HPLT Analytics report

HPLTAnalytics

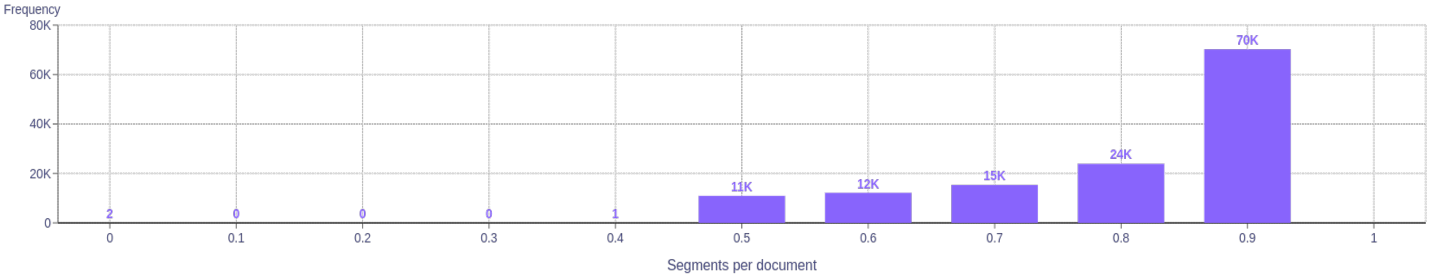
General overview

Corpus	Analytics date	Source language	Target language
HPLT.en-nn	10/26/2023	English (en)	Norwegian Nynorsk (nn)

Volumes

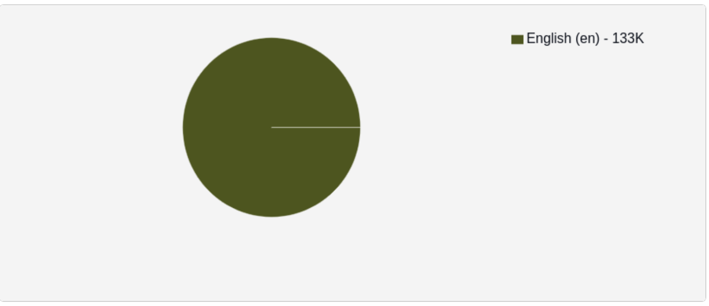
Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size	Src characters	Trg characters
132,540	132,539 (100.00 %)	2.5M	2.2M	12.09 MB	11.77 MB		

Translation likelihood

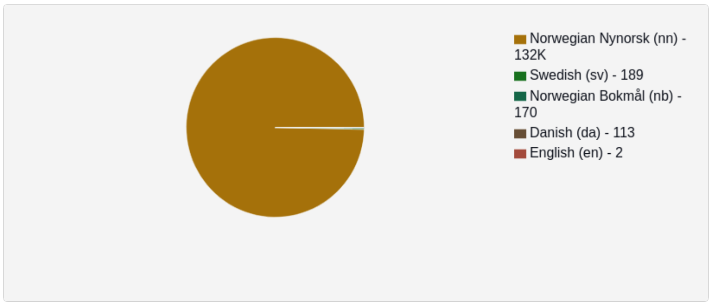


Language Distribution

Source



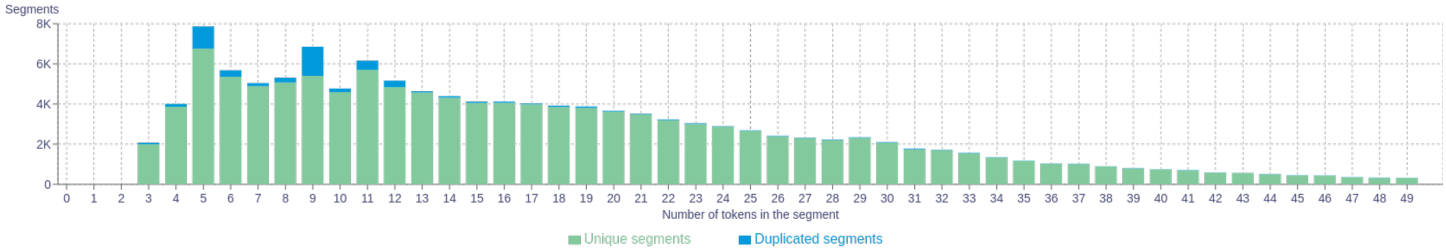
Target



Source segment length distribution by token

<= 49 tokens = **124K** segments | **5.4K** duplicates

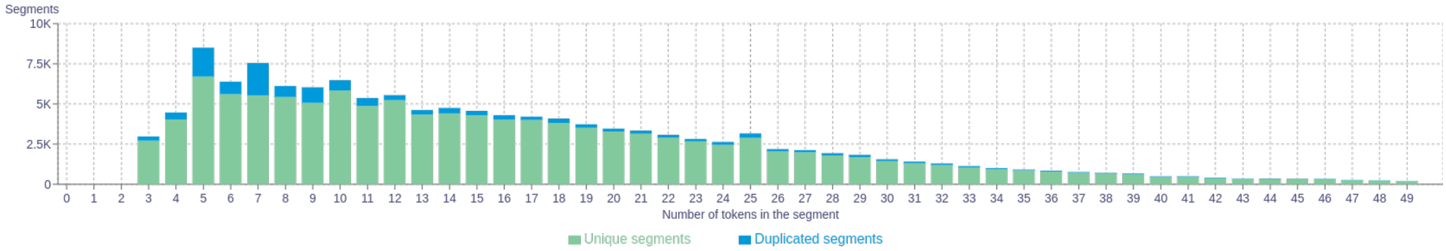
> 50 tokens = **3.6K** segments | **37** duplicates



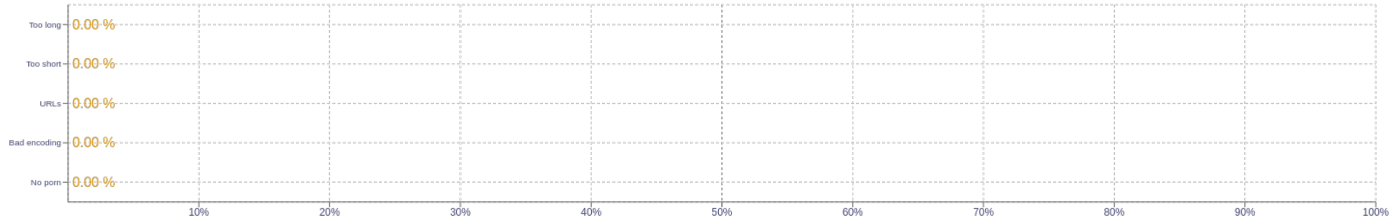
Target segment length distribution by token

<= 49 tokens = **117K** segments | **13K** duplicates

> 50 tokens = **2.5K** segments | **189** duplicates



Segment pair noise distribution



Source n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>time | 4975
one | 4463
new | 4433
also | 4414
map | 3808</td></tr>
<tr><td>2</td><td>larger map | 3088
show larger | 3087
light breeze | 1411
solar time | 1198
day length | 1194</td></tr>
<tr><td>3</td><td>show larger map | 3087
length and solar | 1191
meters per second | 933
click on first | 396
second from south | 380</td></tr>
<tr><td>4</td><td>length and solar time | 1191
day length and solar | 1191
accepted by the municipality | 577
letter of the city | 396
click on first letter | 396</td></tr>
<tr><td>5</td><td>day length and solar time | 1191
first letter of the city | 396
meters per second from south | 380
meters per second from north | 353
accepted by the mapping authority | 285</td></tr>
</table>

Target n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>større | 3732
the | 3660
to | 3587
kart | 3546
vis | 3286</td></tr>
<tr><td>2</td><td>større kart | 3089
vis større | 3087
svak vind | 1201
sann soltid | 1195
lett bris | 990</td></tr>
<tr><td>3</td><td>vis større kart | 3087
daglengde og sann | 1191
meter per sekund | 933
klar for omsetjing | 606
godkjend av kommunen | 572</td></tr>
<tr><td>4</td><td>daglengde og sann soltid | 1191
klikk på første bokstav | 396
første bokstav i byen | 396
per sekund frå nord | 212
per sekund frå nordaust | 154</td></tr>
<tr><td>5</td><td>meter per sekund frå nord | 212
meter per sekund frå nordaust | 154
meter per sekund frå aust | 132
meter per sekund frå søraust | 122
installasjonar i drift tested with | 111</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