HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-sk.tsv	1/28/2025	English (en)	Slovak (sk)

Volumes

Segments	SL tokens	SL characters	SL size
20,056,339	434M	2,264,798,988	2.12 GB

TL tokens	TL characters	TL size
388M	2,260,326,672	2.29 GB

Dataset top 10 domains

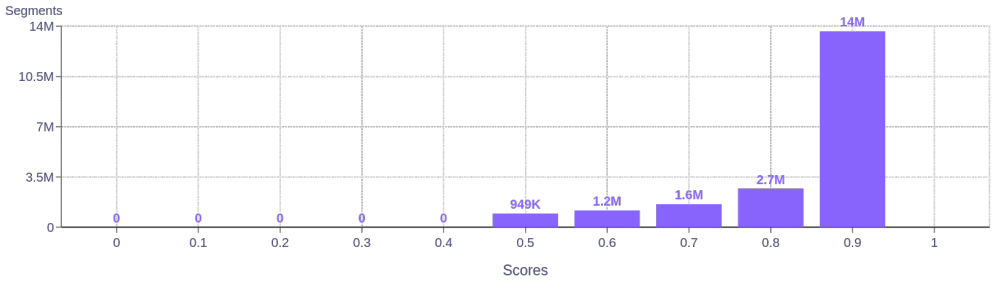
SL domain	Segments	TL domain	Segments
hotels.com	23.1%	hotels.com	8.9%
europa.eu	8.8%	europa.eu	6.9%
google.com	6.1%	google.com	2.9%
booking.com	3.9%	tripadvisor.sk	2.6%
microsoft.com	3.2%	alza.sk	2.5%
travelport.cz	3.1%	travelport.cz	2.5%
wikipedia.org	2.8%	wikipedia.org	2.4%
alza.co.uk	2.7%	booking.com	2.1%
office.com	1.6%	microsoft.com	2.0%
lacroixx.com	1.2%	office.com	1.4%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	105.6%	com	54.5%
eu	14.2%	sk	44.2%
org	10.6%	eu	10.8%
sk	8.5%	org	7.2%
co.uk	6.9%	cz	3.8%
cz	4.9%	net	3.2%
net	4.7%	info	1.6%
ie	2.3%	de	0.6%
de	2.2%	pl	0.4%
info	1.7%	ws	0.3%

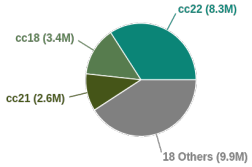
Translation likelihood

≥ 5 = 20M segments | **100.0%**
≥ 8 = 16M segments | **81.5%**
< 5 = 0 segments | **0.0%**



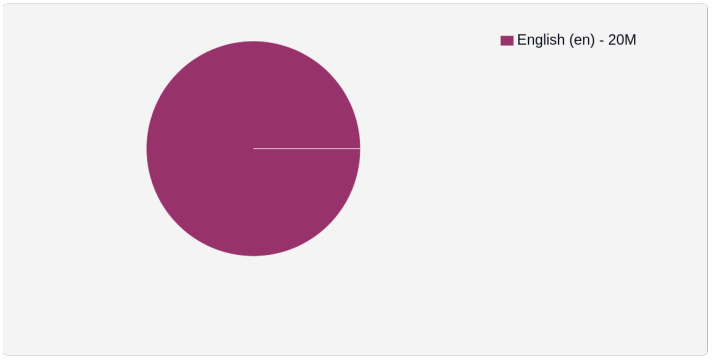
Collections

CC = 69.70%
IA = 30.30%

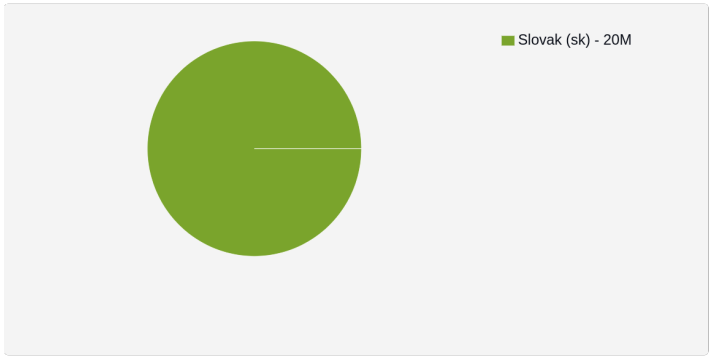


Language Distribution

Source

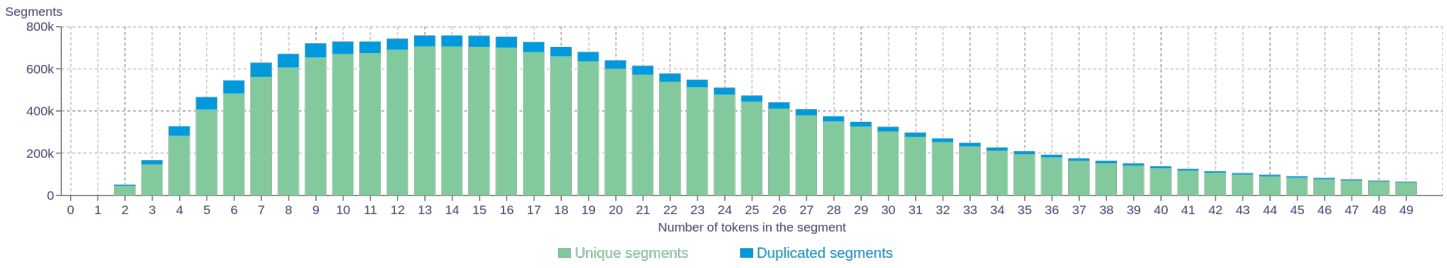


Target



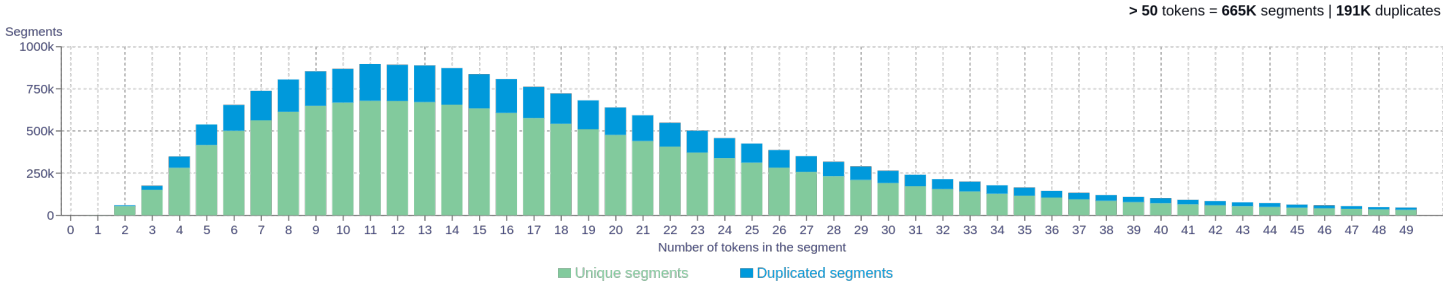
Source segment length distribution by token

<= 49 tokens = **18M** segments | **1.4M** duplicates
> 50 tokens = **989K** segments | **56K** duplicates

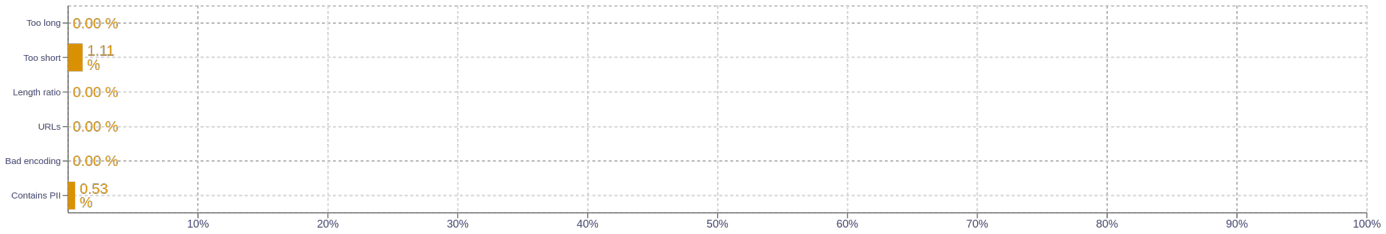


Target segment length distribution by token

<= 49 tokens = **15M** segments | **4.8M** duplicates
> 50 tokens = **665K** segments | **191K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	data | 1808580 · personal | 1068081 · use | 954987 · also | 877525 · information | 828995
2	personal data | 890676 · data protection | 123837 · data subject | 120589 · personal information | 104737 · privacy policy | 92778
3	processing of personal | 99504 · terms and conditions | 71135 · reservations with confidence | 61129 · proud to partner | 61114 · tripadvisor is proud | 61085
4	processing of personal data | 98494 · processing of your personal | 78168 · address is being protected | 48875 · use of the website | 41758 · process your personal data | 37343
5	processing of your personal data | 72685 · tripadvisor is proud to partner | 61085 · email address is being protected | 46129 · parliament and of the council | 28032 · proud to partner with booking.com | 22709

Target n-grams

Size	n-grams
1	údajov | 1086690 · osobných | 799844 · údaje | 747208 · môžete | 738525 · ste | 589520
2	osobných údajov | 776876 · osobné údaje | 484803 · vašich osobných | 153374 · dotknutá osoba | 118132 · webovej stránky | 112627
3	vašich osobných údajov | 149636 · ochrany osobných údajov | 84614 · ochrane osobných údajov | 66031 · vykonávať bezproblémové rezervácie | 61179 · bezproblémové rezervácie hotela | 61179
4	vykonávať bezproblémové rezervácie hotela | 61179 · ktorému budete môcť vykonávať | 61179 · uzatvorenie spolupráce s partnerom | 61125 · tripadvisor s hrdosťou ohlasuje | 61125 · spoločnosť tripadvisor s hrdosťou | 61125
5	vďaka ktorému budete môcť vykonávať | 61179 · ktorému budete môcť vykonávať bezproblémové | 61179 · tripadvisor s hrdosťou ohlasuje uzatvorenie | 61125 · spoločnosť tripadvisor s hrdosťou ohlasuje | 61125 · ohlasuje uzatvorenie spolupráce s partnerom | 61125

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