

General overview

Corpus	Analytics date	Language
ky_1.jsonl.tsv	3/17/2024	Kyrgyz (ky)

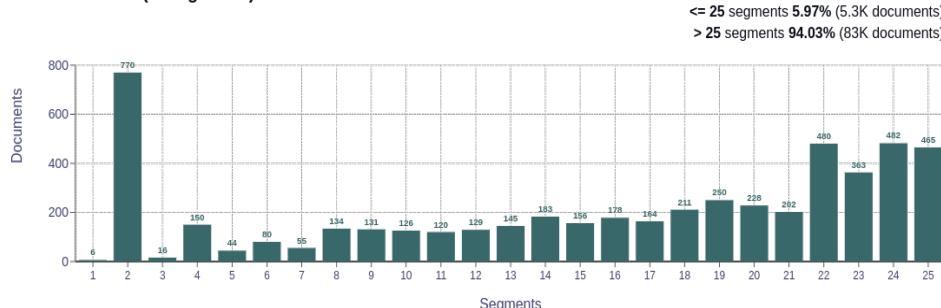
Volumes

Docs	Segments	Unique segments	Tokens	Size
88,322	11,748,340	14,112 (0.12 %)	132M	1.29 GB

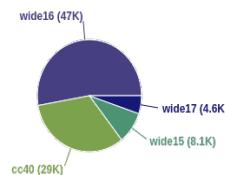
Type-Token Ratio

Kyrgyz (ky)
0.02

Documents size (in segments)

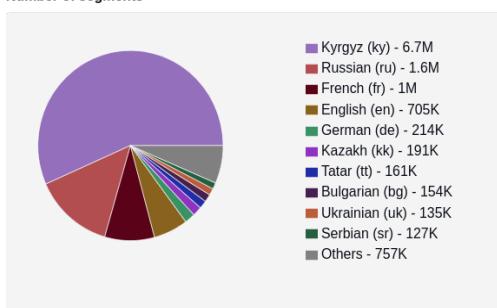


Documents by collection

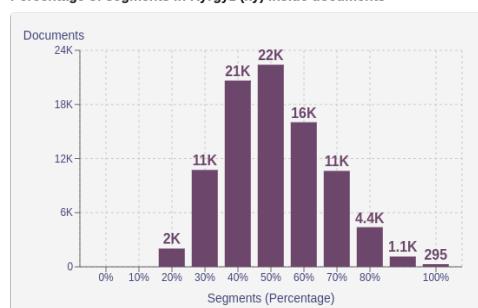


Language Distribution

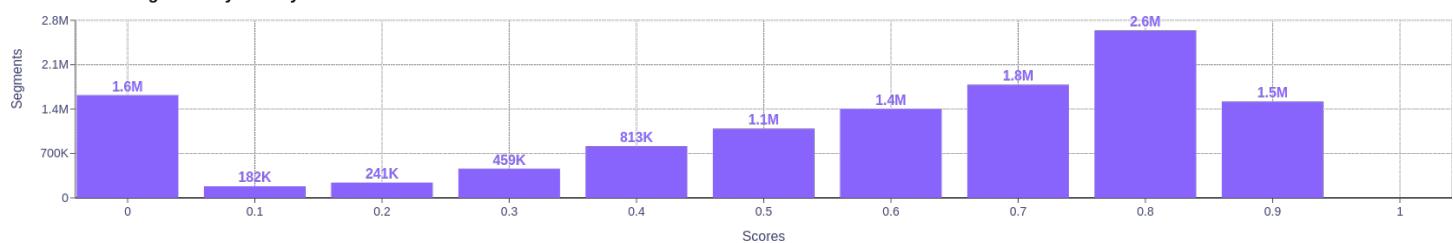
Number of segments



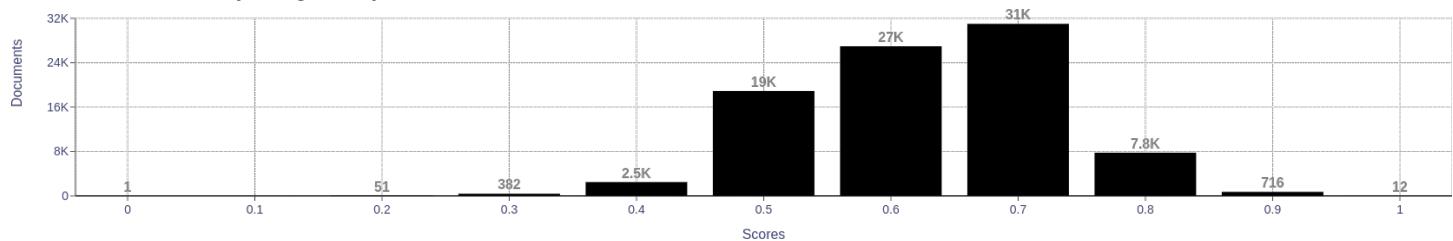
Percentage of segments in Kyrgyz (ky) inside documents



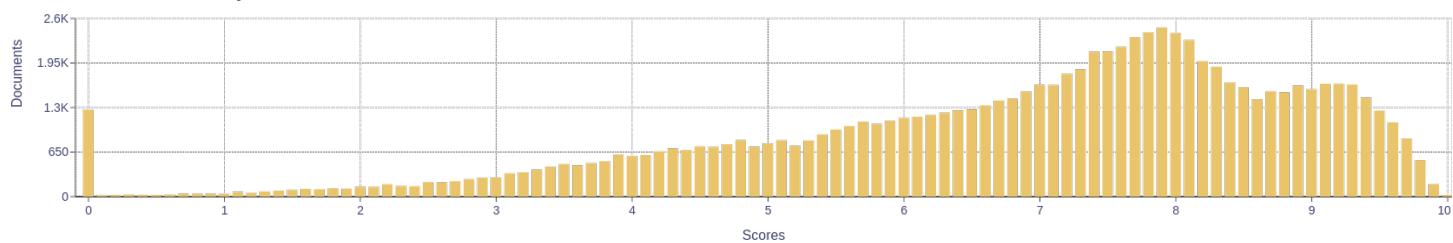
Distribution of segments by fluency score



Distribution of documents by average fluency score



Distribution of documents by document score

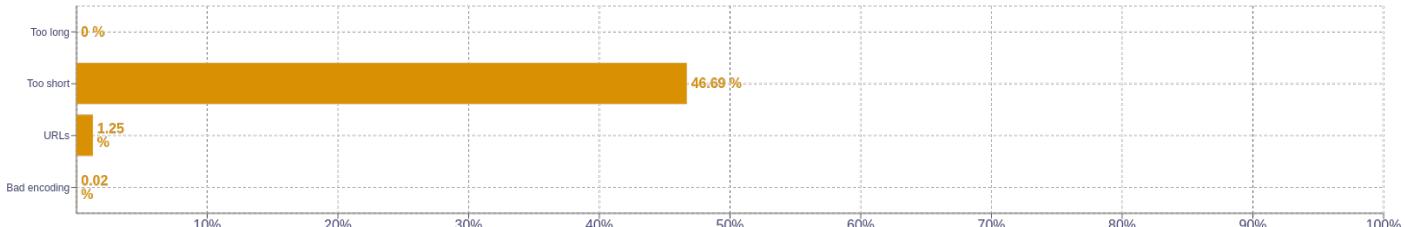


Segment length distribution by token

<= 49 tokens = 2.2M segments | 9M duplicates
 > 50 tokens = 550K segments | 65K duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	бет 382383 kg 234760 кыргызстан 182462 жөнүндө 168895 Республикасынын 150480
2	билим берүү 33791 кат келиптири 25909 курманбек бакиев 25701 аскар акаев 24623 министрліктин жообу 23172
3	министрліктін жаш кадрларынын 23162 жаш кадрларынын жоруктары 23162 унч тарыхый мурас 13710 эркінтоо биримдик пресс 13710 знесай новости иссык 13710
4	министрліктін жаш кадрларынын жоруктары 23162 өкмөтү нур эл де 13710 унч тарыхый мурас обон 13710 эркінтоо биримдик пресс kg 13710 экспресс сааратан diezel айыл 13710
5	өкмөтү нур эл де факт 13710 унч тарыхый мурас обон ош 13710 эркінтоо биримдик пресс kg нур 13710 экспресс сааратан diezel айыл өкмөтү 13710 шамы ош жаңырығы кыргыз рухы 13710

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , 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>