HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-fi.tsv	2/1/2025	English (en)	Finnish (fi)

Volumes

Segments	SL tokens	SL characters	SL size
29,067,875	618M	3,222,108,570	3.01 GB

TL tokens	TL characters	TL size
463M	3,319,814,491	3.22 GB

Dataset top 10 domains

SL domain	Segments	TL domain	Segments
hotels.com	13.1%	europa.eu	5.2%
europa.eu	6.6%	hotels.com	5.0%
google.com	6.3%	wikipedia.org	4.0%
wikipedia.org	4.6%	google.com	2.4%
agoda.com	2.3%	agoda.com	1.7%
booking.com	2.2%	docplayer.fi	1.4%
microsoft.com	1.8%	bibliacatolica.com.br	1.3%
bibliacatolica.com.br	1.2%	tripadvisor.fi	1.3%
docplayer.net	1.1%	microsoft.com	1.2%
doctrinepublishing.com	1.1%	booking.com	1.2%

Dataset top 10 TLDs

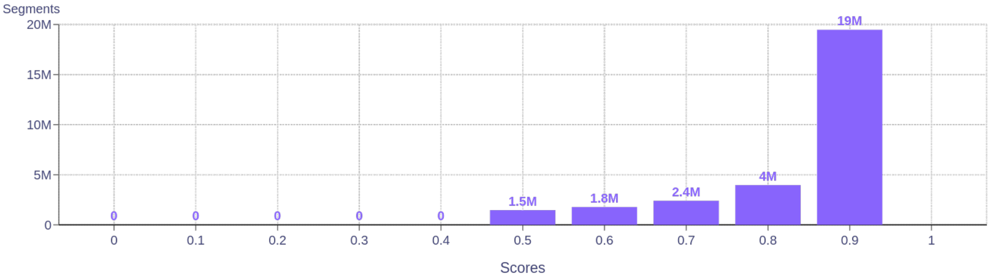
SL domain	Segments	TL domain	Segments
com	103.0%	com	61.1%
fi	19.3%	fi	43.5%
org	11.5%	org	8.5%
eu	8.7%	eu	6.7%
co.uk	5.4%	net	3.2%
net	5.1%	com.br	1.6%
de	2.2%	info	1.4%
ie	2.2%	de	1.0%
com.br	1.4%	ee	0.6%
info	1.4%	nl	0.5%

Translation likelihood

≥ 5 = 29M segments | **100.0%**

≥ 8 = 23M segments | **80.6%**

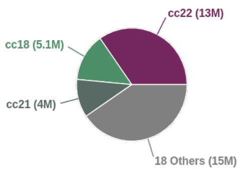
< 5 = 0 segments | **0.0%**



Collections

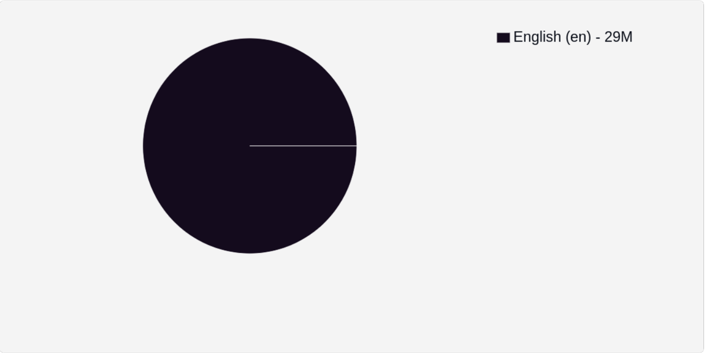
CC = 71.96%

IA = 28.04%

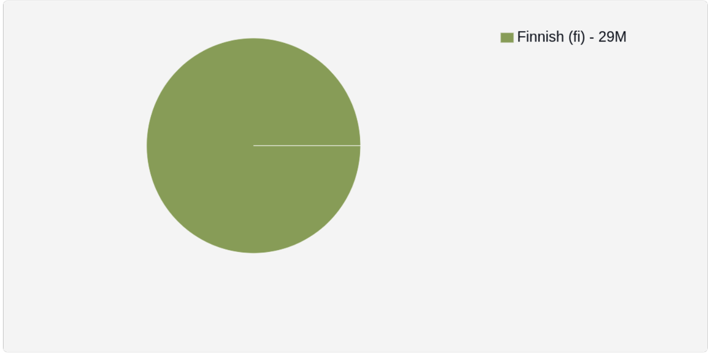


Language Distribution

Source



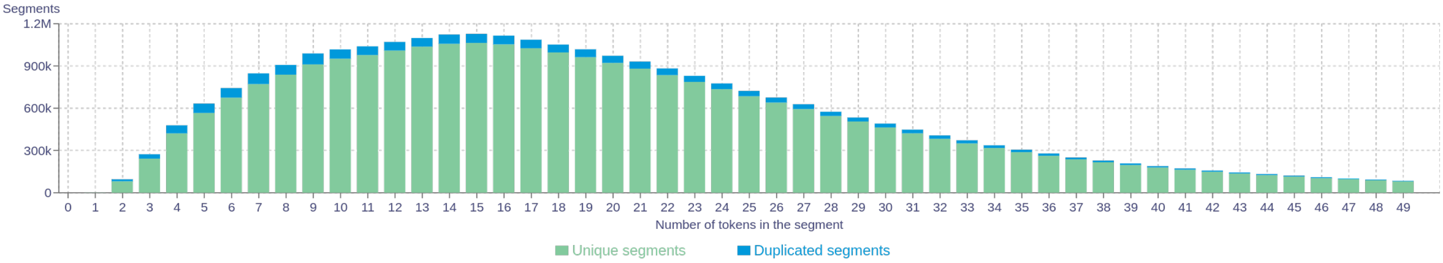
Target



Source segment length distribution by token

<= 49 tokens = **26M** segments | **1.7M** duplicates

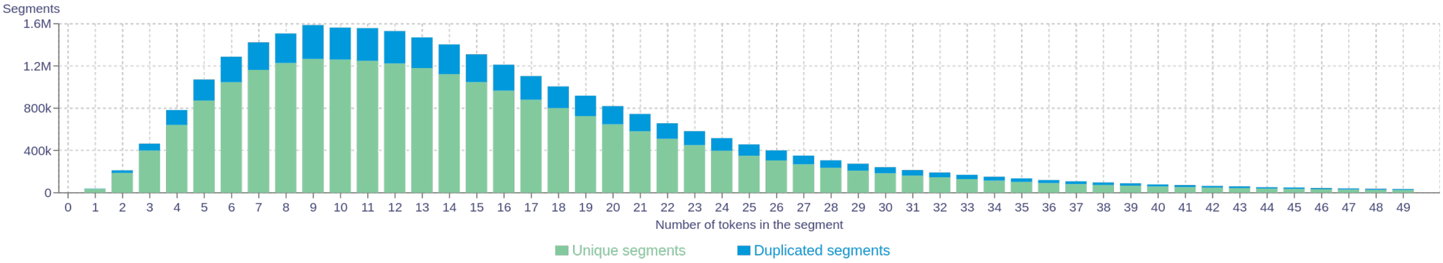
> 50 tokens = **1.2M** segments | **53K** duplicates



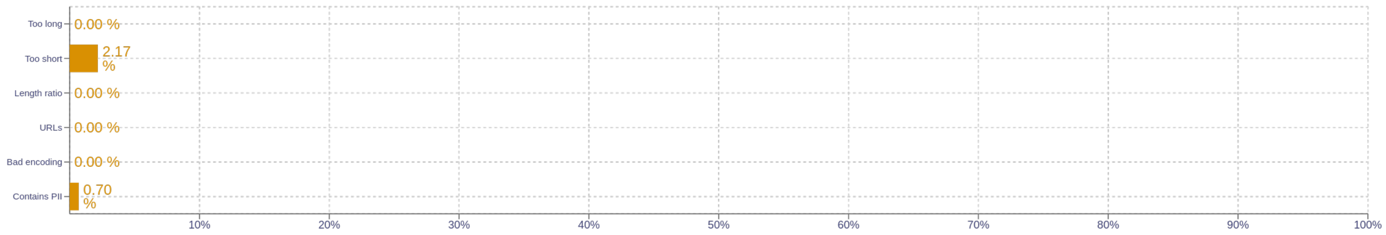
Target segment length distribution by token

<= 49 tokens = **23M** segments | **5.8M** duplicates

> 50 tokens = **459K** segments | **124K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	data | 1378384 also | 1321896 use | 1171037 information | 1117438 hotel | 1033436
2	personal data | 532909 personal information | 142211 air conditioning | 120418 privacy policy | 111089 member states | 102075
3	board of directors | 58877 protected from spambots | 57936 terms and conditions | 52551 processing of personal | 46852 hotel is within | 41383
4	address is being protected | 57986 processing of your personal | 49215 processing of personal data | 45967 wi-fi in public areas | 34651 wi-fi in all rooms | 30915
5	email address is being protected | 54691 processing of your personal data | 44102 tripadvisor is proud to partner | 40015 free wi-fi in all rooms | 30881 people looked at this hotel | 27168

Target n-grams

Size	n-grams
1	myös | 1524616 voit | 1186760 voi | 1028526 n | 976976 kaikki | 736703
2	muun muassa | 133078 kävelymatkan päässä | 103208 minuutin kävelymatkan | 88000 milloin tahansa | 82549 voit tehdä | 76849
3	minuutin kävelymatkan päässä | 86364 sähköpostiosoite on suojattu | 56896 yhteyttä tähän yritykseen | 41539 joten voit tehdä | 40426 voit tehdä varauksesi | 40110
4	sähköpostiosoite on suojattu spamboteilta | 56845 voit tehdä varauksesi kohteessa | 40083 joten voit tehdä varauksesi | 40083 hotellia viimeisen tunnin sisällä | 27471 euroopan parlamentin ja neuvoston | 24667
5	joten voit tehdä varauksesi kohteessa | 40083 tarkasteli tätä hotellia viimeisen tunnin | 18404 henkilöä tarkasteli tätä hotellia viimeisen | 18404 käyttäjää ovat kiinnostuneita tästä majoituksesta.alkaen | 16106 hotellissa on wifi-internetyhteys hotellivieraiden käytettävissä | 15974

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