HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Source language	Target language
HPLT.en-bs	10/26/2023	English (en)	Bosnian (bs)

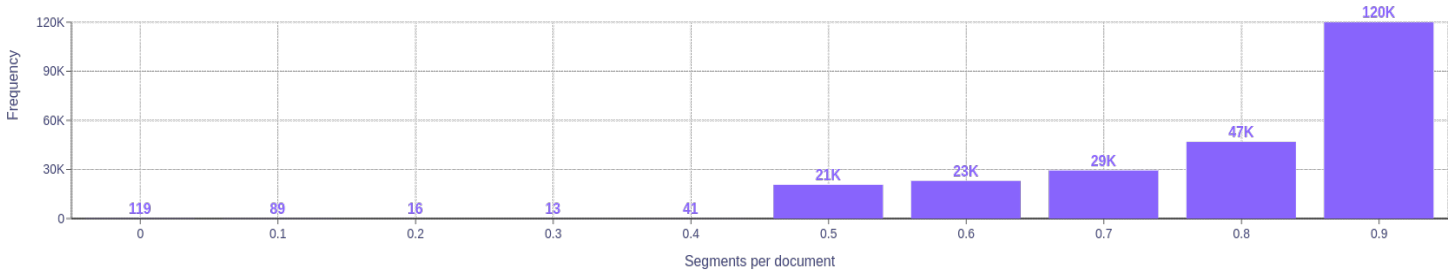
Volumes

Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size
240,015	1,608 (0.67 %)	3.2M	3.2M	16.89 MB	17.76 MB

Type-Token Ratio

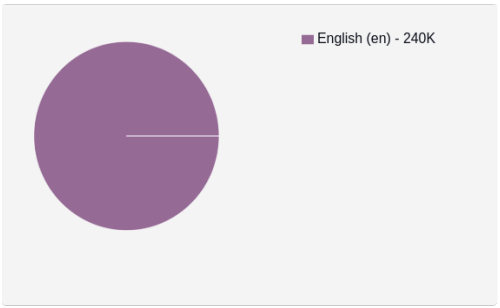
Source	Target
0.04	0.05

Translation likelihood

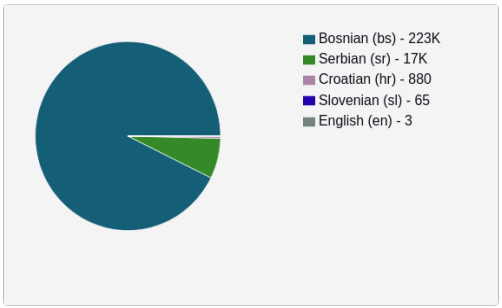


Language Distribution

Source



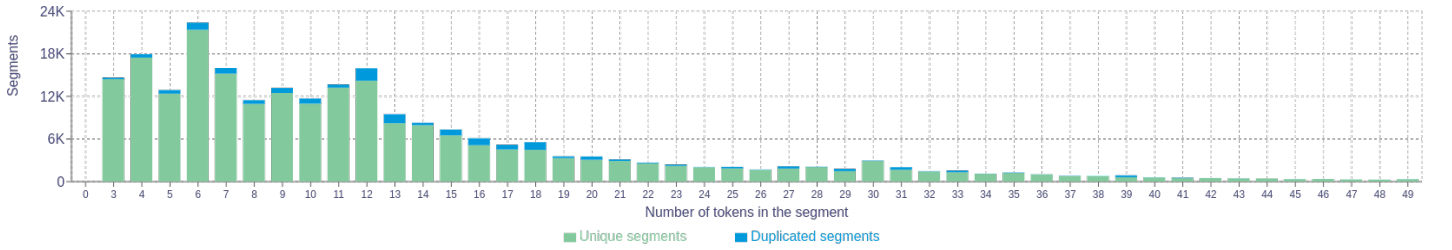
Target



Source segment length distribution by token

<= 49 tokens = **221K** segments | **16K** duplicates

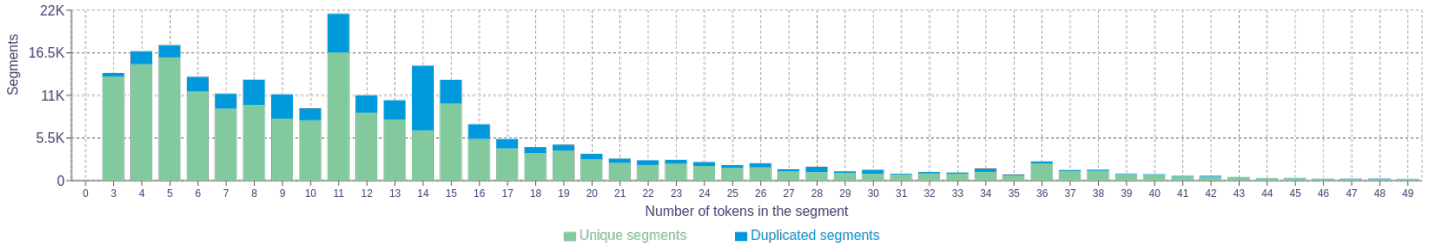
> 50 tokens = **2.8K** segments | **51** duplicates



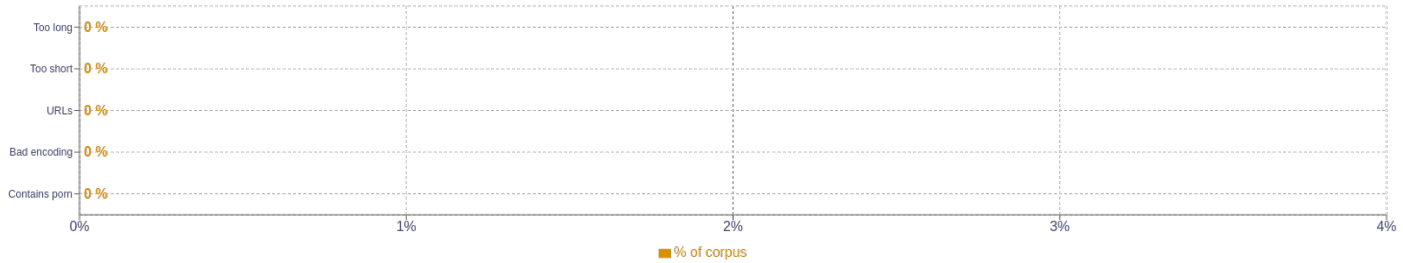
Target segment length distribution by token

<= 49 tokens = **189K** segments | **49K** duplicates

> 50 tokens = **1.9K** segments | **245** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	international | 22926 · climate | 15856 · used | 13578 · united | 12963 · usa | 12768
2	international loads | 6858 · international transportation | 6338 · subtropical climate | 6037 · humid subtropical | 6037 · postal address | 5181
3	humid subtropical climate | 6037 · condition not indicated | 4584 · cities and villages | 4473 · climate humid subtropical | 4023 · köppen climate classification | 4020
4	nearby cities and villages | 4464 · climate humid subtropical climate | 4023 · transport cargoagent.net freight offers | 3609 · cargoagent.net freight offers summary | 3609 · offers summary international loads | 3531
5	transport cargoagent.net freight offers summary | 3609 · freight offers summary international loads | 3531 · cargoagent.net freight offers summary international | 3531 · get full analysis of name | 2635 · exchange- international transportation and spedition | 1687

Target n-grams

Size	n-grams
1	države | 29219 · sjedinjene | 28318 · američke | 27756 · međunarodni | 20832 · prevoz | 14905
2	američke države | 27600 · sjedinjene američke | 27576 · međunarodni transport | 9971 · međunarodni prevoz | 9282 · nije navedeno | 7062
3	sjedinjene američke države | 27566 · vrućim ljetnim mjesecima | 6959 · vlažna suptropska klima | 6959 · klima s vrućim | 6959 · tereti za međunarodni | 6917
4	suptropska klima s vrućim | 6959 · klima s vrućim ljetnim | 6959 · okolnih gradova i sela | 4470 · tereti za međunarodni transport | 3090 · tereti za međunarodni prevoz | 2983
5	vlažna suptropska klima s vrućim | 6959 · suptropska klima s vrućim ljetnim | 6959 · klima s vrućim ljetnim mjesecima | 6959 · berza za međunarodni transport robe | 2056 · kamiona za međunarodni prevoz robe | 1630

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).

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.

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