HPLT Analytics report

HPLTAnalytics

General overview

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

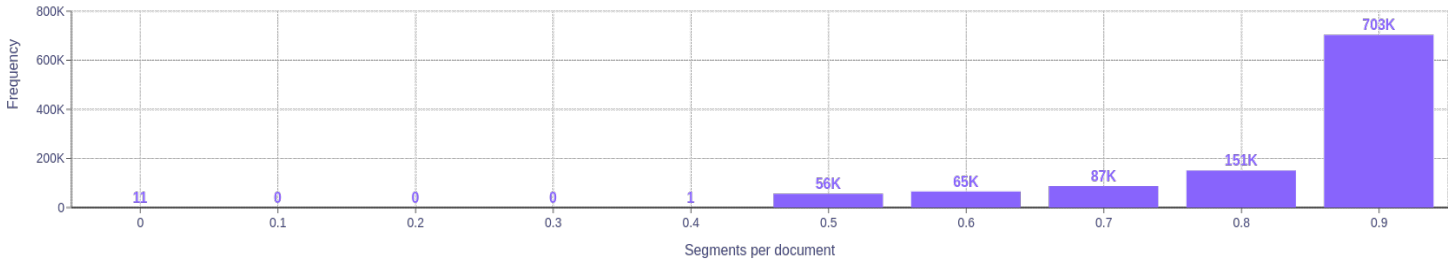
Volumes

Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size
1,063,114	2,096 (0.20 %)	16M	17M	82.49 MB	89.01 MB

Type-Token Ratio

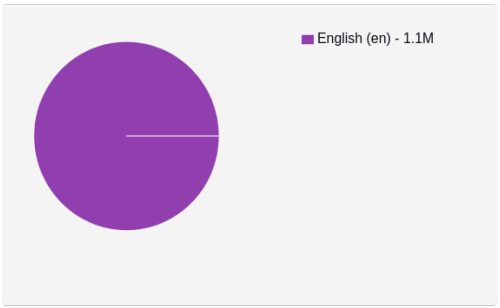
Source	Target
0.03	0.03

Translation likelihood

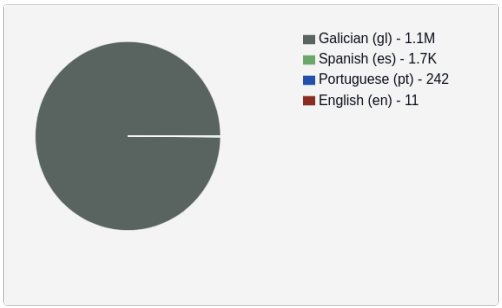


Language Distribution

Source

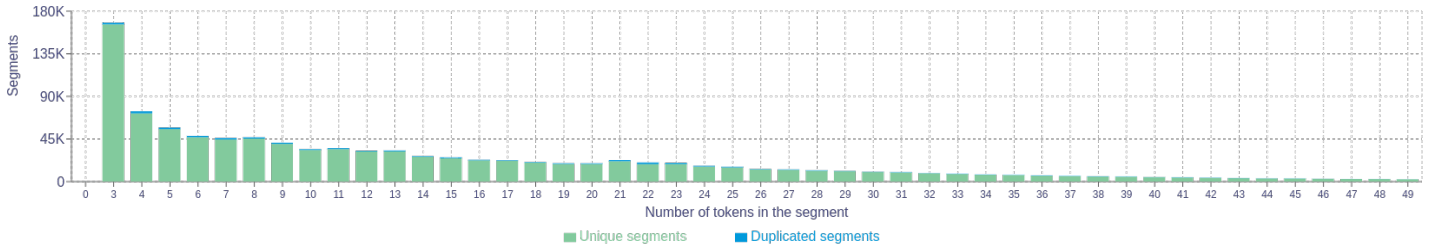


Target



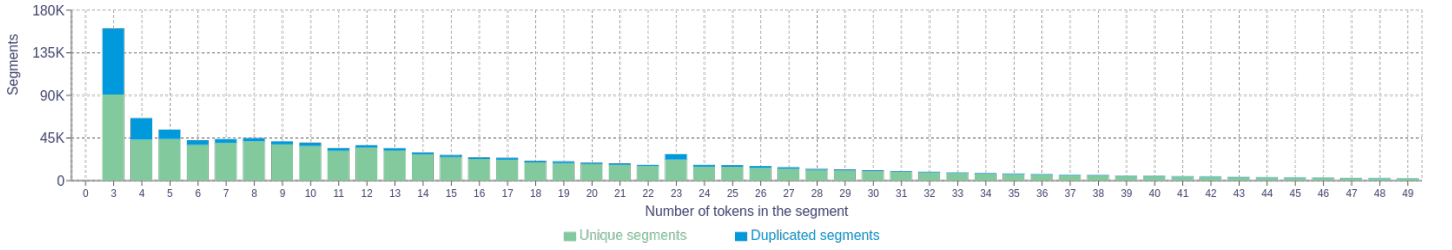
Source segment length distribution by token

<= 49 tokens = **1M** segments | **29K** duplicates
> 50 tokens = **34K** segments | **931** duplicates

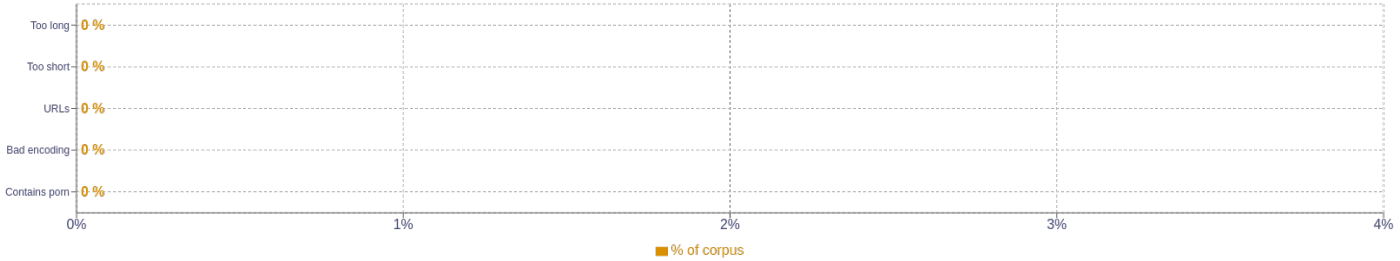


Target segment length distribution by token

<= 49 tokens = **858K** segments | **170K** duplicates
> 50 tokens = **35K** segments | **3.3K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	trails | 104164 · routes | 100458 · de | 28396 · calories | 27050 · one | 27008
2	serving size | 19176 · best trails | 10027 · get full | 8400 · full analysis | 8393 · active installations | 8266
3	get full analysis | 8393 · active installations tested | 8184 · analysis of surname | 4259 · analysis of name | 4133 · please try one | 3749
4	full analysis of surname | 4259 · full analysis of name | 4133 · one of these games | 3745 · male get full analysis | 1440 · uncommon names with surname | 1281
5	get full analysis of surname | 4259 · get full analysis of name | 4133 · try one of these games | 3743 · wikimedia commons has media related | 803 · list of surnames with name | 630

Target n-grams

Size	n-grams
1	rutas | 221335 · calorías | 27055 · información | 26901 · tamaño | 26428 · proteínas | 25458
2	sitio web | 9803 · mellores rutas | 9196 · instalacións activas | 8794 · actualizado fai | 8793 · activas probado | 8655
3	instalacións activas probado | 8655 · editar a fonte | 4956 · unidos de américa | 4488 · completa do apelido | 4459 · obter unha análise | 4425
4	tamaño de la porción | 23805 · estados unidos de américa | 4486 · obter unha análise completa | 4424 · análise completa do apelido | 4423 · obteña unha análise completa | 3969
5	nomes máis comúns e pouco | 1162 · commons ten máis contidos multimedia | 1044 · wikimedia commons ten máis contidos | 1028 · macho obteña unha análise completa | 798 · datos de compañías de australia | 692

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).

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