HPLT Analytics report

General overview

Corpus	Date	Language
kaz_Cyrl.jsonl.tsv	9/21/2024	Kazakh (kk)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
2,637,363	81,006,479	42,567,471 (52.55 %)	1.8B	11,053,418,553	18.78 GB

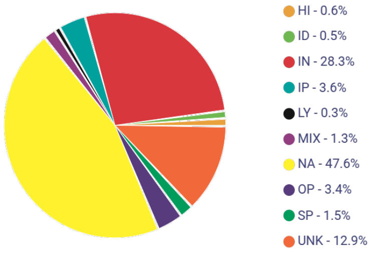
Top 10 domains

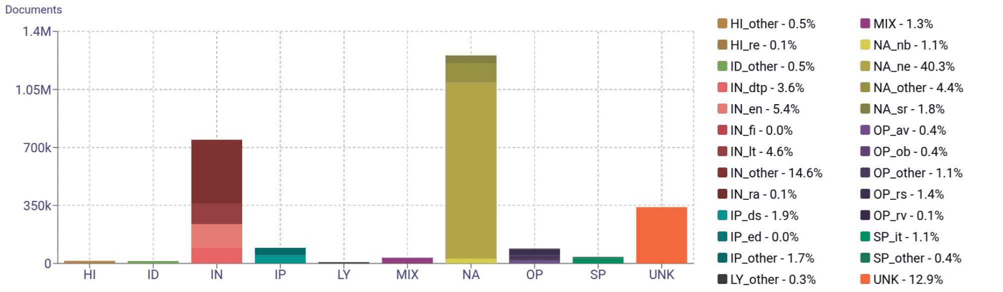
Domain	Docs	% of total
wikipedia.org	154K	5.84%
azattyq.org	143K	5.43%
strategy2050.kz	116K	4.41%
kodeksy-kz.com	62K	2.37%
tengrinews.kz	58K	2.22%
nur.kz	57K	2.17%
inform.kz	50K	1.91%
stud.kz	47K	1.77%
massaget.kz	37K	1.40%
baq.kz	35K	1.32%

Top 10 TLDs

Domain	Docs	% of total
kz	1.8M	67.51%
org	350K	13.25%
com	216K	8.21%
ru	77K	2.93%
gov.kz	64K	2.41%
net	29K	1.10%
info	29K	1.08%
edu.kz	23K	0.86%
mobi	6.4K	0.24%
uz	6K	0.23%

Register labels

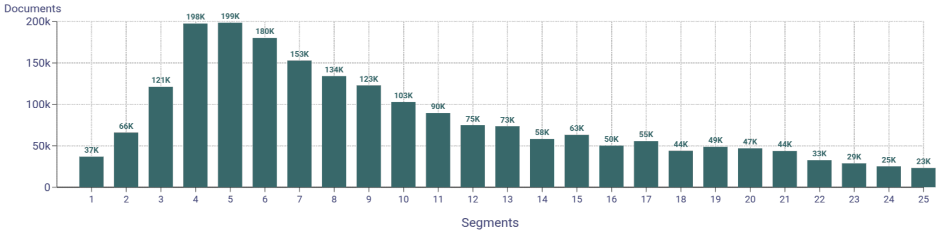




🤖 **MT**:7.4% | 194K Documents

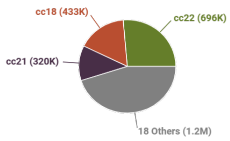
Documents size (in segments)

<= 25 segments **78.51%** (2.1M documents)
> 25 segments **21.49%** (567K documents)



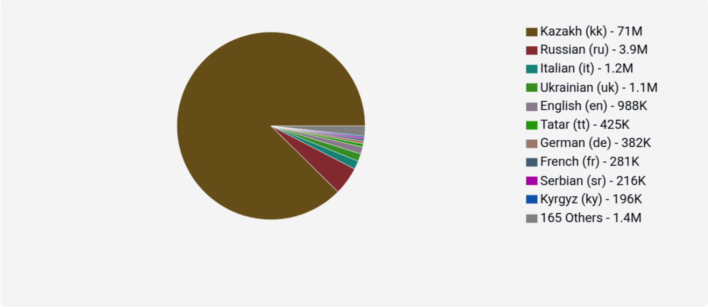
Documents by collection

CC = 68.75%
IA = 31.25%

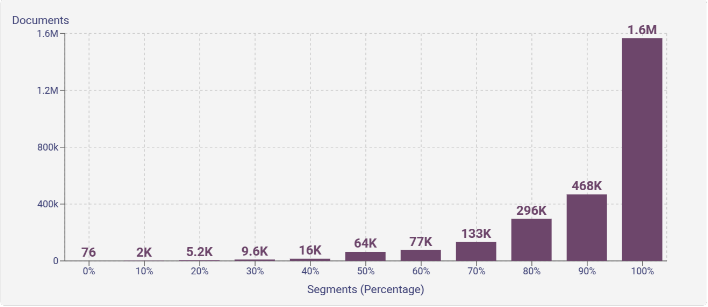


Language Distribution

Number of segments in the Kazakh (kk) corpus

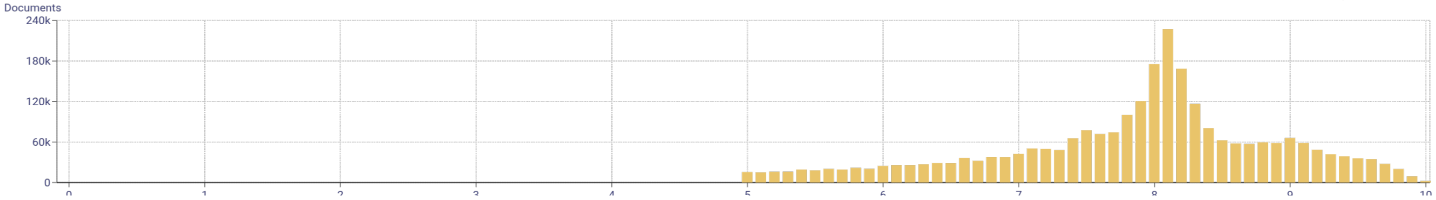


Percentage of segments in Kazakh (kk) inside documents



Distribution of documents by document score

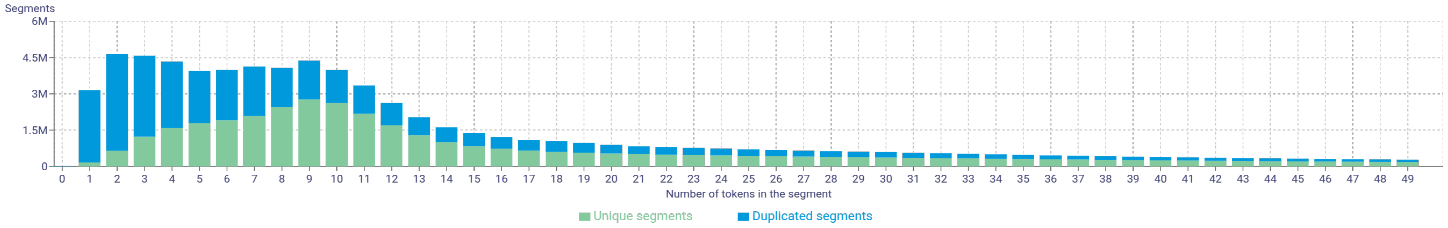
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (2.6M documents)



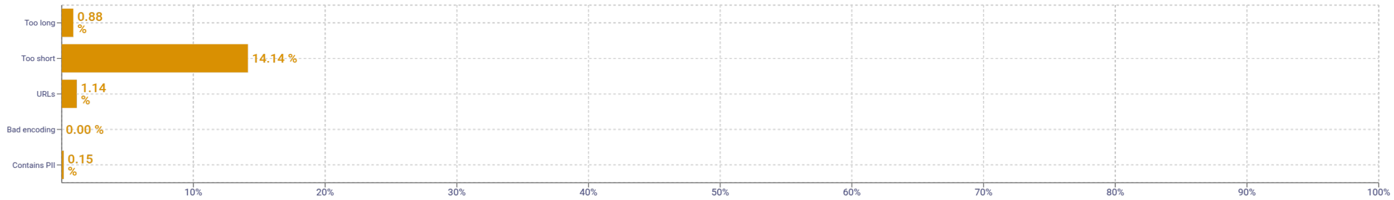
Segment length distribution by token

≤ 49 tokens = **37M** segments | **36M** duplicates

> 50 tokens = **8.9M** segments | **2.9M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	жəне | 17900838 да | 5290698 бір | 5108395 бойынша | 4820816 қазақстан | 4770376
2	қазақстан республикасының | 1407502 болып табылады | 1009956 қазақстан республикасы | 946957 білім беру | 703405 басқа да | 560917
3	өткен соң қолданысқа | 193307 жəне басқа да | 191300 он күн өткен | 166733 күнтізбелік он күн | 146253 ресми жарияланған күнінен | 144430
4	күн өткен соң қолданысқа | 182442 өткен соң қолданысқа енгізіледі | 158912 күнтізбелік он күн өткен | 137602 алғашқы ресми жарияланған күнінен | 134652 жарияланған күнінен кейін күнтізбелік | 93685
5	он күн өткен соң қолданысқа | 166096 күн өткен соң қолданысқа енгізіледі | 150197 ресми жарияланған күнінен кейін күнтізбелік | 93542 күнінен кейін күнтізбелік он күн | 85030 жарияланған күнінен кейін күнтізбелік он | 84740

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

Register labels

Name	Abbr.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av