HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-nb.tsv	1/29/2025	English (en)	Norwegian Bokmål (nb)

Volumes

Segments	SL tokens	SL characters	SL size
22,912,722	481M	2,494,061,609	2.33 GB

TL tokens	TL characters	TL size
446M	2,464,850,087	2.34 GB

Dataset top 10 domains

SL domain	Segments	TL domain	Segments
google.com	7.8%	wikipedia.org	5.6%
wikipedia.org	6.5%	tripadvisor.com	4.3%
hotels.com	6.1%	venere.com	3.4%
venere.com	4.8%	hotels.com	3.0%
tripadvisor.com	2.0%	google.com	2.9%
agoda.com	1.9%	jetcost.no	1.4%
booking.com	1.3%	agoda.com	1.4%
hostelbookers.com	1.3%	expedia.no	1.3%
microsoft.com	1.3%	biblegateway.com	1.1%
biblegateway.com	1.2%	hostelworld.com	1.0%

Dataset top 10 TLDs

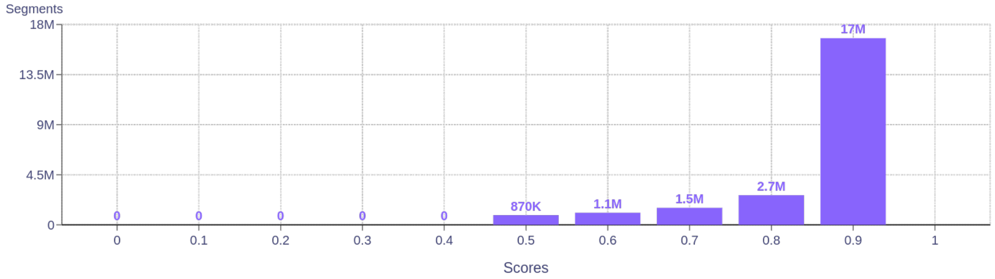
SL domain	Segments	TL domain	Segments
com	104.7%	com	68.8%
org	13.9%	no	34.3%
no	10.7%	org	10.9%
net	6.8%	net	5.7%
co.uk	6.2%	info	1.9%
ie	3.0%	eu	1.3%
eu	2.3%	me	0.8%
info	2.1%	de	0.5%
com.au	1.9%	ru	0.5%
ca	1.6%	se	0.4%

Translation likelihood

≥ 5 = 23M segments | **100.0%**

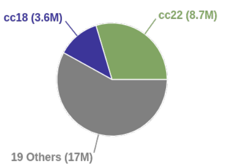
≥ 8 = 19M segments | **84.8%**

< 5 = 0 segments | **0.0%**



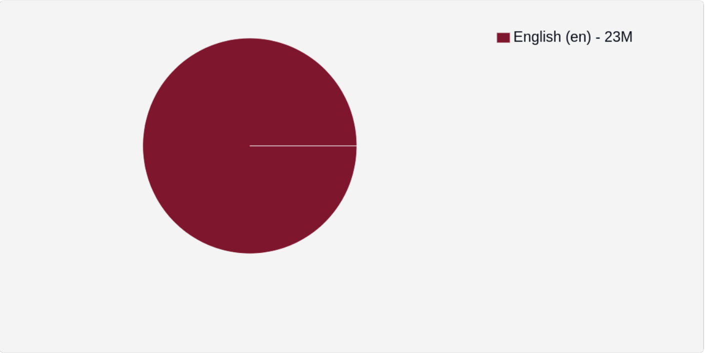
Collections

CC = 66.73%
IA = 33.27%

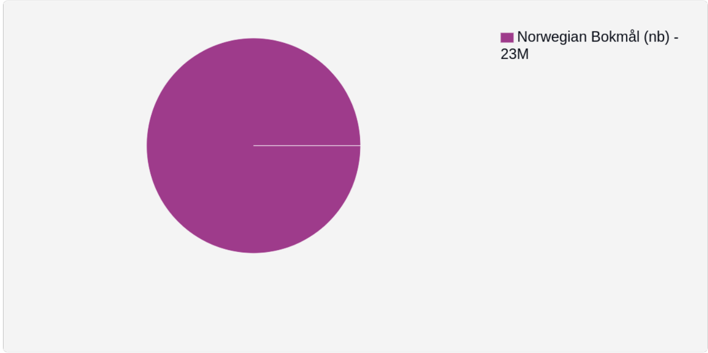


Language Distribution

Source



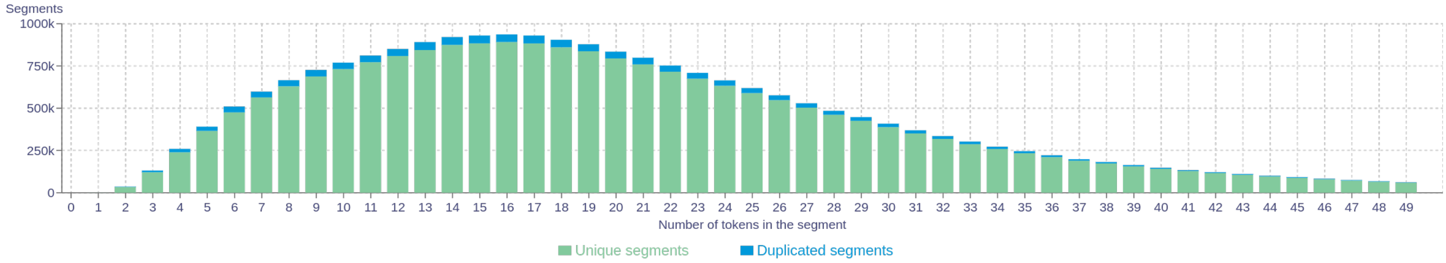
Target



Source segment length distribution by token

<= 49 tokens = **21M** segments | **1.1M** duplicates

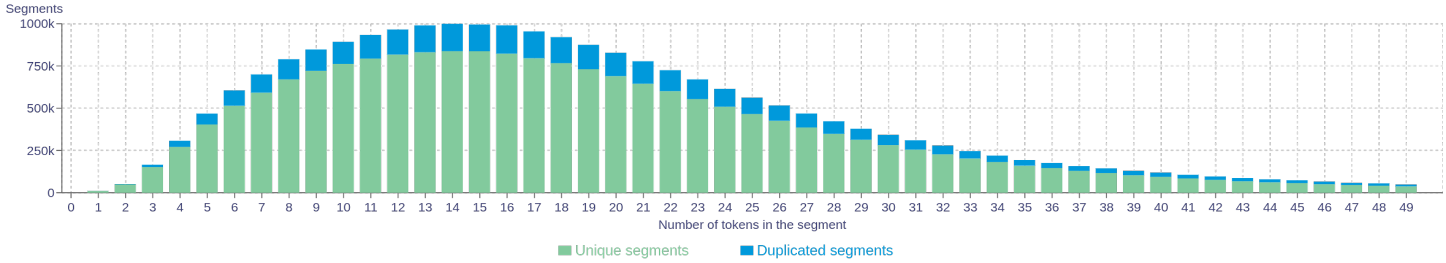
> 50 tokens = **664K** segments | **23K** duplicates



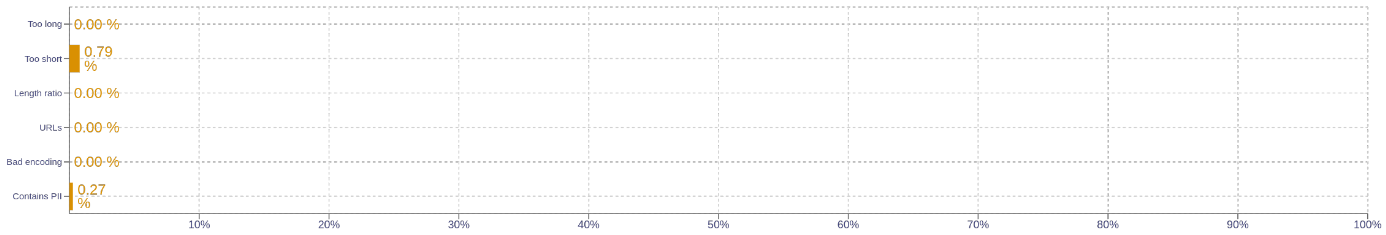
Target segment length distribution by token

<= 49 tokens = **19M** segments | **3.7M** duplicates

> 50 tokens = **491K** segments | **95K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	hotel | 1221802 also | 1136860 one | 924217 use | 789345 time | 699614
2	personal data | 115276 free wi-fi | 111749 hotel details | 87978 personal information | 76166 make sure | 72917
3	proud to partner | 145293 tripadvisor is proud | 145223 reservations with confidence | 145004 users who viewed | 55268 partner with booking.com | 46587
4	looked at this hotel | 30752 hotel in the last | 30387 one of the best | 24569 hotel reservations with confidence | 22771 wi-fi in public areas | 20298
5	tripadvisor is proud to partner | 145223 proud to partner with booking.com | 46587 proud to partner with expedia | 36615 venere so you can book | 31881 people looked at this hotel | 30381

Target n-grams

Size	n-grams
1	andre | 812602 mer | 769836 ligger | 602713 få | 579349 dine | 532454
2	gratis wi-fi | 106788 flere opplysninger | 96308 blant annet | 95020 hotellet ligger | 78890 triponline sa | 70787
3	tripadvisor er stolt | 145267 trygt kan bestille | 144599 partner med booking.com | 50512 flere opplysninger priser | 49987 alternativer for reservasjon | 41467
4	brukere som har sett | 55786 løpet av den siste | 31191 sett på dette hotellet | 30464 bør du se nærmere | 22000 wifi på alle rom | 15415
5	stolt av å være partner | 145267 venere slik at du trygt | 33518 løpet av den siste timen | 30749 otel slik at du trygt | 25705 tingo slik at du trygt | 23741

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