HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-heb_Hebr	9/18/2025	Hebrew

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
26,082,588	647,272,554	410,742,381 (63.46 %)	16B	78,486,012,143	129.13 GB

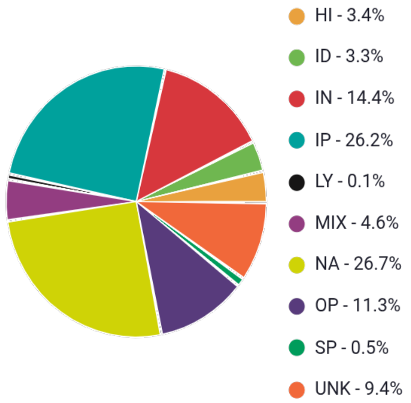
Top 10 domains

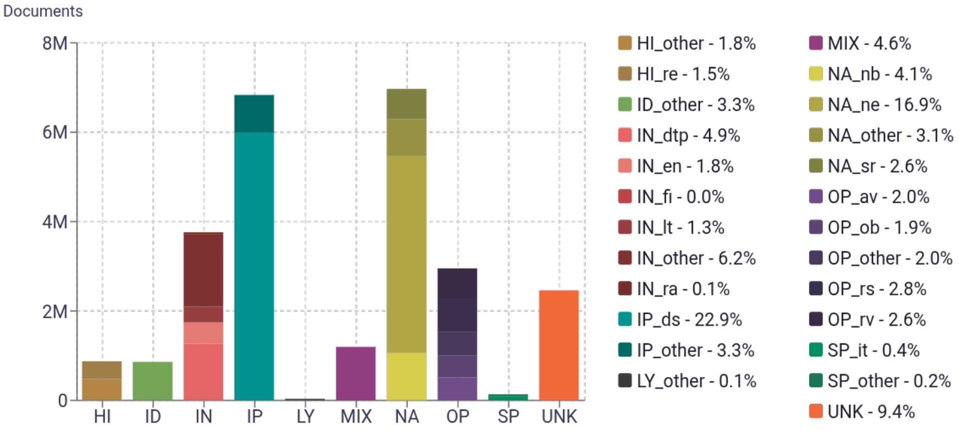
Domain	Docs	% of total
walla.co.il	931K	3.57%
ynet.co.il	387K	1.48%
wikipedia.org	347K	1.33%
haaretz.co.il	256K	0.98%
maariv.co.il	250K	0.96%
blogspot.com	228K	0.88%
mako.co.il	219K	0.84%
wordpress.com	218K	0.84%
psakdin.co.il	194K	0.74%
calcalist.co.il	180K	0.69%

Top 10 TLDs

Domain	Docs	% of total
co.il	15M	58.30%
com	5.4M	20.84%
org	1.6M	6.10%
org.il	1.4M	5.50%
net	806K	3.09%
ac.il	346K	1.32%
tv	145K	0.56%
info	118K	0.45%
muni.il	63K	0.24%
gov.il	52K	0.20%

Register labels

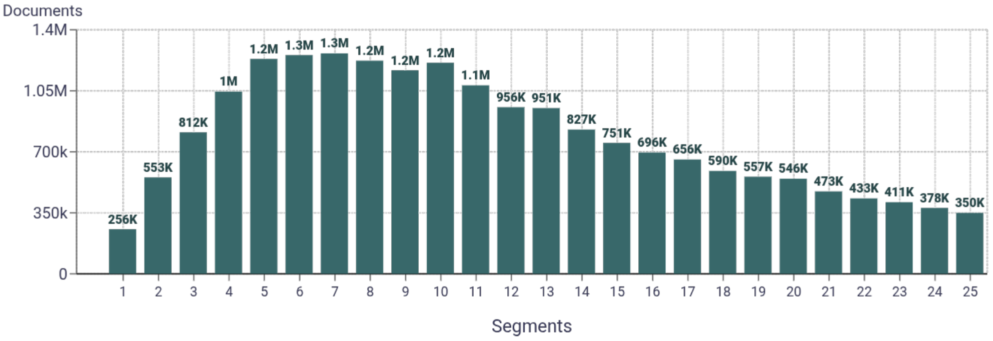




MT:6.5% | 1.7M Documents

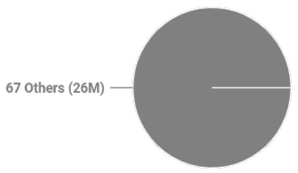
Documents size (in segments) ⓘ

≤ **25** segments **75.43%** (20M documents)
> **25** segments **24.57%** (6.4M documents)



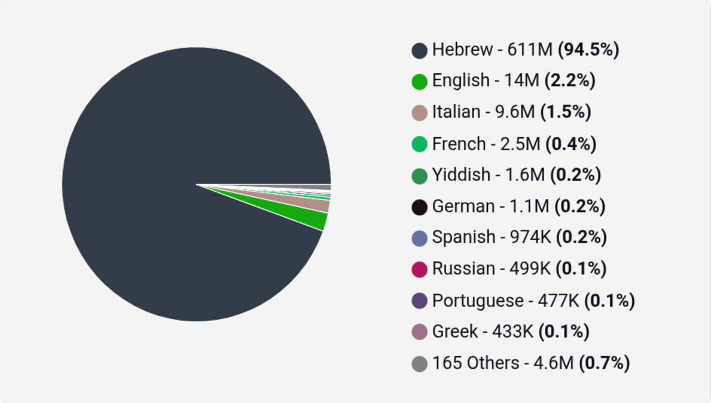
Document collections

CC = 88.07%
IA = 11.93%

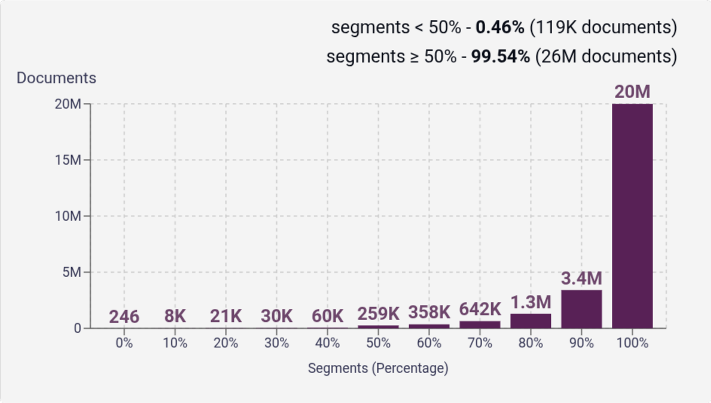


Language Distribution

Number of segments in the Hebrew corpus



Percentage of segments in Hebrew inside documents

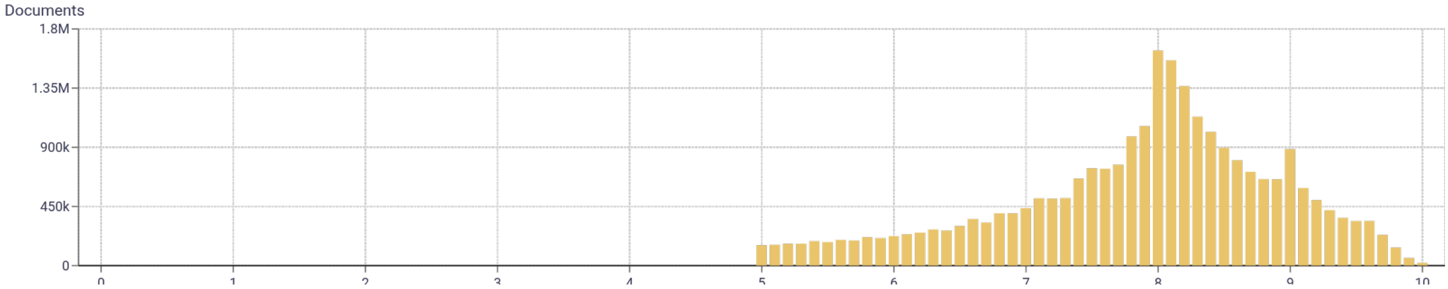

segments < 50% - **0.46%** (119K documents)

segments ≥ 50% - **99.54%** (26M documents)

Distribution of documents by document score

score < 5 - **0%** (0 documents)

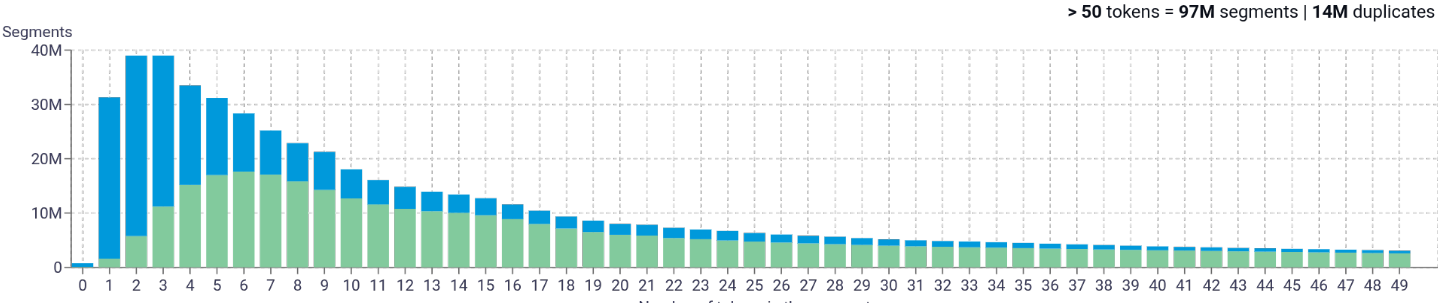
score ≥ 5 - **100%** (26M documents)



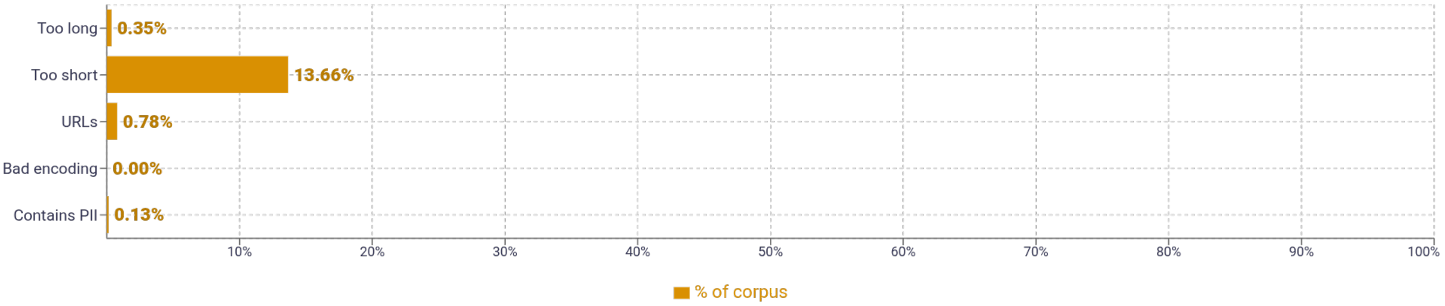
Segment length distribution by token

≤ **49** tokens = **551M** segments | **223M** duplicates

> **50** tokens = **97M** segments | **14M** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	26,307,297 | כדי 25,200,621 | ב 22,614,309 | אחד 20,435,743 | ה 18,650,872 | בכל
2	2,946,106 | תל אביב 2,509,445 | בית המשפט 2,056,397 | לאחר מכן 1,679,738 | בתל אביב 1,645,229 | חוות דעת
3	1,433,400 | בסופו של דבר 1,249,605 | עריכת קוד מקור 399,233 | בתי מלון ב 372,262 | בית המשפט העליון 316,527 | תגובתך נקלטה בהצלחה
4	316,278 | ותפורסם על פי מדיניות 162,066 | ראש הממשלה בנימין נתניהו 160,738 | נסה שנית במועד מאוחר 160,738 | אנא נסה שנית במועד 159,206 | התראה בדוא"ל כאשר תגובתך
5	316,278 | ותפורסם על פי מדיניות המערכת 160,737 | אנא נסה שנית במועד מאוחר 159,206 | לקבל התראה בדוא"ל כאשר תגובתך 159,206 | התראה בדוא"ל כאשר תגובתך תאושר 159,206 | בדוא"ל כאשר תגובתך תאושר ותפורסם

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

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	**How-to or instructions**	**HI**	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	**Informational persuasion**	**IP**	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	**Opinion**	**OP**
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	**Informational description**	**IN**	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				