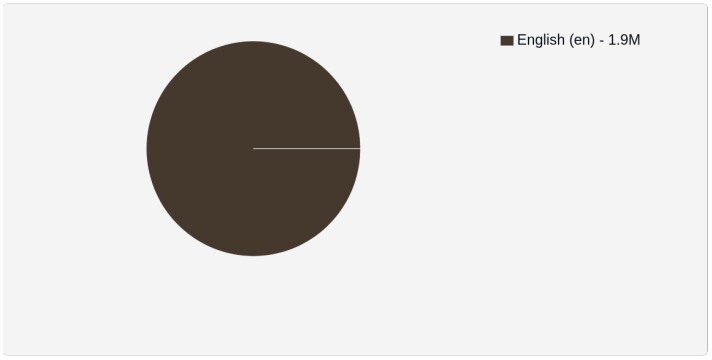
Collections

CC = 71.77%
IA = 28.23%

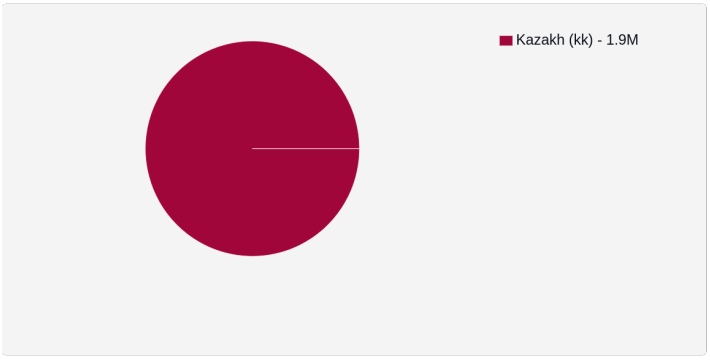


Language Distribution

Source

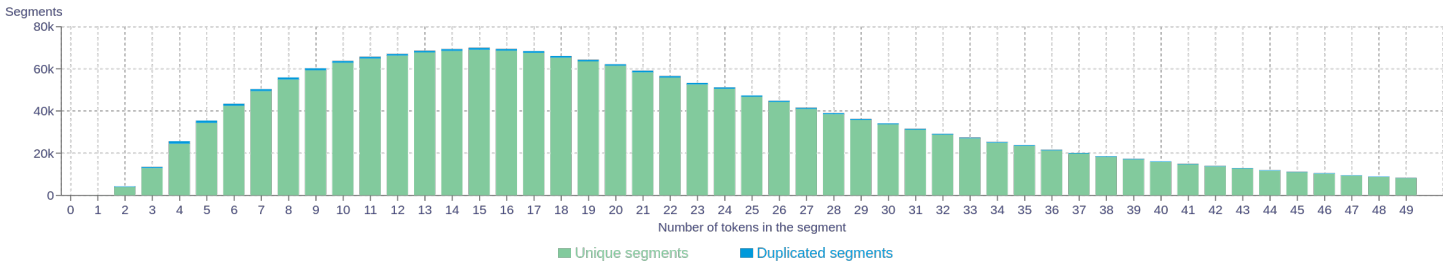


Target



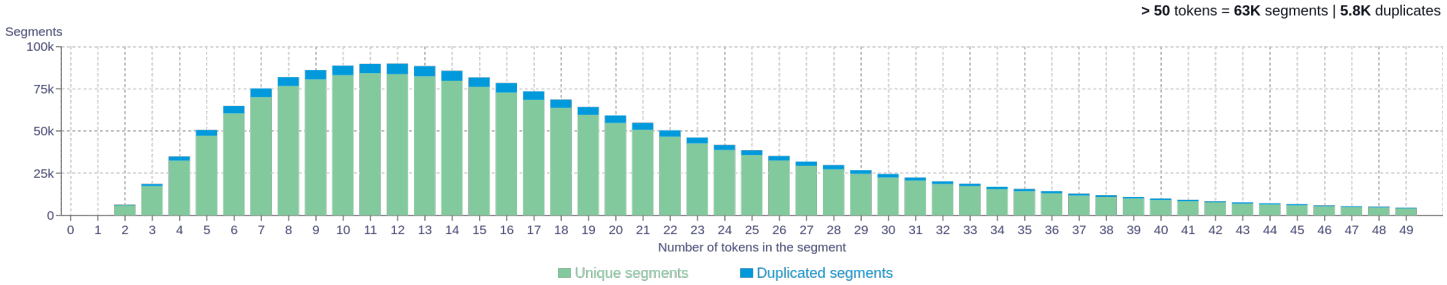
Source segment length distribution by token

<= 49 tokens = **1.8M** segments | **27K** duplicates
> 50 tokens = **124K** segments | **1.7K** duplicates

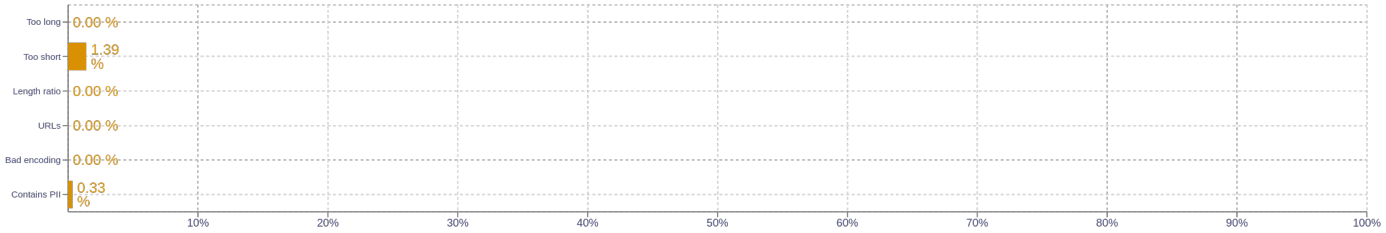


Target segment length distribution by token

<= 49 tokens = **1.7M** segments | **134K** duplicates
> 50 tokens = **63K** segments | **5.8K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	kazakhstan | 159493 · republic | 75383 · also | 71652 · one | 70052 · state | 67320
2	personal data | 12436 · nursultan nazarbayev | 8748 · prime minister | 6120 · personal information | 6009 · united states | 5805
3	republic of kazakhstan | 61557 · head of state | 6352 · peace and blessings | 6095 · president of kazakhstan | 5487 · board of directors | 4910
4	president of the republic | 6557 · legislation of the republic | 5668 · law of the republic | 4768 · republic of kazakhstan dated | 3492 · government of the republic | 3355
5	blessings of allaah be upon | 3582 · peace and blessings of allaah | 3568 · ministry of education and science | 2125 · expiry of ten calendar days | 1789 · upon expiry of ten calendar | 1751

Target n-grams

Size	n-grams
1	және | 729183 · немесе | 168594 · бойынша | 127327 · қазақстан | 103815 · болады | 100981
2	болып табылады | 50238 · қазақстан республикасының | 31898 · кез келген | 21812 · білім беру | 21330 · қазақстан республикасы | 20401
3	және басқа да | 7400 · қандай да бір | 4030 · болып табылады және | 3366 · бірі болып табылады | 3313 · білім және ғылым | 2891
4	игілігі мен сәлемі болсын | 2825 · күн өткен соң қолданысқа | 2529 · өткен соң қолданысқа енгізіледі | 2384 · күнтізбелік он күн өткен | 2124 · алғашқы ресми жарияланған күнінен | 1794
5	он күн өткен соң қолданысқа | 2416 · күн өткен соң қолданысқа енгізіледі | 2217 · алланың игілігі мен сәлемі болсын | 1663 · ресми жарияланған күнінен кейін күнтізбелік | 1341 · күнінен кейін күнтізбелік он күн | 1251

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