HPLT Analytics report

HPLTAnalytics

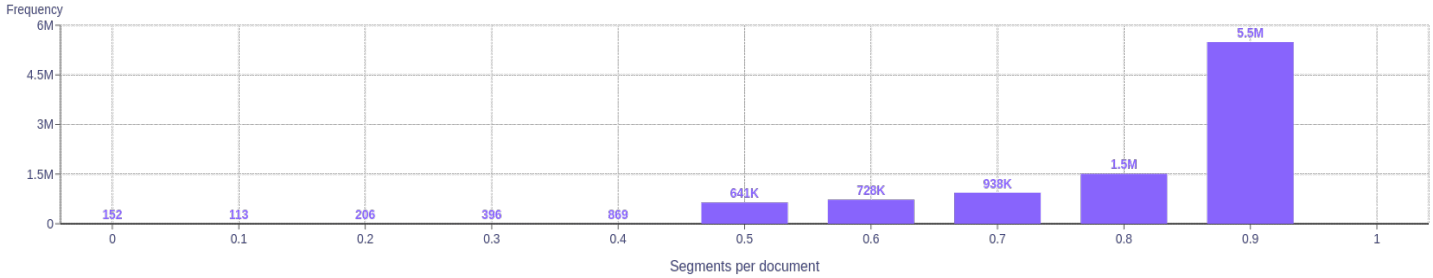
General overview

Corpus	Analytics date	Source language	Target language
HPLT.en-hr	10/27/2023	English (en)	Croatian (hr)

Volumes

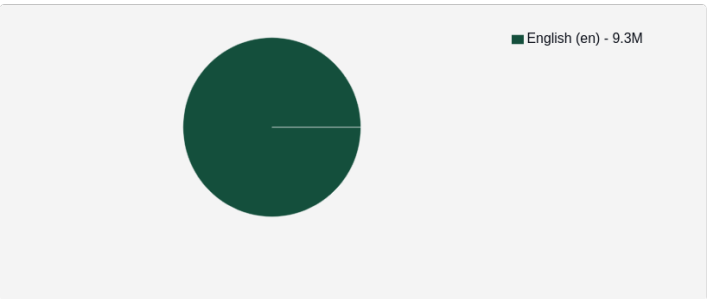
Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size	Src characters	Trg characters
9,310,369	9,310,276 (100.00 %)	162M	152M	815.9 MB	852.71 MB		

Translation likelihood

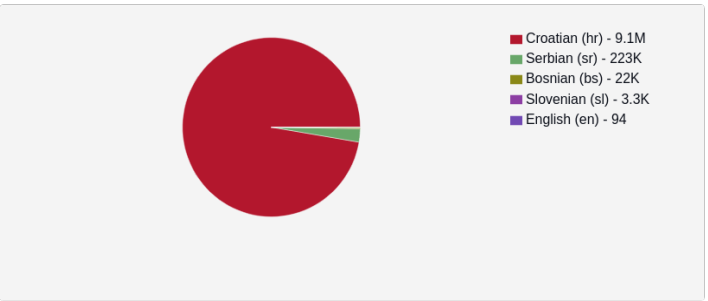


Language Distribution

Source



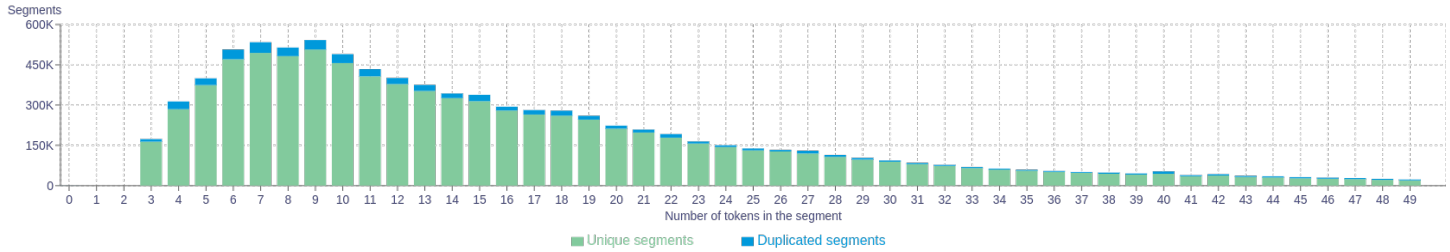
Target



Source segment length distribution by token

<= 49 tokens = **8.4M** segments | **587K** duplicates

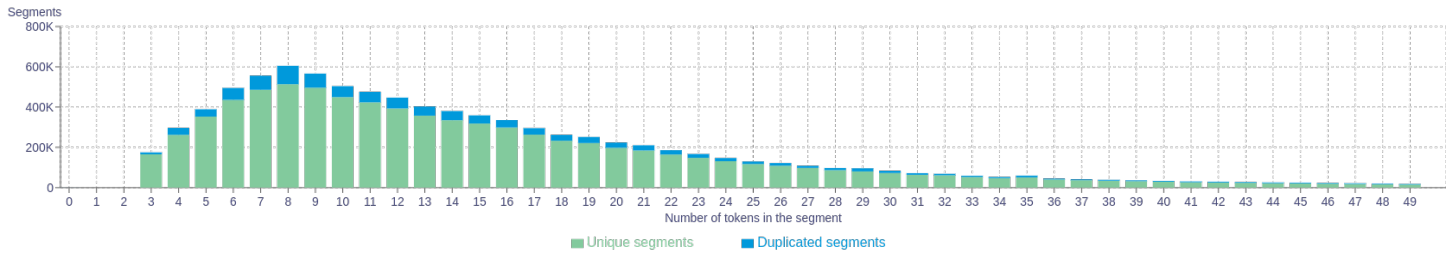
> 50 tokens = **281K** segments | **21K** duplicates



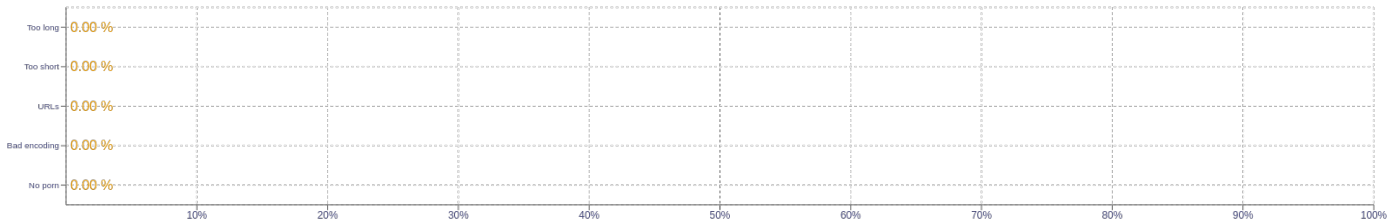
Target segment length distribution by token

<= 49 tokens = **8M** segments | **1.1M** duplicates

> 50 tokens = **206K** segments | **27K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	hotel | 370035 · weather | 326122 · new | 306112 · free | 288669 · used | 285819
2	personal data | 74918 · local time | 71682 · best prices | 71369 · buy used | 68441 · operating hours | 68097
3	year of manufacture | 153819 · prices from either | 58907 · either machinery dealers | 58907 · dealers or private | 58907 · freemeteo hotel bookings | 44971
4	prices from either machinery | 58907 · machinery dealers or private | 58907 · dealers or private sellers | 58907 · best prices from either | 58907 · clouds freemeteo hotel bookings | 44597
5	prices from either machinery dealers | 58907 · machinery dealers or private sellers | 58907 · either machinery dealers or private | 58907 · best prices from either machinery | 58907 · snow clouds freemeteo hotel bookings | 44570

Target n-grams

Size	n-grams
1	vrijeme | 350121 · više | 286272 · hotel | 269926 · godina | 223323 · može | 182135
2	godina proizvodnje | 156248 · zračna luka | 90361 · vremenska prognoza | 86402 · engleskom jeziku | 85952 · radni sati | 80648
3	dugoročna vremenska prognoza | 66747 · freemeteo rezervacije hotela | 44967 · više o smještajnom | 44879 · oblaci freemeteo rezervacije | 44574 · snijeg oblaci freemeteo | 44537
4	strojeva bilo od privatnih | 51508 · cijenama bilo od distributera | 51508 · više o smještajnom objektu | 44879 · oblaci freemeteo rezervacije hotela | 44574 · snijeg oblaci freemeteo rezervacije | 44537
5	strojeva bilo od privatnih prodavača | 51508 · najboljim cijenama bilo od distributera | 51508 · distributera strojeva bilo od privatnih | 51508 · cijenama bilo od distributera strojeva | 51508 · snijeg oblaci freemeteo rezervacije hotela | 44537

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