

## General overview

Corpus	Analytics date	Language
nn_1.jsonl.tsv	3/19/2024	Norwegian Nynorsk (nn)

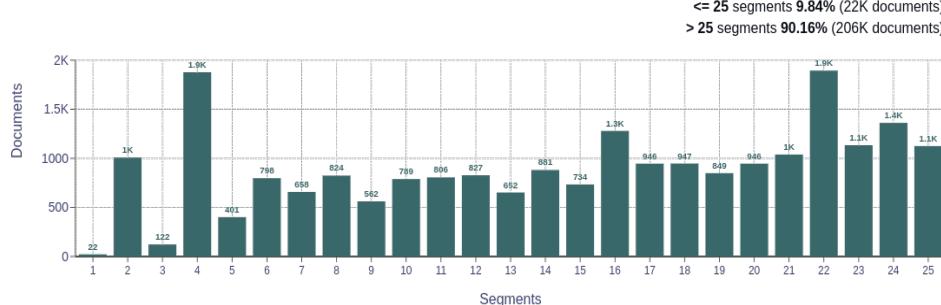
## Volumes

Docs	Segments	Unique segments	Tokens	Size
228,480	28,787,245	24,621 (0.09 %)	350M	1.77 GB

## Type-Token Ratio

Norwegian Nynorsk (nn)
0.01

## Documents size (in segments)



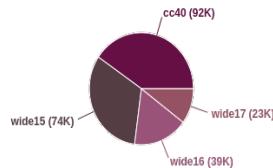
## Top 10 domains

Domain	Docs	% of total
blogspot.no	26K	11.50
wikipedia.org	14K	6.15
docplayer.me	9.9K	4.31
ndla.no	9.1K	3.99
blogspot.com	8.8K	3.84
framtida.no	3.7K	1.64
lokalhistoriewiki.no	2.8K	1.21
wordpress.com	2.2K	0.94
midsiden.no	1.8K	0.77
uib.no	1.6K	0.68

## Top 10 TLDs

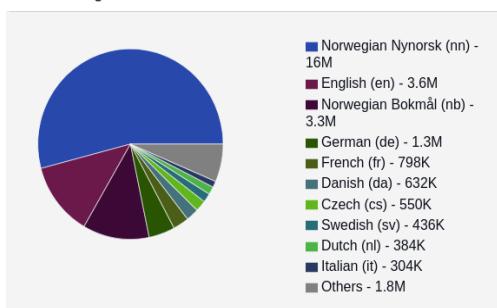
Domain	Docs	% of total
no	147K	64.30
com	31K	13.68
org	19K	8.45
me	9.9K	4.32
kommune.no	5.2K	2.27
net	3.5K	1.54
info	2K	0.86
vgs.no	1.1K	0.49
fr	871	0.38
se	772	0.34

## Documents by collection

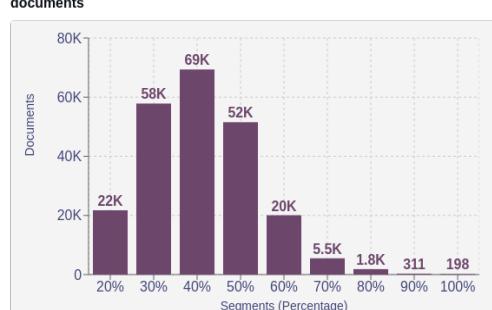


## Language Distribution

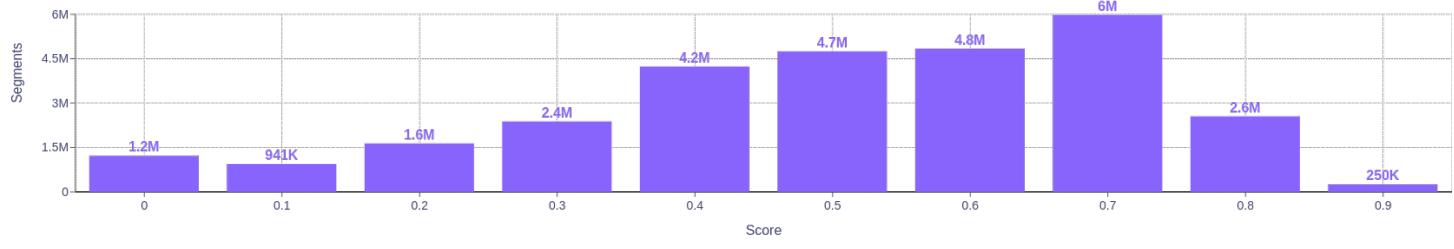
## Number of segments



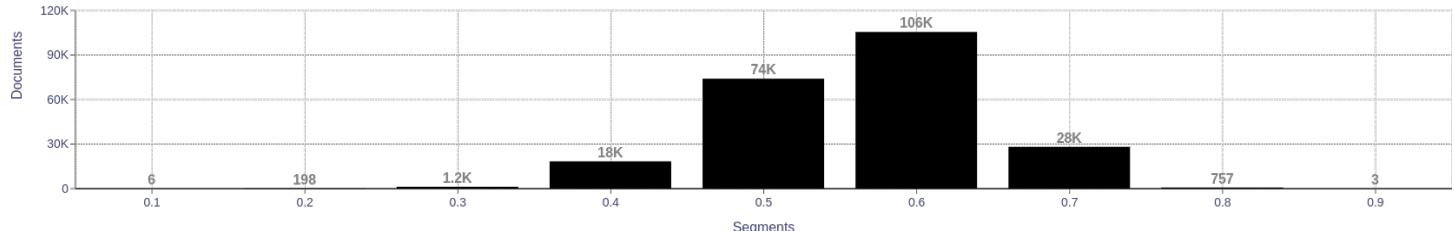
## Percentage of segments in Norwegian Nynorsk (nn) inside documents



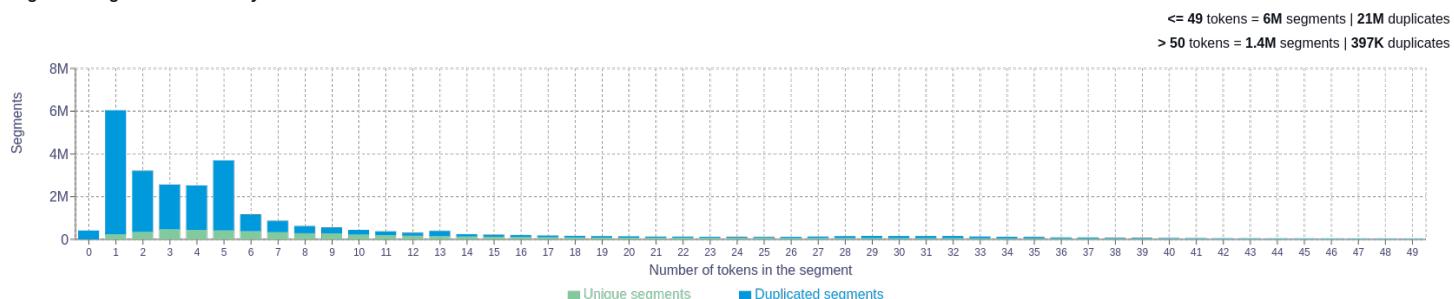
## Distribution of segments by fluency score



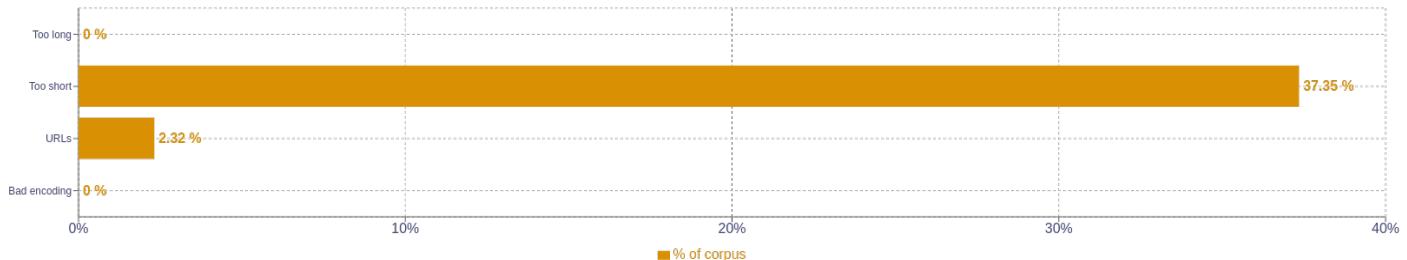
## Distribution of documents by average fluency score



## Segment length distribution by token



## Segment noise distribution



## Frequent n-grams

Size	n-grams
1	(detaljer   895215) (kommune   733749) (år   692325) (to   514673) (the   456217)
2	(kommune møteprotokoll   110302) (via e-postblogg   91549) (funksjon representerer   63615) (møteprotokoll utval   59567) (utval møtedato   52739)
3	(send dette via   91644) (del på twitterdel   91550) (facebookdel på pinterest   91313) (twitterdel på facebookdel   91312) (sogn og fjordane   77264)
4	(send dette via e-postblogg   91549) (nyere innlegg eldre innlegg   33711) (innlegg eldre innlegg start   25492) (nynorsk/bokmål nynorsk eksamensinformasjon eksamenstid   22900) (følgjande faste medlemmer møtte   21141)
5	(twitterdel på facebookdel på pinterest   91312) (del på twitterdel på facebookdel   91312) (nyere innlegg eldre innlegg start   25492) (faste medlemmer var til stades   14925) (ronk ronk ronk ronk ronk   13092)

## 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>