HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-ars_Arab	10/3/2025	Najdi Arabic (ars)

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
1,810	38,346	29,627 (77.26 %)	22.74%	525K	2,726,160	4.69 MB

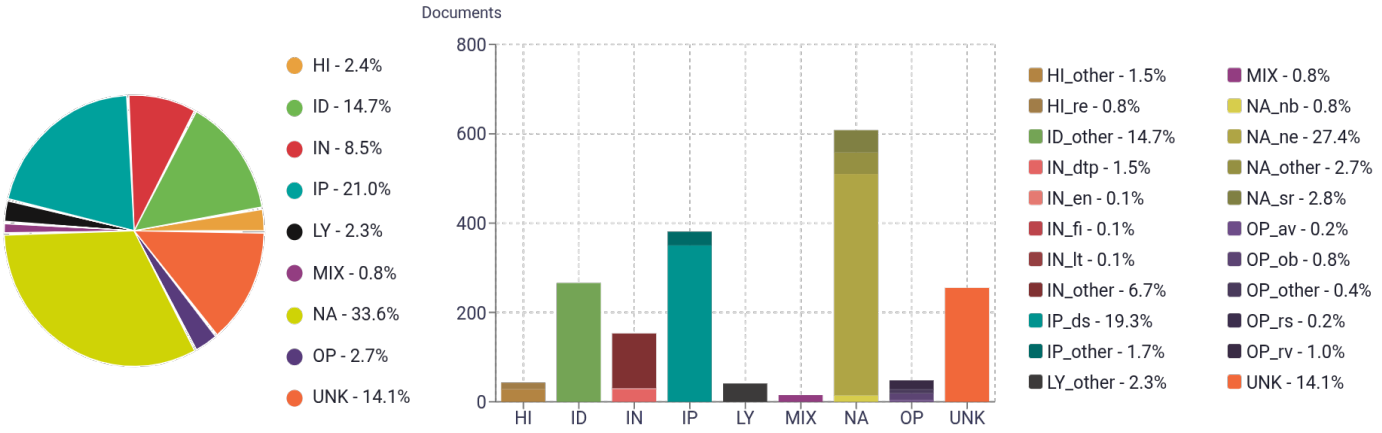
Top 10 domains

Domain	Docs	% of total
unlimit-tech.com	51	2.82%
otlobmehany.com	47	2.60%
alriyadh.com	31	1.71%
asir.me	29	1.60%
jobs-arab.com	28	1.55%
blogspot.com	27	1.49%
alweeam.com.sa	23	1.27%
a1ash.com	23	1.27%
mokhtsar.net	22	1.22%
dubaisession.com	22	1.22%

Top 10 TLDs

Domain	Docs	% of total
com	1.2K	68.84%
net	285	15.75%
org	63	3.48%
com.sa	41	2.27%
me	33	1.82%
info	19	1.05%
net.sa	12	0.66%
news	9	0.50%
co	9	0.50%
us	7	0.39%

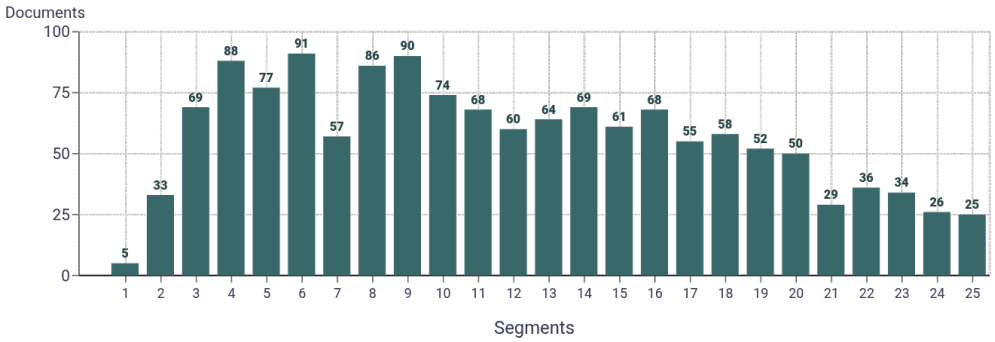
Register labels



MT:3.9% | 70 Documents

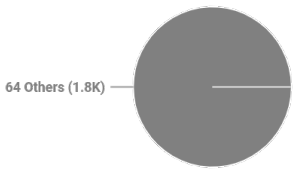
Documents size (in segments)

≤ 25 segments **78.73%** (1.4K documents)
> 25 segments **21.27%** (385 documents)



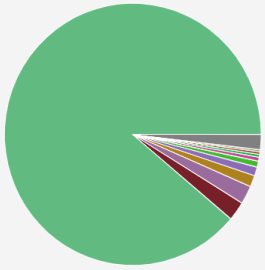
Document collections

CC = 72.60%
IA = 27.40%



Language Distribution

Number of segments in the Najdi Arabic (ars) corpus

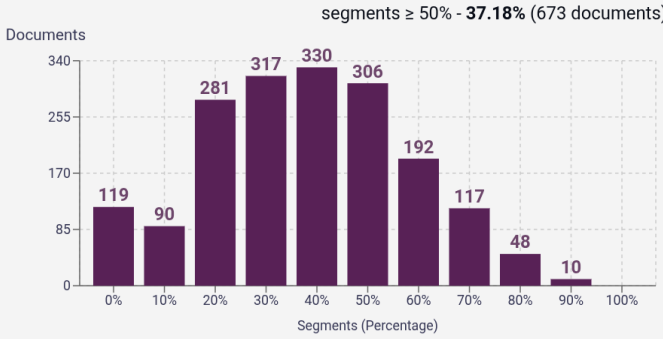


- Arabic (ar) - 34K **(88.9%)**
- English (en) - 907 **(2.4%)**
- Egyptian Arabic (arz) - 870 **(2.3%)**
- French (fr) - 542 **(1.4%)**
- Persian (fa) - 416 **(1.1%)**
- Italian (it) - 283 **(0.7%)**
- German (de) - 185 **(0.5%)**
- Central Kurdish (ckb) - 139 **(0.4%)**
- Korean (ko) - 113 **(0.3%)**
- Swedish (sv) - 97 **(0.3%)**
- 62 Others - 723 **(1.9%)**

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

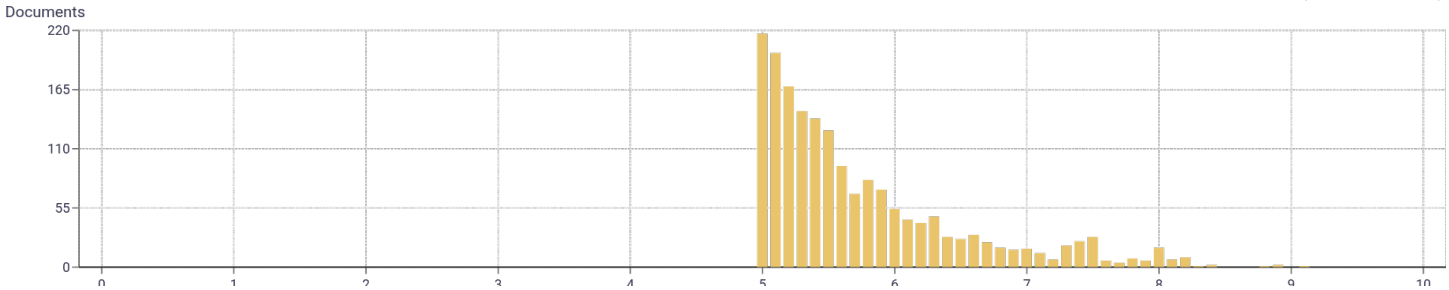
Percentage of segments in Najdi Arabic (ars) inside documents

segments < 50% - **62.82%** (1.1K documents)
segments ≥ 50% - **37.18%** (673 documents)



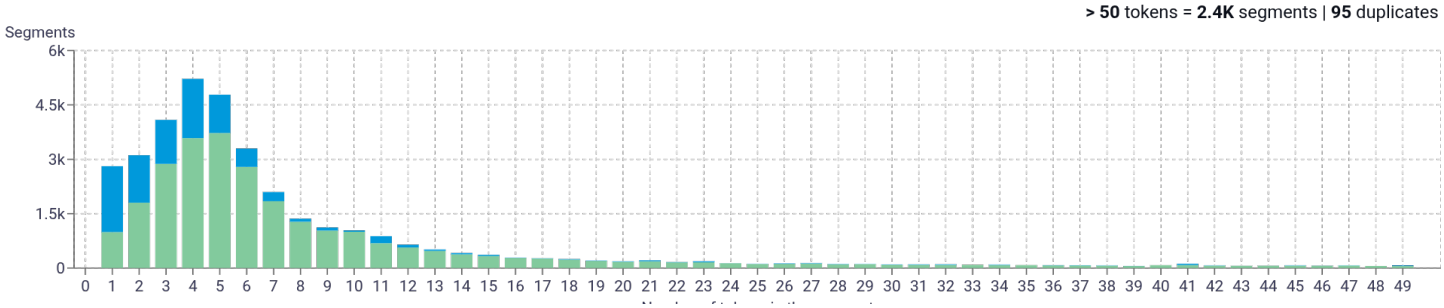
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (1.8K documents)

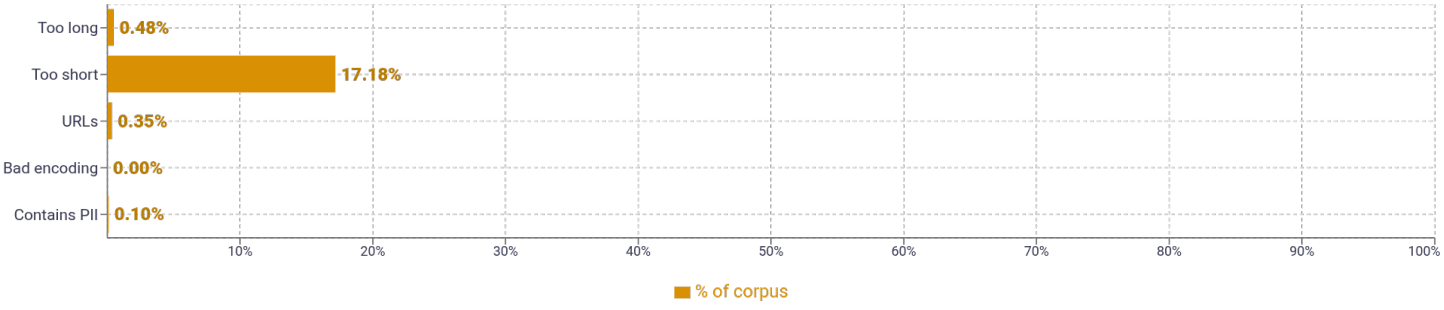


Segment length distribution by token

≤ 49 tokens = **36K** segments | **8.6K** duplicates
> 50 tokens = **2.4K** segments | **95** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	شركة | 3,622 · بالرياض | 3,044 · الله | 2,990 · تنظيف | 2,153 · الرياض | 1,781
2	شركة تنظيف | 1,649 · نقل عفش | 477 · بنك الراجحي | 425 · شركة نقل | 374 · يمن موبايل | 366
3	المملكة العربية السعودية | 247 · فتح في نافذة | 244 · نص نص نص | 236 · شركة نقل عفش | 234 · نقل عفش بالرياض | 222
4	فتح في نافذة جديدة | 244 · نص نص نص نص | 225 · شكراااا شكراااا شكراااا | 171 · شركة تنظيف فلل بالرياض | 110 · شركة تنظيف خزانات بالرياض | 96
5	نص نص نص نص نص | 215 · شكراااا شكراااا شكراااا شكراااا | 170 · فقط الأعضاء المسجلين والمفعلين يمكنهم | 64 · الأعضاء المسجلين والمفعلين يمكنهم رؤية | 64 · السلام عليكم ورحمة الله وبركاته | 52

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				