

## General overview

Corpus	Analytics date	Language
gl_1.jsonl.tsv	3/21/2024	Galician (gl)

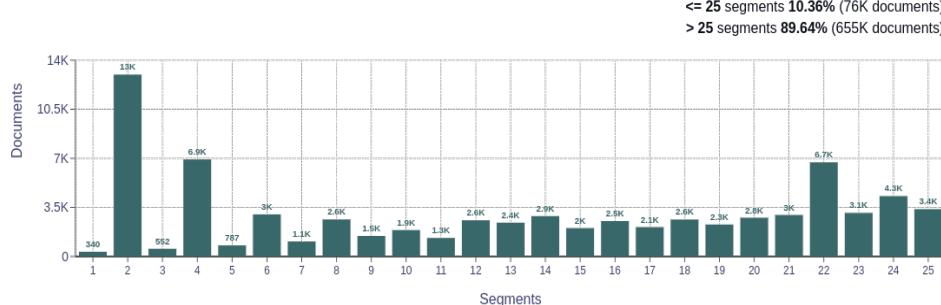
## Volumes

Docs	Segments	Unique segments	Tokens	Size
731,356	92,682,759	45,786 (0.05 %)	1B	5.03 GB

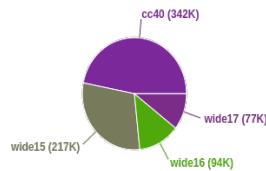
## Type-Token Ratio

Galician (gl)
0.01

## Documents size (in segments)

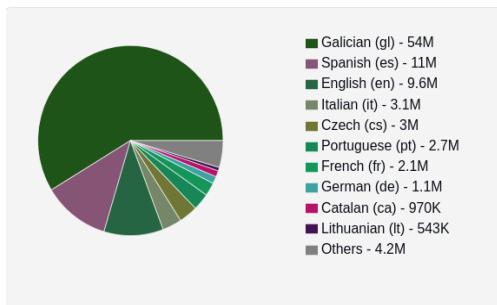


## Documents by collection

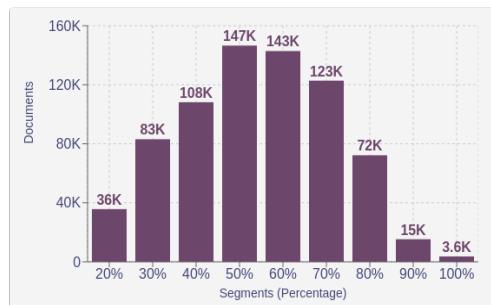


## Language Distribution

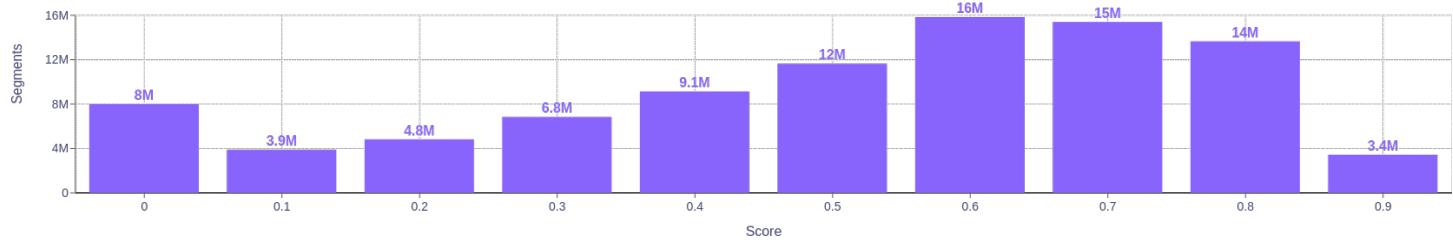
## Number of segments



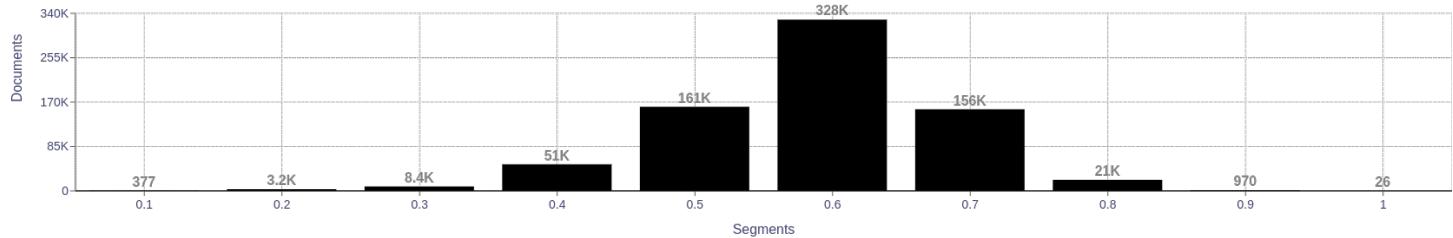
## Percentage of segments in Galician (gl) inside documents



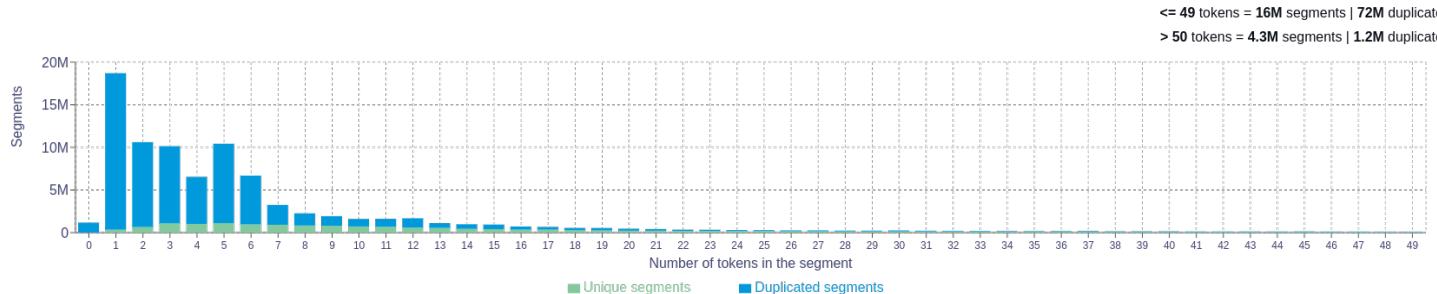
## Distribution of segments by fluency score



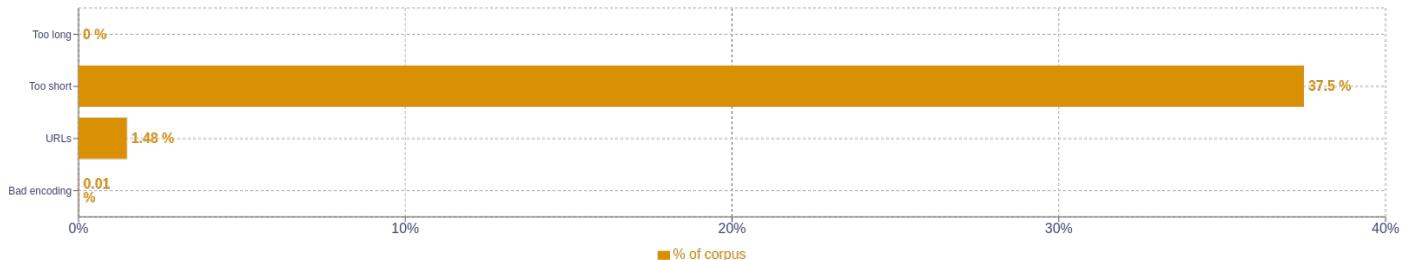
## Distribution of documents by average fluency score



## Segment length distribution by token



## Segment noise distribution



## Frequent n-grams

Size	n-grams
1	(y   2371096 galicia   1898093 día   1110181 abril   1085104 marzo   1051215)
2	(hay comentarios   300026 correo electrónico escribe   292932 correo electrónico   231870 aviso legal   220674 sitio web   215834)
3	(enviar por correo   450385 facebookcompartir en pinterest   433535 blogcompartir con twittercompartir   292942 electrónico escribe un blogcompartir   292931 twittercompartir con facebookcompartir   289851)
4	(enviar por correo electrónico escribe   292932 correo electrónico escribe un blogcompartir   292931 enviar por correo electrónico blogthis   143634 enlaces a esta entrada   68538 entradas antiguas página principal   53522)
5	(electrónico escribe un blogcompartir con twittercompartir   292931 blogcompartir con twittercompartir con facebookcompartir   289851 twittercompartir con facebookcompartir en pinterest   289850 compartir en twittercompartir en facebookcompartir   143696 twittercompartir en facebookcompartir en pinterest   143685)

## 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/mabanon/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>