

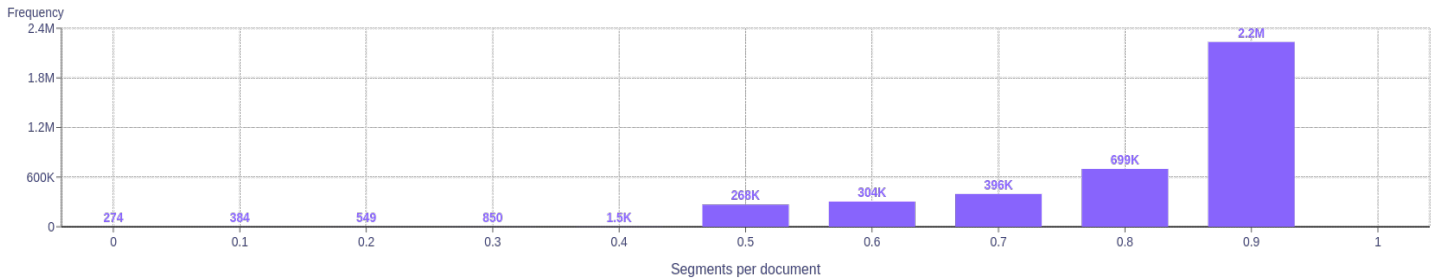
General overview

Corpus	Analytics date	Source language	Target language
HPLT.en-sr	10/24/2023	English (en)	Serbian (sr)

Volumes

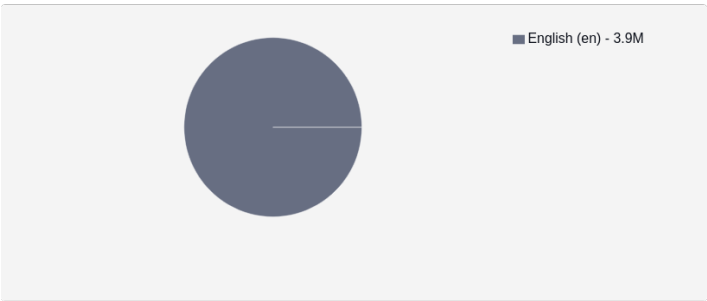
Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size	Src characters	Trg characters
3,904,423	3,904,384 (100.00 %)	63M	61M	325.13 MB	343.9 MB		

Translation likelihood

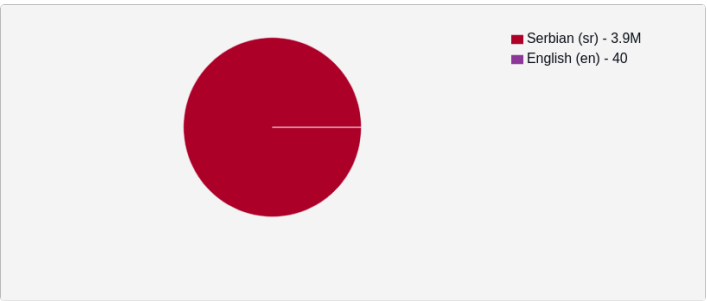


Language Distribution

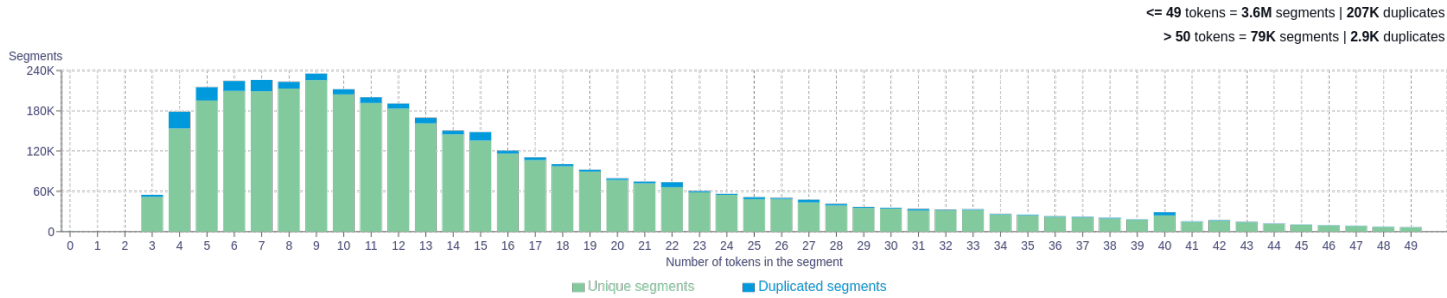
Source



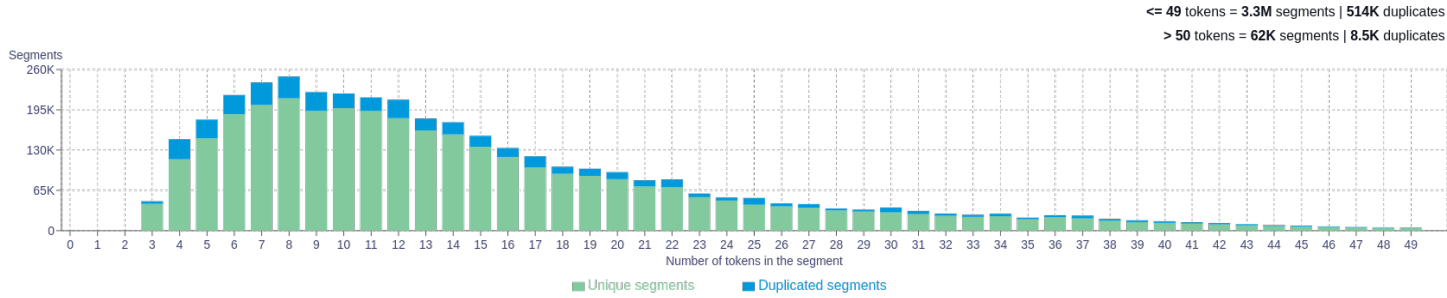
Target



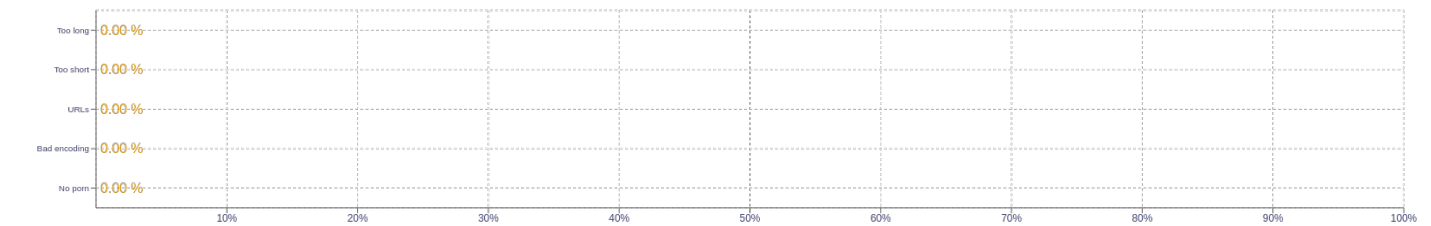
Source segment length distribution by token



Target segment length distribution by token



Segment pair noise distribution



Source n-grams

Size	n-grams
1	hotels   429345 hotel   314216 best   213308 near   168628 new   142956
2	hotels near   98205 traveller rating   58694 see availability   39633 best hotels   39523 personal data   36686
3	opens in new   33796 reservations with confidence   33545 proud to partner   33275 tripadvisor is proud   33272 reserves the right   27892
4	opens in new window   33780 temporarily hold an amount   25831 right to temporarily hold   25831 hold an amount prior   25831 amount prior to arrival   25831
5	tripadvisor is proud to partner   33272 reserves the right to temporarily   25835 temporarily hold an amount prior   25831 cards and reserves the right   25830 accepts these cards and reserves   25830

Target n-grams

Size	n-grams
1	u   1727729 i   1605161 za   982618 na   837510 da   804972
2	u blizini   203397 hoteli u   179046 u mestu   122558 u gradu   82214 da li   75543
3	hoteli u blizini   89365 u blizini znamenitosti   70223 porodični hoteli u   37051 u novom prozoru   33818 se u novom   33683
4	hoteli u blizini znamenitosti   63477 se u novom prozoru   33655 otvara se u novom   33655 pravite bezbedne rezervacije za   33642 možete da pravite bezbedne   33642
5	otvara se u novom prozoru   33655 možete da pravite bezbedne rezervacije   33642 jer možete da pravite bezbedne   33642 da pravite bezbedne rezervacije za   33642 tripadvisor je ponosan na partnerstvo   33273

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 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>, <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>