HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-apc_Arab	10/3/2025	Levantine Arabic (apc)

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
253	3,494	3,044 (87.12 %)	12.88%	47K	234,940	413.79 KB

Top 10 domains

Domain	Docs	% of total
palestinerememb...	18	7.11%
art-en.com	11	4.35%
hellooha.com	9	3.56%
babycenter.com	7	2.77%
gidny.com	6	2.37%
fxsolve.com	5	1.98%
bukja.net	5	1.98%
wordpress.com	3	1.19%
nawaret.com	3	1.19%
ilcode.com	3	1.19%

Top 10 TLDs

Domain	Docs	% of total
com	187	73.91%
net	26	10.28%
org	13	5.14%
ps	3	1.19%
me	2	0.79%
co.il	2	0.79%
ws	1	0.40%
tv	1	0.40%
today	1	0.40%
tech	1	0.40%

Register labels

- HI - 2.0%
- ID - 14.2%
- IN - 4.3%
- IP - 14.2%
- LY - 15.0%
- MIX - 0.8%
- NA - 33.6%
- OP - 2.4%
- UNK - 13.4%

- HI_other - 0.4%
- HI_re - 1.6%
- ID_other - 14.2%
- IN_dtp - 1.2%
- IN_lt - 0.4%
- IN_other - 2.8%
- IP_ds - 10.7%
- IP_other - 3.6%
- LY_other - 15.0%
- MIX - 0.8%
- NA_nb - 1.2%
- NA_ne - 30.4%
- NA_other - 2.0%
- OP_ob - 1.2%
- OP_other - 1.2%
- UNK - 13.4%
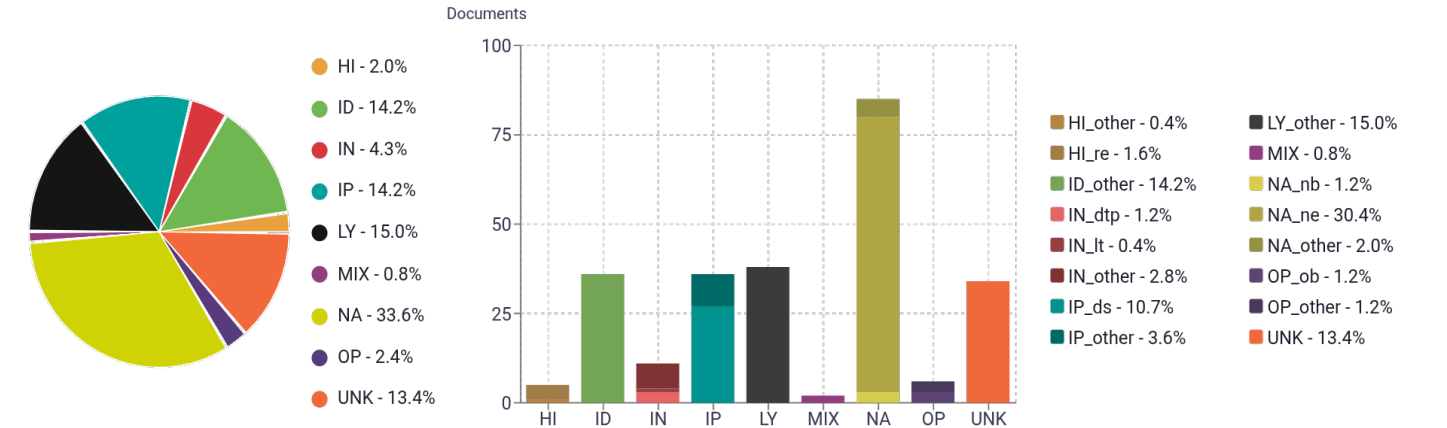
MT:1.2% | 3 Documents

Documents size (in segments) ⓘ

≤ 25 segments **85.77%** (217 documents)
> 25 segments **14.23%** (36 documents)

Document collections

CC = 71.94%
IA = 28.06%

61 Others (253)
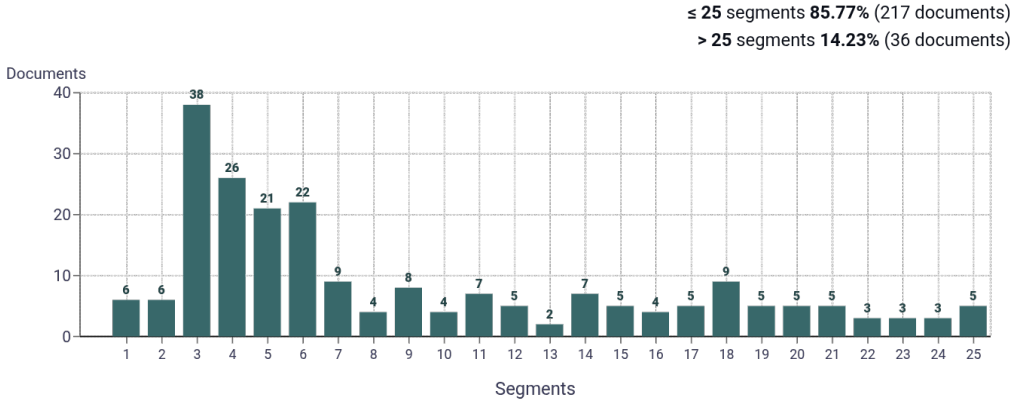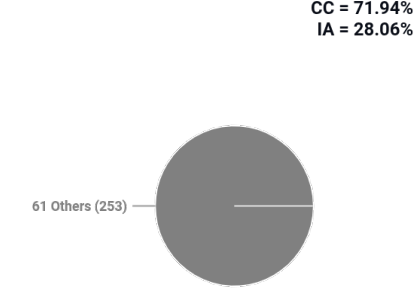
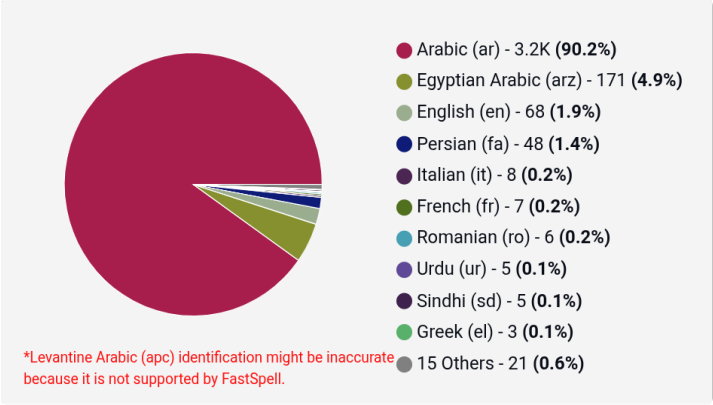## Language Distribution

Number of segments in the Levantine Arabic (apc) corpus

- Arabic (ar) - 3.2K **(90.2%)**
- Egyptian Arabic (arz) - 171 **(4.9%)**
- English (en) - 68 **(1.9%)**
- Persian (fa) - 48 **(1.4%)**
- Italian (it) - 8 **(0.2%)**
- French (fr) - 7 **(0.2%)**
- Romanian (ro) - 6 **(0.2%)**
- Urdu (ur) - 5 **(0.1%)**
- Sindhi (sd) - 5 **(0.1%)**
- Greek (el) - 3 **(0.1%)**
- 15 Others - 21 **(0.6%)**

*Levantine Arabic (apc) identification might be inaccurate because it is not supported by FastSpell.

Percentage of segments in Levantine Arabic (apc) inside documents
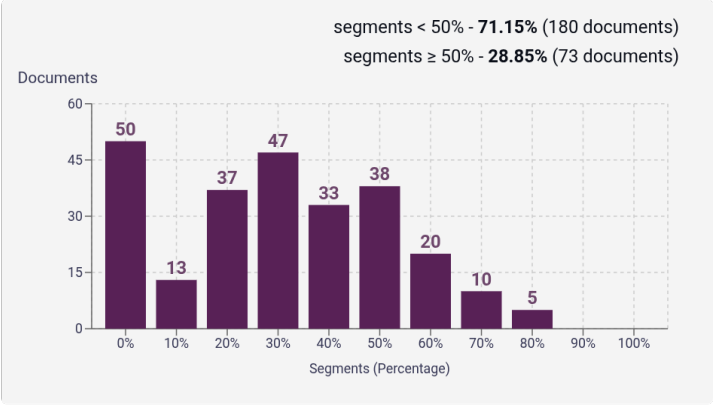
segments < 50% - **71.15%** (180 documents)
segments ≥ 50% - **28.85%** (73 documents)

Documents

Segments (Percentage)	
0%	50
10%	13
20%	37
30%	47
40%	33
50%	38
60%	20
70%	10
80%	5

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (253 documents)

Documents

Segment length distribution by token

≤ **49** tokens = **3.3K** segments | **440** duplicates
> **50** tokens = **204** segments | **10** duplicates

Segments

Segment noise distribution

	% of corpus
Too long	0.31%
Too short	11.02%
URLs	0.34%
Bad encoding	0.00%
Contains PII	0.00%

■ % of corpus
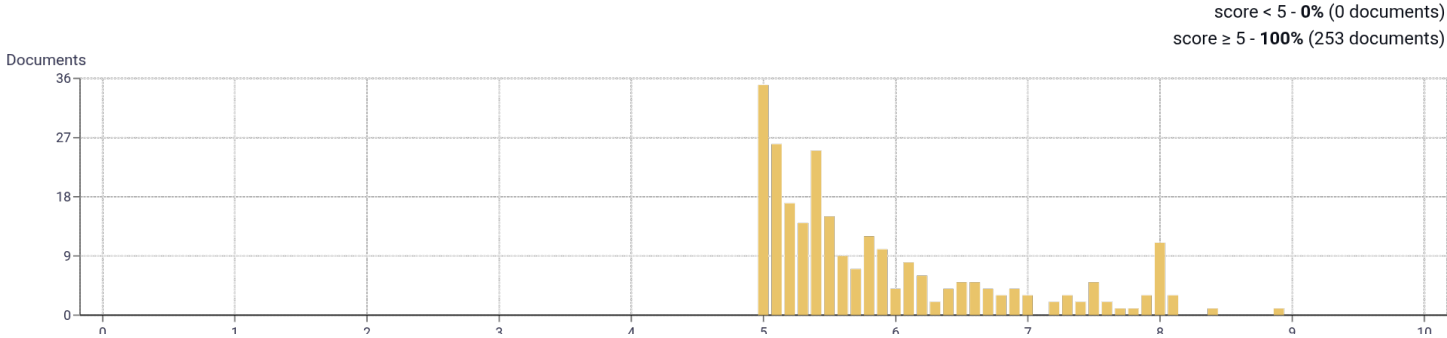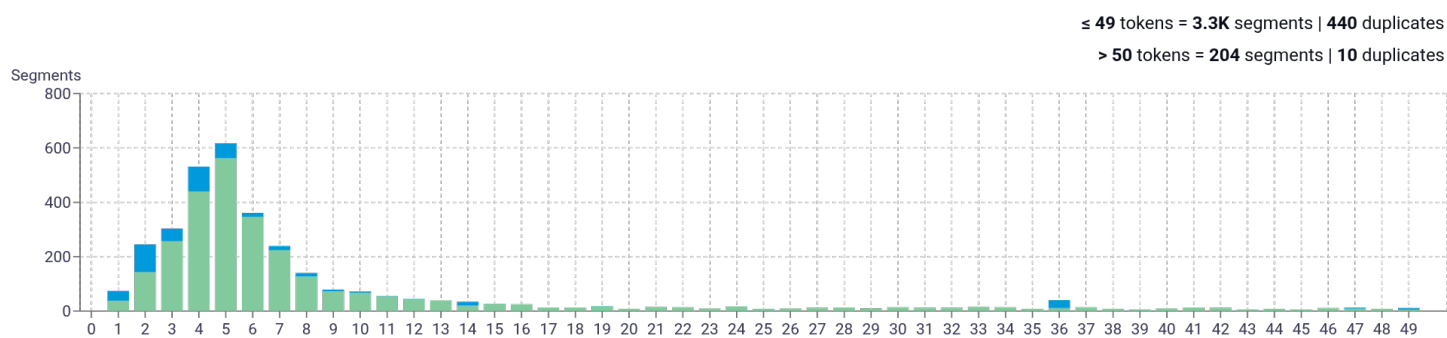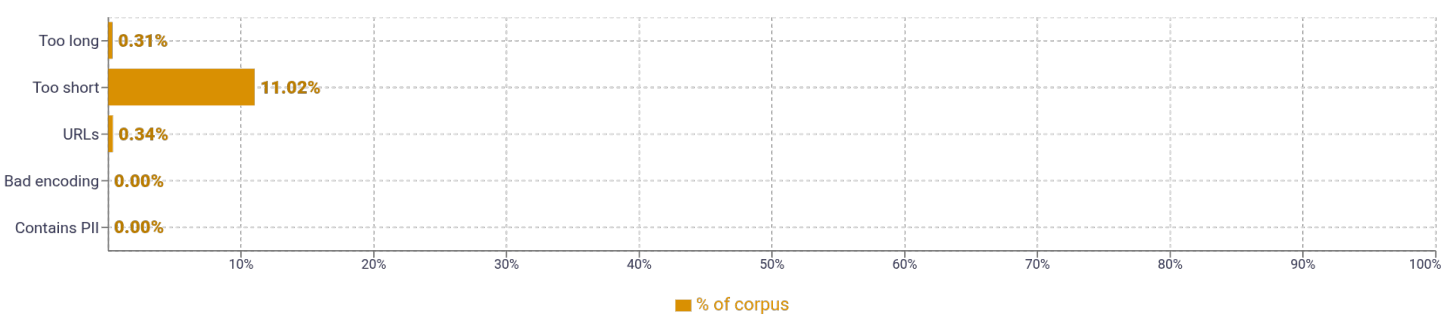
Frequent n-grams

SIZE	N-GRAMS
1	الله | 125 عم | 118 ان | 103 خبر | 95 علي | 94
2	اجمل تقرير | 40 لجهاز منزلي | 44 تقرير لجهاز | 51 تلميع سيارة | 67 خبر عاجل | 84
3	تقرير لجهاز منزلي | 44 المنتدى لا تستطيع | 44 اجمل تقرير لجهاز | 30 بعمولة عبر شركة | 23 التسويق بعمولة عبر | 23
4	اجمل تقرير لجهاز منزلي | 27 التسويق بعمولة عبر شركة | 23 مشاركتك في هذا المنتدى | 22 يعبر عن الرأي الشخصي | 18 مسؤولة عن هذه الآراء | 18
5	يعبر عن الرأي الشخصي لمؤلفها | 18 وفلسطين في الذاكرة غير مسؤولة | 18 صحة المعلومات ولكن لا تضمن | 18 تحاول فلسطين في الذاكرة التدقيق | 18 المعلومات ولكن لا تضمن صحتها | 18

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				