HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-nn.tsv	1/31/2025	English (en)	Norwegian Nynorsk (nn)

Volumes

Segments	SL tokens	SL characters	SL size
563,791	14M	69,346,941	66.38 MB

TL tokens	TL characters	TL size
12M	65,155,921	63.43 MB

Dataset top 10 domains

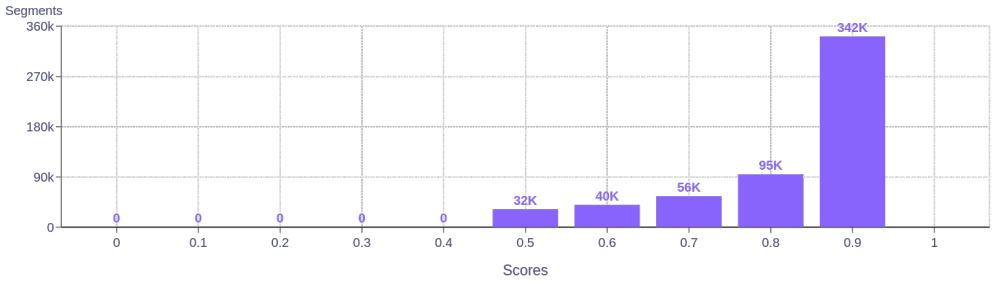
SL domain	Segments	TL domain	Segments
wikipedia.org	86.3%	wikipedia.org	72.1%
libreoffice.org	3.4%	uib.no	2.6%
gimp.org	2.8%	bible.com	2.6%
uib.no	2.7%	honsi.org	2.5%
schools-wikipedia.org	2.5%	gimp.org	2.1%
honsi.org	2.4%	ssb.no	2.0%
ssb.no	2.2%	libreoffice.org	1.9%
bible.com	2.2%	visitnorway.no	1.4%
encyclopedia.org	1.4%	npd.no	1.2%
visitnorway.com	1.3%	skatteetaten.no	1.1%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
org	103.5%	org	81.2%
no	32.3%	no	35.4%
com	20.7%	com	12.6%
net	1.8%	net	1.1%
ws	1.2%	ws	0.6%
edu	1.0%	info	0.4%
mobi	0.9%	mobi	0.3%
fm	0.6%	edu	0.2%
info	0.5%	me	0.2%
co.uk	0.2%	museum.no	0.2%

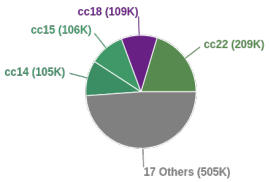
Translation likelihood

≥ 5 = 564K segments | **100.0%**
≥ 8 = 436K segments | **77.4%**
< 5 = 0 segments | **0.0%**



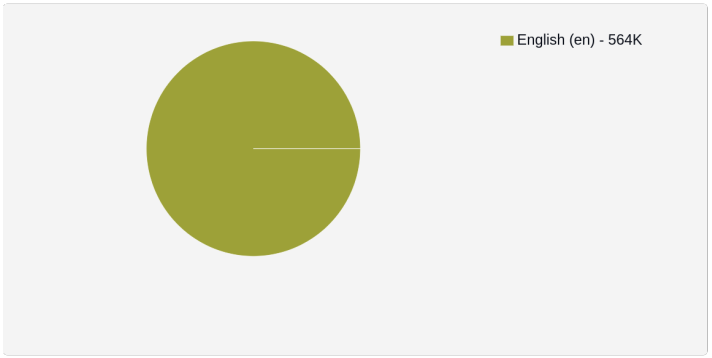
Collections

CC = 67.71%
IA = 32.29%

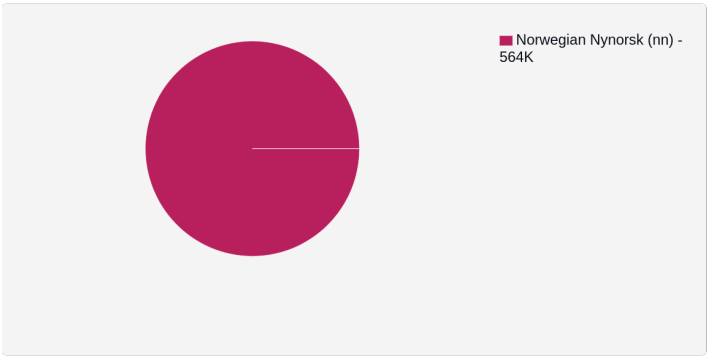


Language Distribution

Source

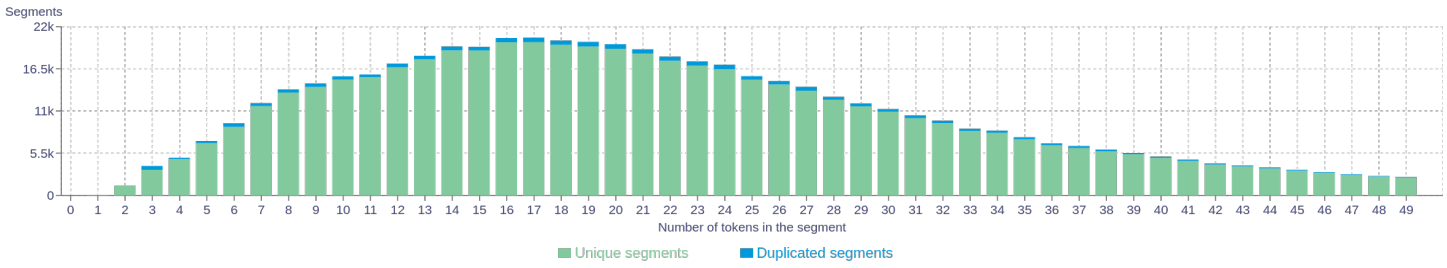


Target



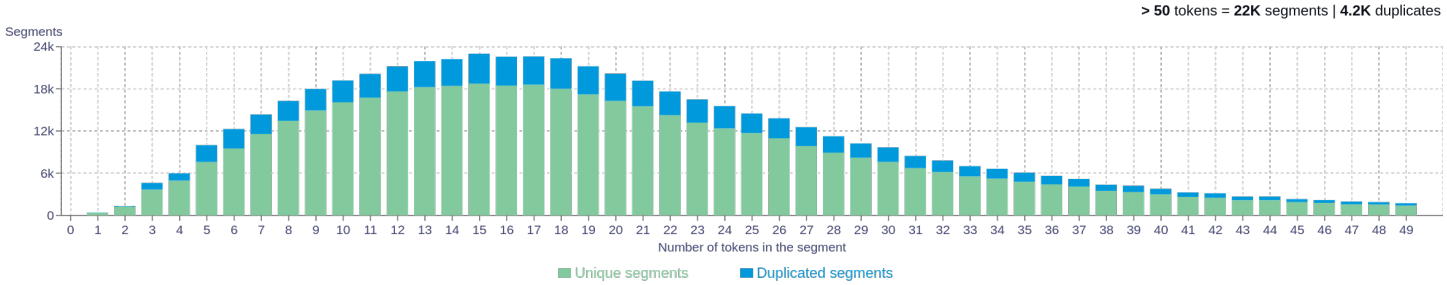
Source segment length distribution by token

<= 49 tokens = **514K** segments | **16K** duplicates
> 50 tokens = **33K** segments | **873** duplicates

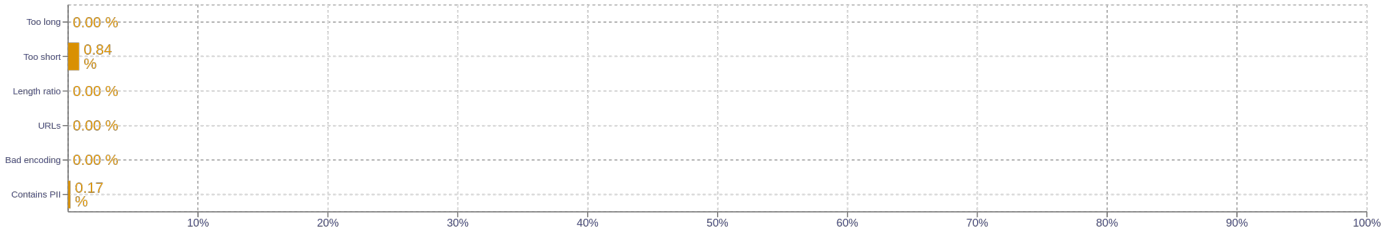


Target segment length distribution by token

<= 49 tokens = **439K** segments | **103K** duplicates
> 50 tokens = **22K** segments | **4.2K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	also | 26911 · one | 25121 · first | 20699 · new | 19634 · city | 16311
2	hotels near | 5153 · united states | 3641 · per cent | 2817 · world war | 2381 · citation needed | 2208
3	world war ii | 1252 · rock and roll | 803 · video in good | 796 · new full version | 795 · watch and download | 707
4	video in good quality | 796 · hd quality for free | 707 · download in hd quality | 707 · united states geological survey | 593 · incorporates public domain material | 564
5	watch and download in hd | 707 · article incorporates public domain material | 564 · material from the united states | 546 · domain material from the united | 546 · united states geological survey document | 545

Target n-grams

Size	n-grams
1	endre | 22422 · the | 22406 · under | 21757 · byen | 18275 · ligg | 18073
2	endre wikiteksten | 9229 · hotell nær | 5354 · new zealand | 1840 · of the | 1650 · new york | 1624
3	wikipedia på engelsk | 1246 · meter over havet | 1242 · video i god | 958 · eurovision song contest | 778 · se og laste | 702
4	video i god kvalitet | 958 · laste ned i hd-kvalitet | 702 · geographic names information system | 562 · united states geological survey-artikkelen | 557 · stoff som er offentleg | 557
5	stoff som er offentleg eige | 557 · offentleg eige frå united states | 557 · innhald frå geographic names information | 557 · inneheld stoff som er offentleg | 557 · eige frå united states geological | 557

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