HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Analytics date	Language
lim_Latn.jsonl.tsv	12/5/2024	Limburgish (li)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
367,929	7,139,712	2,734,723 (38.30 %)	225M	1.06 GB	1,118,709,976

Top 10 domains

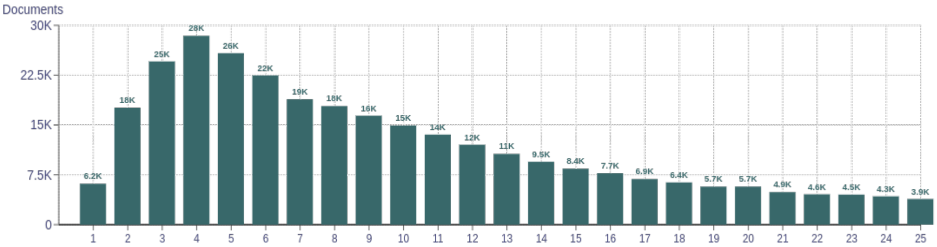
Domain	Docs	% of total
wikipedia.org	198K	53.86
omropfryslan.nl	24K	6.41
vv-sds.nl	5.1K	1.40
itnijs.frl	3K	0.82
demoanne.nl	2.2K	0.60
sirkwy.frl	2.2K	0.59
dbnl.org	1.9K	0.51
limburgslied.nl	1.4K	0.39
vsaduidoma.com	1.3K	0.35
stuft.nl	1.3K	0.34

Top 10 TLDs

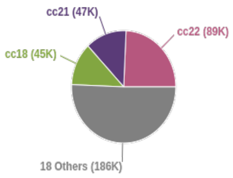
Domain	Docs	% of total
org	205K	55.67
nl	103K	27.92
com	27K	7.38
frl	11K	3.01
be	5.1K	1.39
eu	3.4K	0.92
de	3.2K	0.88
net	2.3K	0.63
zone	1.1K	0.30
info	749	0.20

Documents size (in segments)

<= 25 segments **82%** (302K documents)
> 25 segments **18%** (66K documents)

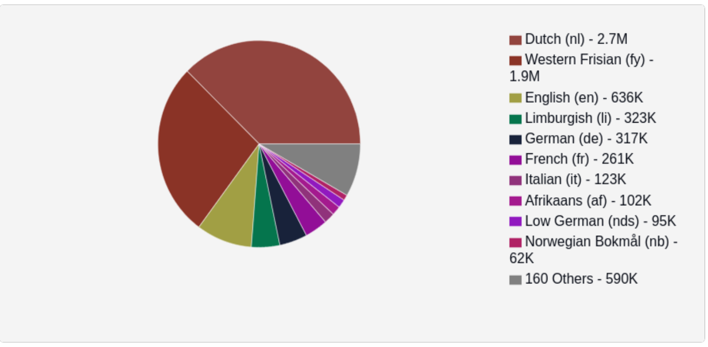


Documents by collection

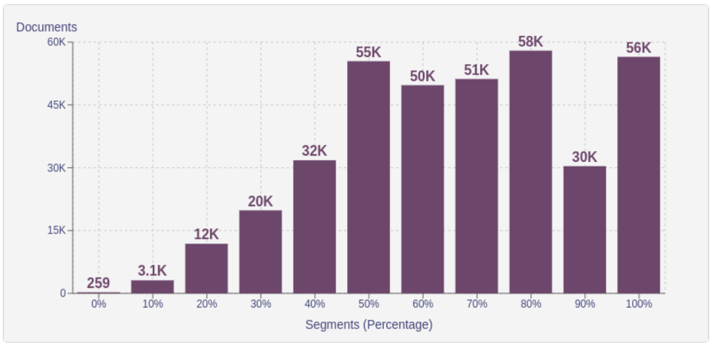


Language Distribution

Number of segments

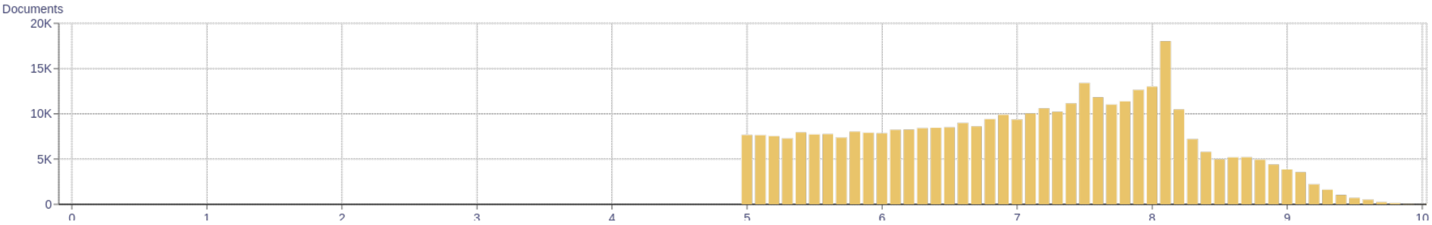


Percentage of segments in Limburgish (li) inside documents



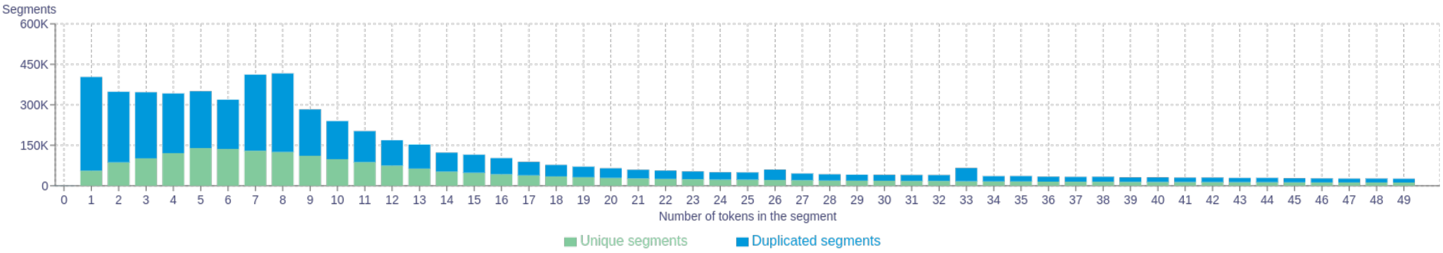
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (368K documents)

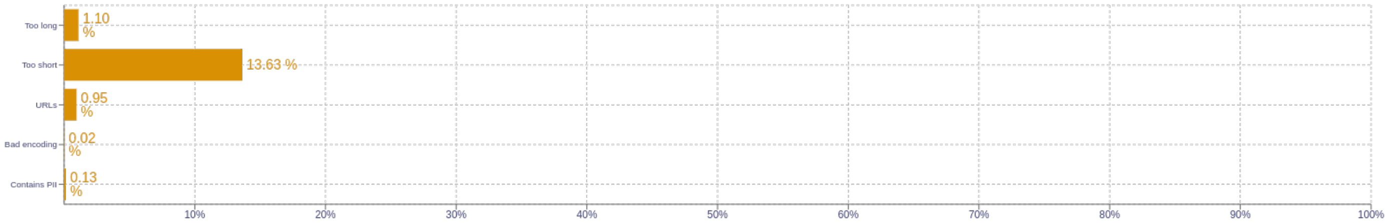


Segment length distribution by token

<= 49 tokens = **2.1M** segments | **3.6M** duplicates
> 50 tokens = **1.4M** segments | **760K** duplicates



Segment noise distribution



Frequent n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>in | 4255962
fan | 2952713
it | 2794728
t | 2753873
van | 2552892</td></tr>
<tr><td>2</td><td>yn it | 307308
fan it | 295853
bewurkje seksje | 235340
van der | 152569
boarne bewurkje | 148729</td></tr>
<tr><td>3</td><td>dit artikel is | 56055
artikel is gesjreve | 47812
is gesjreve in | 33063
aan te gaeve | 30573
aan te hauwe | 30471</td></tr>
<tr><td>4</td><td>dit artikel is gesjreve | 47805
artikel is gesjreve in | 33024
t weurt gewaardeerd óm | 30110
of aan te gaeve | 30104
gaeve welk anger dialek | 30104</td></tr>
<tr><td>5</td><td>dit artikel is gesjreve in | 33021
of aan te gaeve welk | 30104
aan te gaeve welk anger | 30104
hauwe of aan te gaeve | 30102
aan te hauwe of aan | 30102</td></tr>
</table>

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).

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