HPLT Analytics report

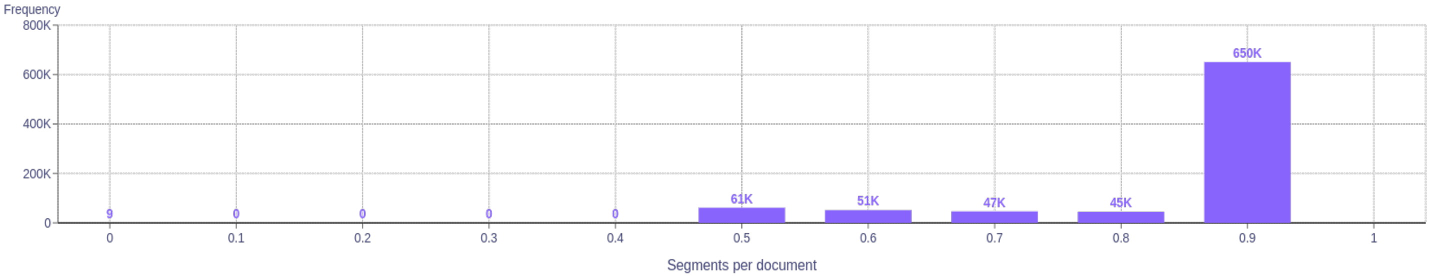
General overview

Corpus	Analytics date	Source language	Target language
HPLT.en-mt	10/23/2023	English (en)	Maltese (mt)

Volumes

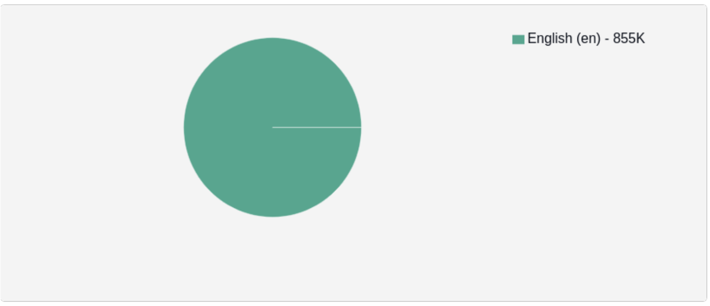
Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size	Src characters	Trg characters
854,829	854,821 (100.00 %)	22M	21M	112.91 MB	128.36 MB		

Translation likelihood

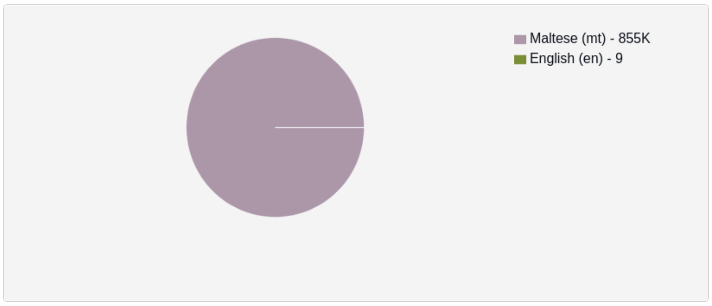


Language Distribution

Source



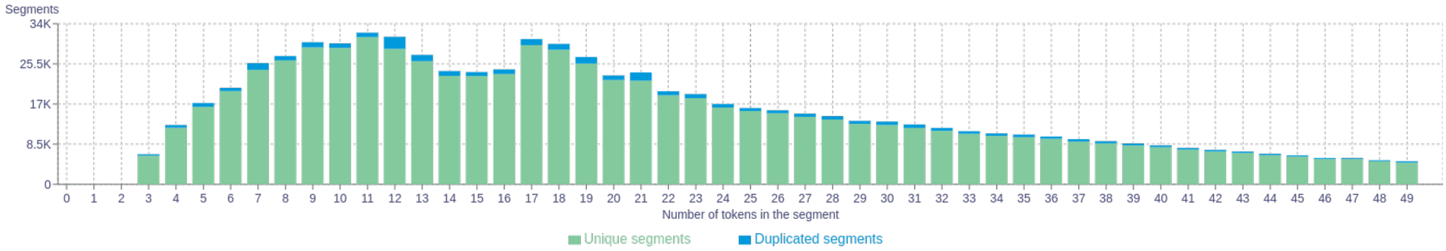
Target



Source segment length distribution by token

<= 49 tokens = **734K** segments | **35K** duplicates

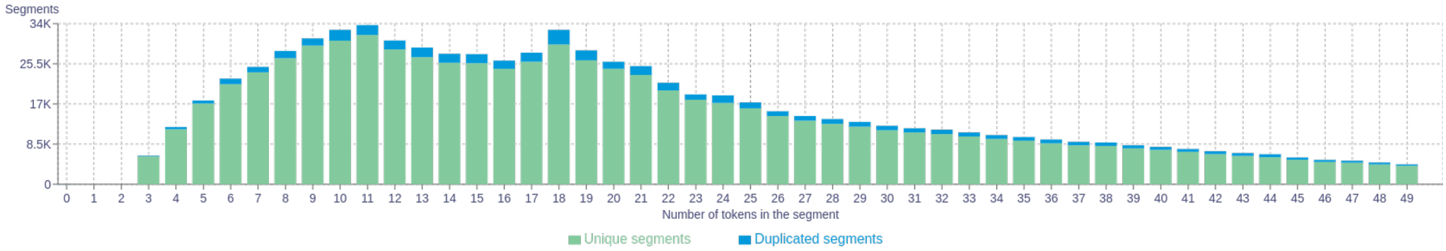
> 50 tokens = **86K** segments | **3.2K** duplicates



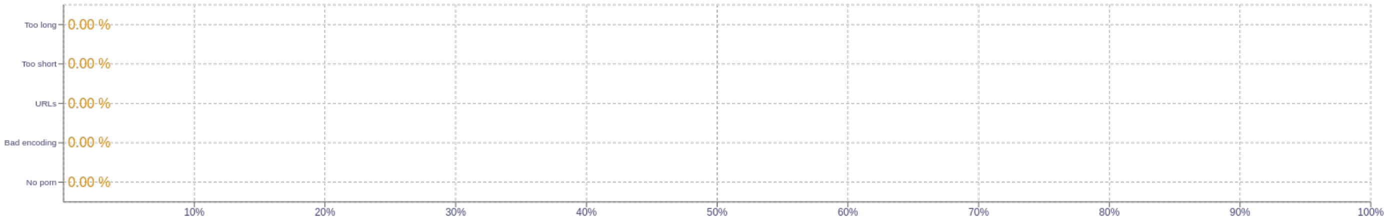
Target segment length distribution by token

<= 49 tokens = **732K** segments | **53K** duplicates

> 50 tokens = **70K** segments | **5.7K** duplicates



Segment pair noise distribution



Source n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>article | 63918
european | 59923
member | 53497
eu | 53095
commission | 46273</td></tr>
<tr><td>2</td><td>archive entry | 34547
member states | 30631
member state | 19597
european parliament | 15416
offer archive | 15234</td></tr>
<tr><td>3</td><td>compare every offer | 15260
offer archive entry | 15234
add to cart | 11166
call new offers | 7569
accordance with article | 5941</td></tr>
<tr><td>4</td><td>compare every offer archive | 15234
referred to in article | 7423
referred to in paragraph | 3807
price not including vat | 2509
starting price not including | 2365</td></tr>
<tr><td>5</td><td>compare every offer archive entry | 15234
parliament and of the council | 8018
starting price not including vat | 2365
ec of the european parliament | 2089
rated hosts who are committed | 1738</td></tr>
</table>

Target n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>ta | 809041
u | 527492
li | 472797
jew | 149785
b | 139257</td></tr>
<tr><td>2</td><td>b 'mod | 37841
id-dħul ta | 35589
u li | 16234
kull offerta | 15283
kotba kollha | 15268</td></tr>
<tr><td>3</td><td>qabbel kull offerta | 15266
offerta id-dħul ta | 15265
kull offerta id-dħul | 15265
b 'mod partikolari | 12250
offerti ġodda għal | 8432</td></tr>
<tr><td>4</td><td>qabbel kull offerta id-dħul | 15265
kull offerta id-dħul ta | 15265
tal-parlament ewropew u tal-kunsill | 8140
is-sejħa offerti ġodda għal | 7556
tat-tluq mhux inkluża l-vat | 2410</td></tr>
<tr><td>5</td><td>qabbel kull offerta id-dħul ta | 15265
prezz tat-tluq mhux inkluża l-vat | 2410
ke tal-parlament ewropew u tal-kunsill | 1971
talba is-sejħa offerti ġodda għal | 1889
huma impenjati li jipprovdu l-aqwa | 1738</td></tr>
</table>

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