The Arabic E-Book Corpus

SND-ID: 2024-145. **Version**: 1. **DOI**: https://doi.org/10.5878/7rbh-gy93

Download data

corpus-html.zip *(223.53 MB)*

corpus-txt.zip *(196.54 MB)*

Associated documentation

metadata.tsv *(306.12 KB)*

README.md *(6.43 KB)*

README.md.pdf *(55.54 KB)*

Download all files

2024-145-1.zip *(~420.42 MB)*

Citation

Hallberg, A. (2024) The Arabic E-Book Corpus (Version 1) [Data set]. University of Gothenburg. Available at: https://doi.org/10.5878/7rbh-gy93

Alternative title

مدونة لغوية للكتب العربية الإلكترونية

Creator/Principal investigator(s)

Andreas Hallberg - University of Gothenburg, Department of Languages and Literatures

Research principal

University of Gothenburg - Department of Languages and Literatures

Description

The Arabic E-Book Corpus is a freely available collection of 1,745 books (81.5 million words) published in by the Hindawi foundation between 2008 and 2024. The books are of various genres, including non-fiction, novels, children's literature, poetry, and plays. The corpus is provided in two versions: html and unformatted plain text. The latter version will be appropriate for most purposes.

Data contains personal data

Yes

Type of personal data

The data container names of copyright holders, such as authors and translators, as well as historical, political, and other public figures mentioned in the works.

Language

Arabic

Time period(s) investigated

2008 – 2024

Data format / data structure

Text

Resource type

Corpus

Foreseen use

NLP application, Human use

Text corpus

- Linguality

  Monolingual
- Language
  - Arabic (ara)
- Modality

  Written Language
- Size

  Words: 80.5 million

  Files: 1,745
- Original source

  http://www.hindawi.org
- Link to other media

  Text: https://www.hindawi.org

Geographic spread

Geographic location: North Africa, The Middle East

Responsible department/unit

Department of Languages and Literatures

Research area

Language technology (computational linguistics) (Standard för svensk indelning av forskningsämnen 2011)

Specific languages (Standard för svensk indelning av forskningsämnen 2011)

Keywords

Corpus linguistics, Corpus-based research, Arabic language, Arabic alphabet

Accessibility level

Access to data through SND
Data are freely accessible

Use of data

Things to consider when using data shared through SND

License

CC BY 4.0

Versions

Version 1. 2024-12-11

Contact for questions about the data

Andreas Hallberg

andreas.hallberg@sprak.gu.se

This resource has the following relations

Compiles Hindawi

CLARIN Virtual Collection Registry

Add to collection

A virtual collection is connected to a specific research purpose and contains links to data resources from various digital archives. It is easy to create, access, and cite the collection.

Read more about virtual collections on the CLARIN website.

Download metadata

DataCite

MetaShare

MetaShare-CMDI

DDI 2.5

DDI 3.3

DCAT-AP-SE 2.0

JSON-LD

PDF

Citation (CSL)

File overview (CSV)

Published: 2024-12-11