HPLT Analytics report

HPLTAnalytics

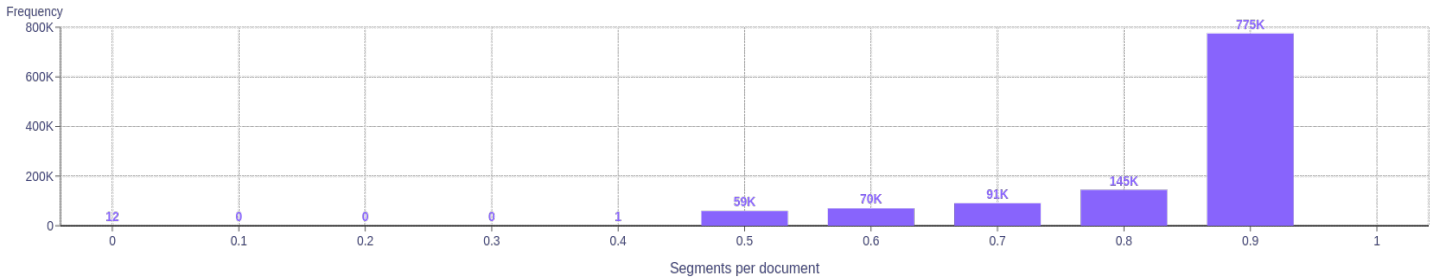
General overview

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

Volumes

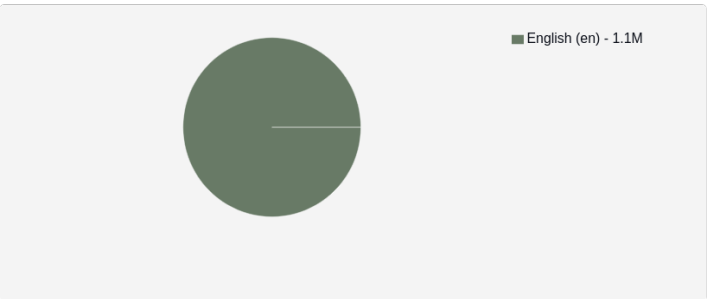
Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size	Src characters	Trg characters
1,139,063	1,139,052 (100.00 %)	21M	21M	109.74 MB	200.95 MB		

Translation likelihood

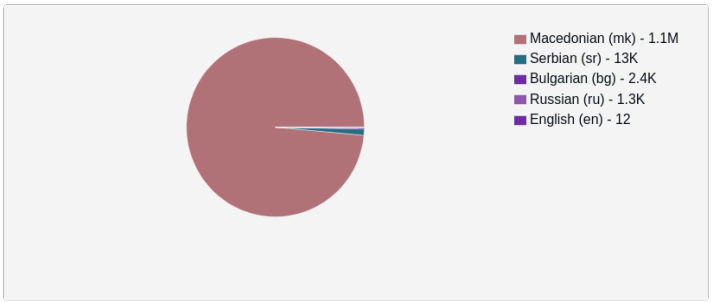


Language Distribution

Source



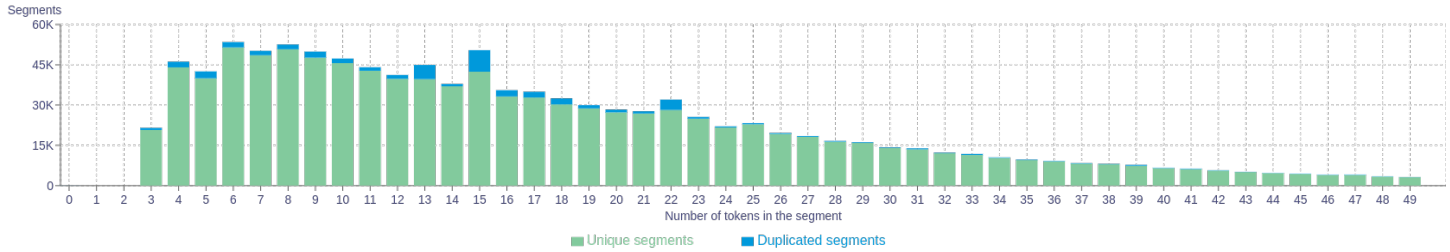
Target



Source segment length distribution by token

<= 49 tokens = **1M** segments | **52K** duplicates

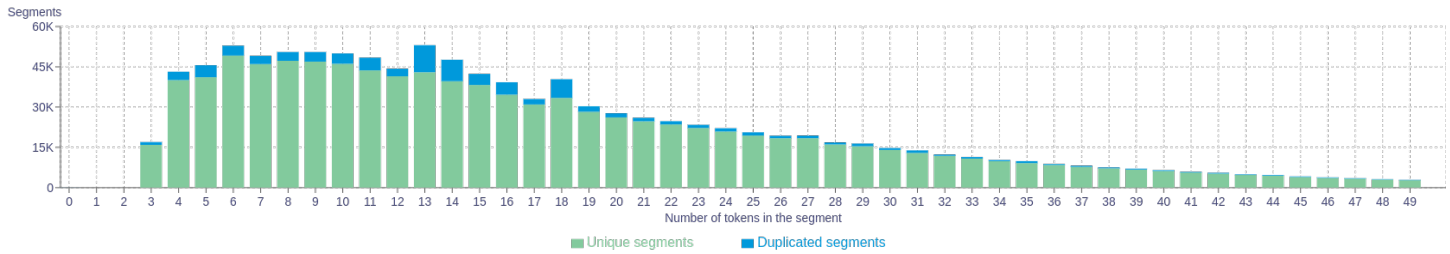
> 50 tokens = **40K** segments | **629** duplicates



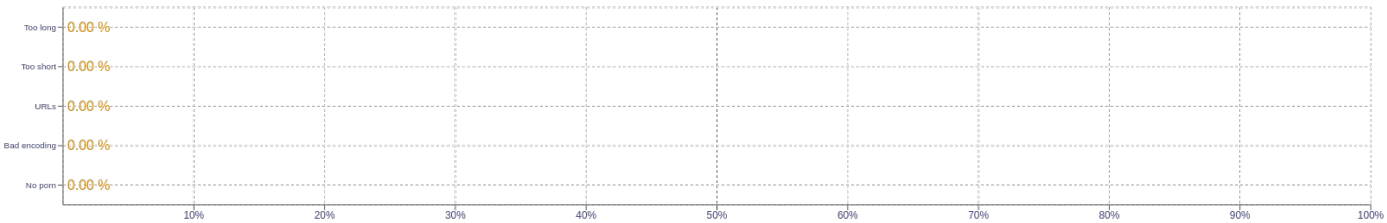
Target segment length distribution by token

<= 49 tokens = **1M** segments | **91K** duplicates

> 50 tokens = **36K** segments | **1.5K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	new | 54656 · one | 31981 · also | 31829 · offers | 30943 · first | 30210
2	new offers | 23628 · immeasurably grateful | 9446 · working hours | 8582 · search term | 8172 · important details | 8083
3	sure to appreciate | 9446 · appreciate this gesture | 9446 · clicking the button | 7271 · republic of macedonia | 6394 · call new offers | 5597
4	gesture and be immeasurably | 9446 · position on the main | 7261 · machines offered at machineseeker | 5140 · visit epoch and segpay | 4927 · price not including vat | 4289
5	sure to appreciate this gesture | 9446 · gesture and be immeasurably grateful | 9446 · position on the main page | 7261 · first position on the main | 7261 · please visit epoch and segpay | 4927

Target n-grams

Size	n-grams
1	година | 49110 · нови | 32875 · македонија | 29512 · понуди | 27777 · време | 26443
2	нови понуди | 23657 · оцени гестот | 9457 · бескрајно благодарна | 9457 · работни часови | 8444 · важни детали | 8099
3	зборот за пребарување | 8171 · позиција на главната | 7262 · кликнеш на копчето | 7261 · резултати од пребарувањето | 6900 · категорија на фирмата | 6206
4	дефинитивно ќе го оцени | 9457 · позиција на главната страница | 7262 · првата позиција на главната | 7261 · машини понудени на machineseeker | 5278 · овластен застапник за продажба | 4928
5	дефинитивно ќе го оцени гестот | 9457 · првата позиција на главната страница | 7261 · посетете ги epoch и segpay | 4928 · нашиот овластен застапник за продажба | 4928 · фиксна цена не вклучувајќи ддв | 2726

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