

## General overview

Corpus	Analytics date	Language
is_1.jsonl.tsv	3/22/2024	Icelandic (is)

## Volumes

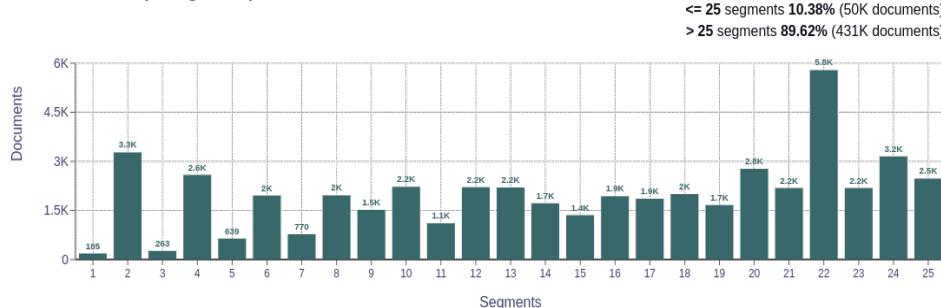
Docs	Segments	Unique segments	Tokens	Size
481,328	62,190,599	40,366 (0.06 %)	662M	3.66 GB

## Type-Token Ratio

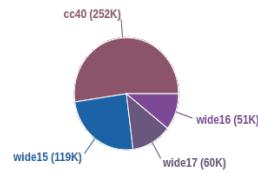
Icelandic (is)

0.01

## Documents size (in segments)

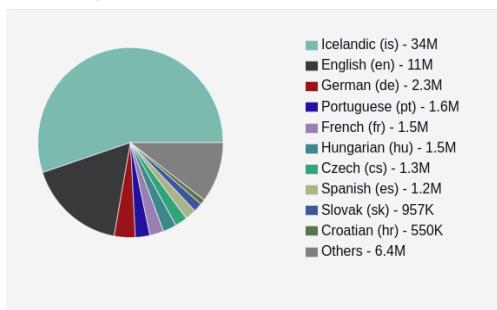


## Documents by collection

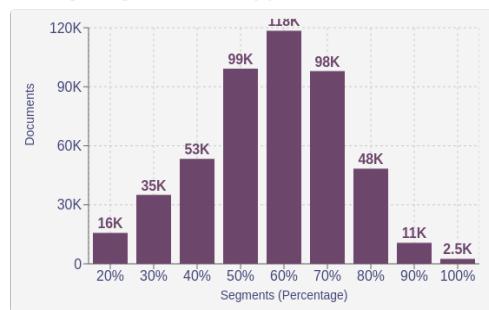


## Language Distribution

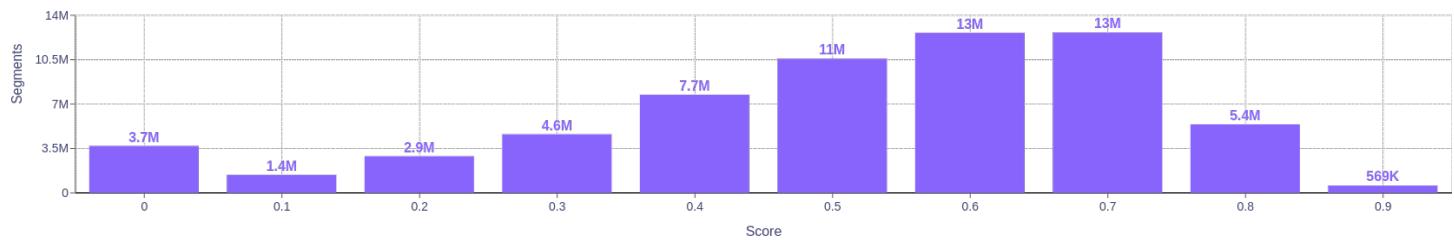
## Number of segments



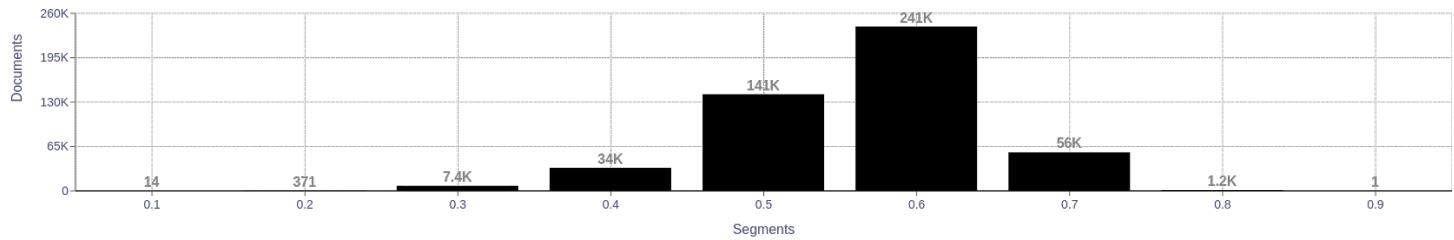
## Percentage of segments in Icelandic (is) inside documents



## Distribution of segments by fluency score

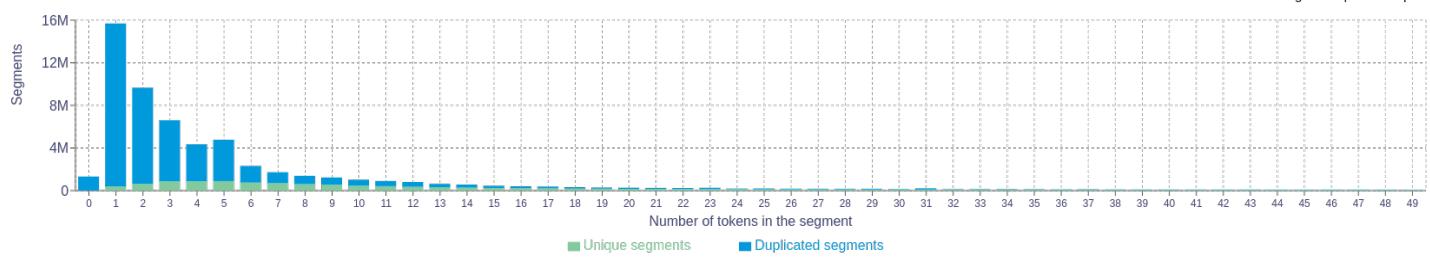


## Distribution of documents by average fluency score

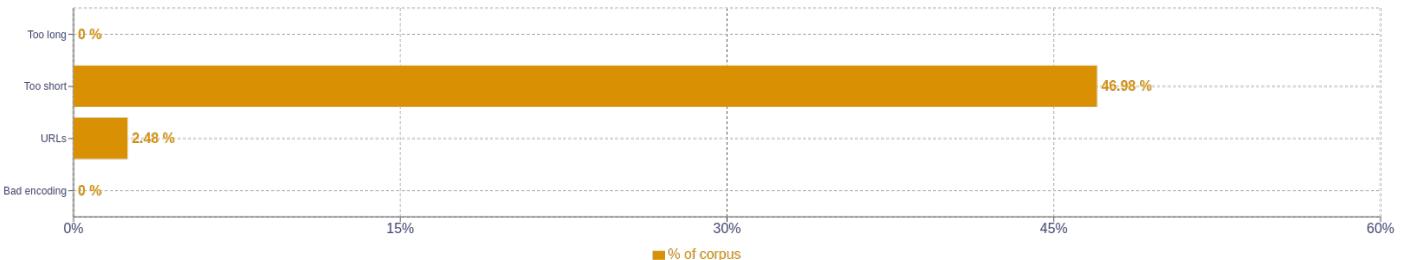


## Segment length distribution by token

<= 49 tokens = 12M segments | 47M duplicates  
> 50 tokens = 3M segments | 708K duplicates



## Segment noise distribution



## Frequent n-grams

Size	n-grams
1	var   2408989   the   1902422   a   1596154   hafa   1271463   and   1097022
2	hafa samband   193086   gististaðnum mynd   172266   in the   167965   lesa meira   160579   eyða breyta   112195
3	mynd af gististaðnum   185686   sýna meira sýna   79671   fær góða einkunn   78014   meðalverð á nótt   67449   hér á landi   62376
4	mynd af gististaðnum mynd   172266   gististaðnum mynd af gististaðnum   172266   cookie is set by   34405   opnast í nýjum glugga   33409   the cookies in the   29360
5	gististaðnum mynd af gististaðnum mynd   167795   twitterdeila á facebookdeila á pinterest   41666   (deila á twitterdeila á facebookdeila   41666)   user consent for the cookies   29359   the user consent for the   29359

## 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/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>