

General overview

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

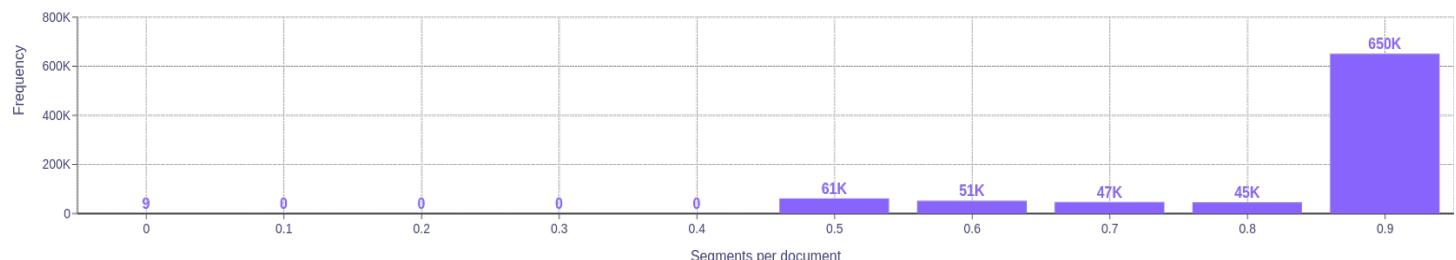
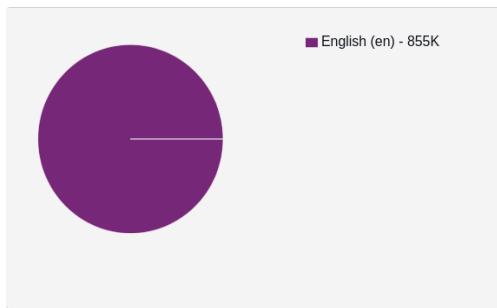
Volumes

Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size
854,829	2,621 (0.31 %)	22M	21M	112.91 MB	128.36 MB

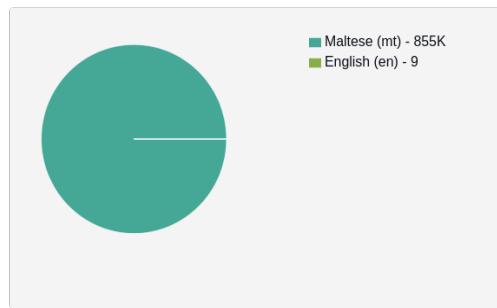
Type-Token Ratio

Source	Target
0.01	0.03

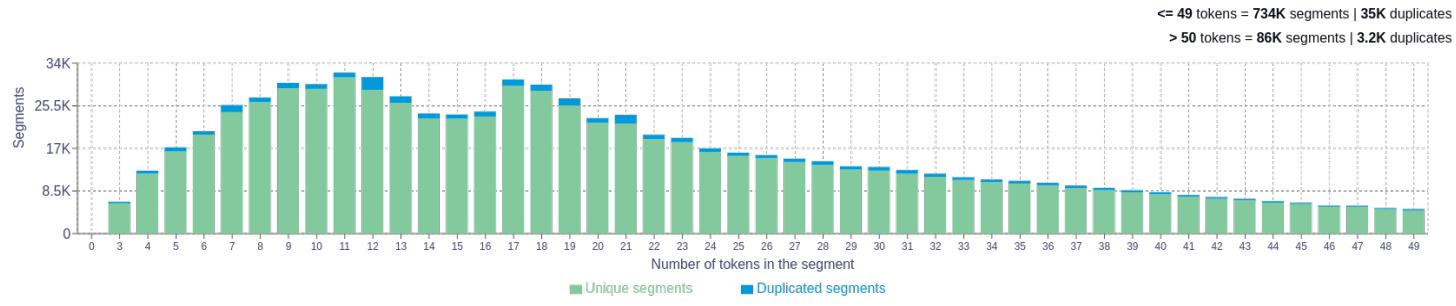
Translation likelihood

Language Distribution
Source

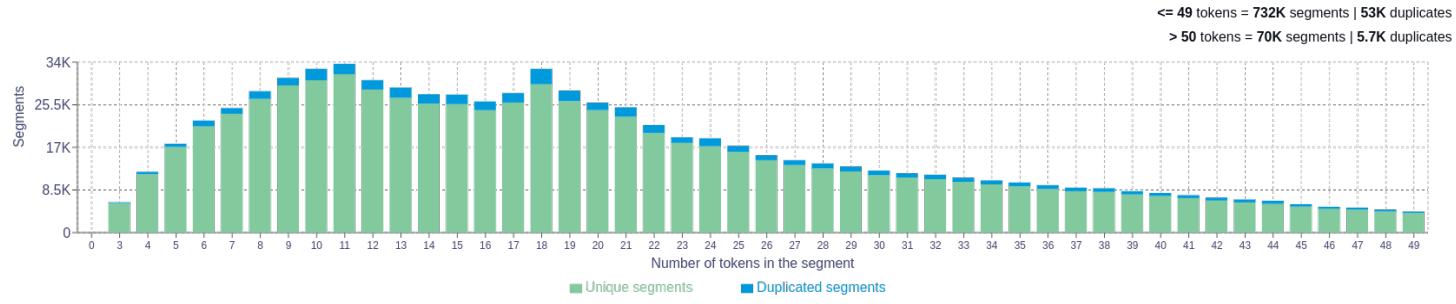
Target



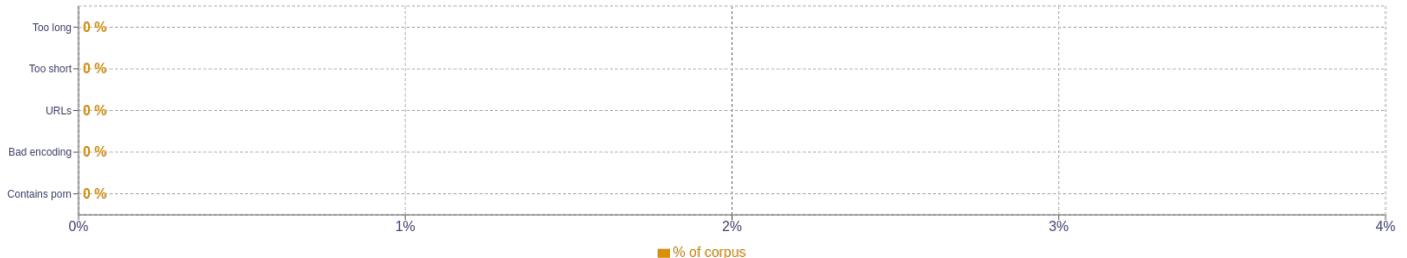
Source segment length distribution by token



Target segment length distribution by token



Segment pair noise distribution



Source n-grams

Size	n-grams
1	(article 63918) (european 59923) (member 53497) (eu 53095) (commission 46273)
2	(archive entry 34547) (member states 30631) (member state 19597) (european parliament 15416) (offer archive 15234)
3	(compare every offer 15260) (offer archive entry 15234) (add to cart 11166) (call new offers 7569) (accordance with article 5941)
4	(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)
5	(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)

Target n-grams

Size	n-grams
1	(ta 809041) (u 527492) (li 472797) (jew 149785) (b 139257)
2	(b 'mod 37841) (id-dhul ta 35589) (u li 16234) (kull offerta 15283) (kotba kollha 15268)
3	(gabbel kull offerta 15266) (offerta id-dhul ta 15265) (kull offerta id-dhul 15265) (b 'mod partikolari 12250) (offerti godda għal 8432)
4	(gabbel kull offerta id-dhul 15265) (kull offerta id-dhul ta 15265) (tal-parlament ewropew u tal-kunsill 8140) (is-sejha offerti godda għal 7556) (tat-tluq mhux inkluža l-vat 2410)
5	(gabbel kull offerta id-dhul ta 15265) (prezz tat-tluq mhux inkluža l-vat 2410) (ke tal-parlament ewropew u tal-kunsill 1971) (talba is-sejha offerti godda għal 1889) (huma impenjati li jipprova l-aqwa 1738)

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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>, , , 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>