

General overview

Corpus	Analytics date	Language
eu_1.jsonl.tsv	3/20/2024	Basque (eu)

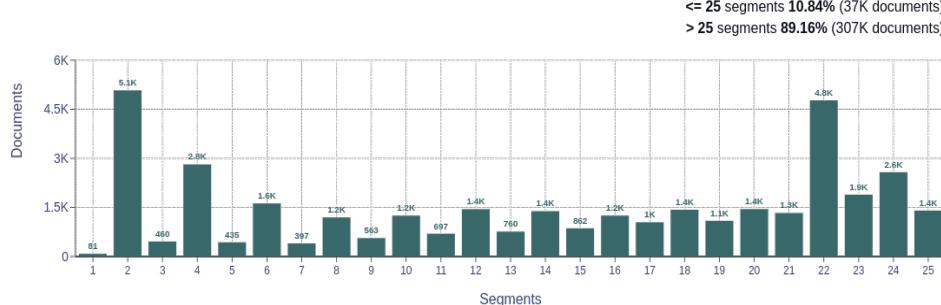
Volumes

Docs	Segments	Unique segments	Tokens	Size
343,947	37,226,281	31,743 (0.09 %)	412M	2.3 GB

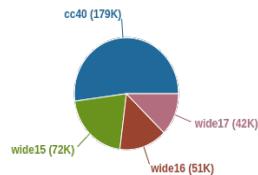
Type-Token Ratio

Basque (eu)
0.02

Documents size (in segments)

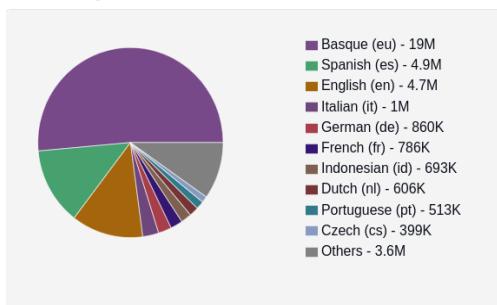


Documents by collection

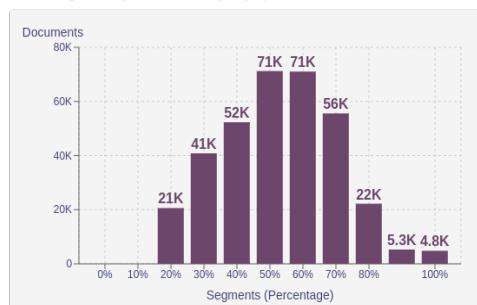


Language Distribution

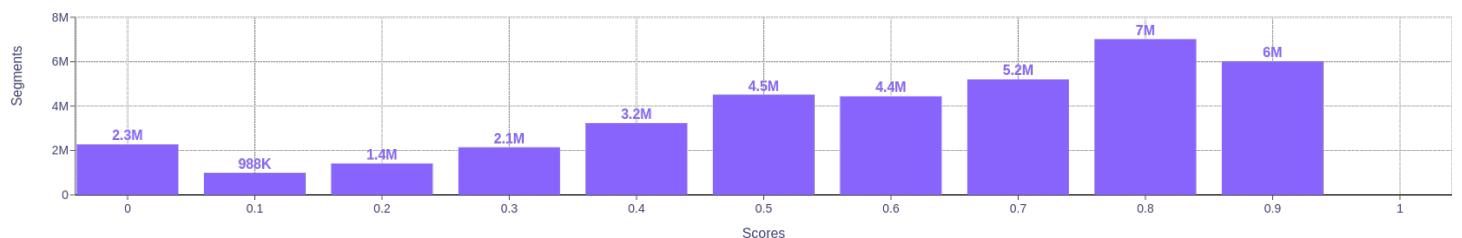
Number of segments



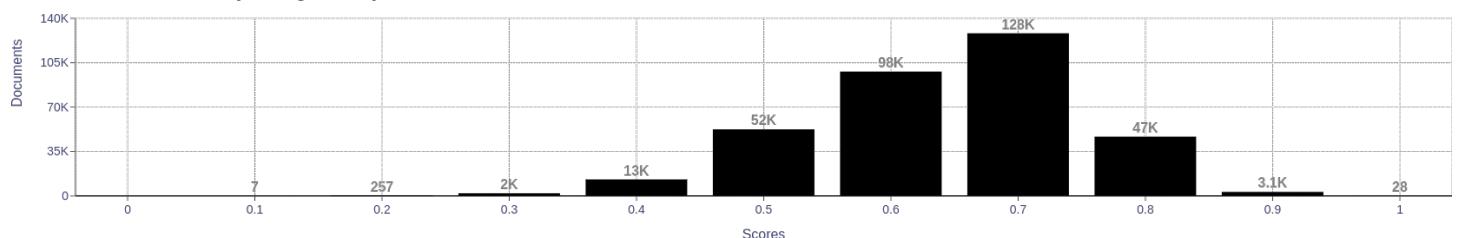
Percentage of segments in Basque (eu) inside documents



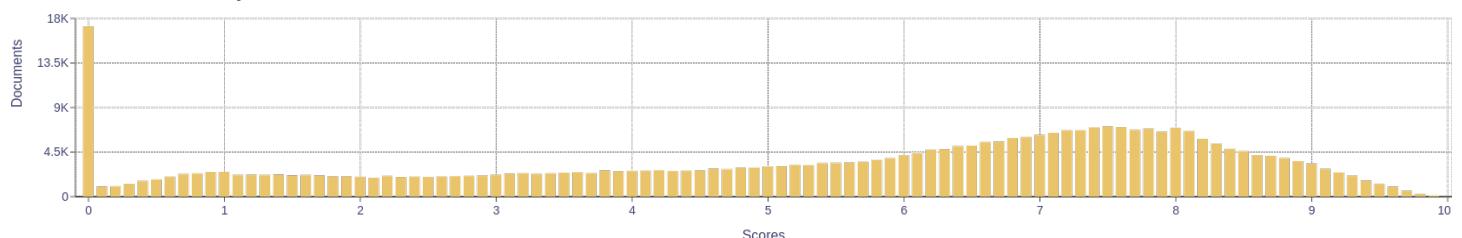
Distribution of segments by fluency score



Distribution of documents by average fluency score

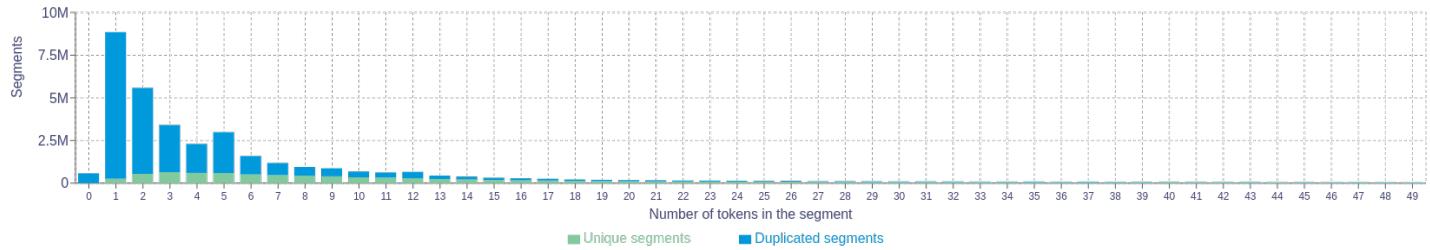


Distribution of documents by document score

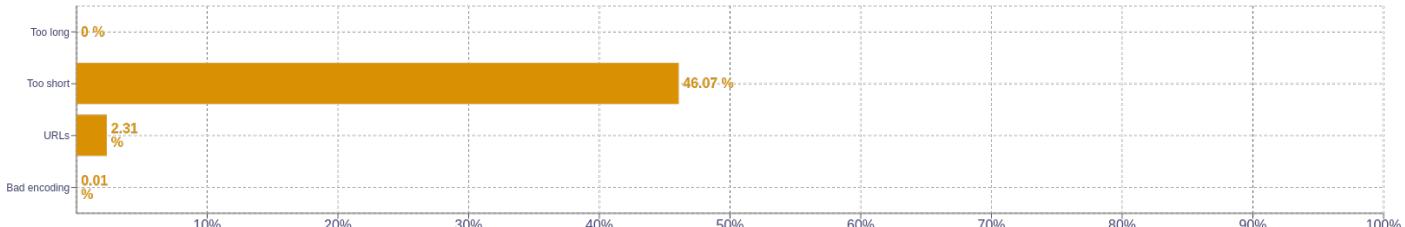


Segment length distribution by token

<= 49 tokens = 8.8M segments | 27M duplicates
 > 50 tokens = 1.8M segments | 416K duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	(de 4070563) (la 1959228) (en 1634840) (y 1349054) (el 1304858)
2	(de la 497473) (en el 218700) (en la 189454) (de los 165211) (a la 148796)
3	(no hay comentarios 93711) (enviar por correo 87288) (con twittercompartir con 86480) (un blogcompartir con 86479) (blogcompartir con twittercompartir 86479)
4	(un blogcompartir con twittercompartir 86479) (blogcompartir con twittercompartir con 86479) (por correo electrónico escribe un 86472) (enviar por correo electrónico escribe 86472) (electrónico escribe un blogcompartir con 86472)
5	(un blogcompartir con twittercompartir con 86479) (por correo electrónico escribe un blogcompartir 86472) (enviar por correo electrónico escribe un 86472) (electrónico escribe un blogcompartir con twittercompartir 86472) (correo electrónico escribe un blogcompartir con 86472)

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>).

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>