

Arabisk e-bokskorpus

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

Ladda ner data

corpus-html.zip (223.53 MB)

corpus-txt.zip (196.54 MB)

Tillhörande dokumentation

metadata.tsv (306.12 KB)

README.md (6.43 KB)

README.md.pdf (55.54 KB)

Ladda ner alla filer

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

Citering

Hallberg, A. (2024) Arabisk e-bokskorpus (Version 1) [Dataset]. Göteborgs universitet. Tillgänglig via: <https://doi.org/10.5878/7rbh-gy93>

Alternativ titel

مدونة لغوية للكتب العربية الإلكترونية

Skapare/primärforskare

[Andreas Hallberg](#) - Göteborgs universitet, Institutionen för språk och litteraturer

Forskningshuvudman

[Göteborgs universitet](#) - Institutionen för språk och litteraturer

Beskrivning

Arabisk e-bokskorpus är en fritt tillgänglig samling av 1 745 böcker på arabiska, publicerade av Hindawi Foundation mellan 2008 och 2024. Böckerna är av olika genrer, bland annat, facktext, romaner, barnlitteratur, poesi och pjäser. Korpusen är tillgänglig i två versioner: html och icke-formaterad ren text. Den senare bäst lämpad för de flesta syften.

Data innefattar personuppgifter

Ja

Typ av personuppgifter

Datan innehåller namn på upphovsrättsinnehavare, såsom författare och översättare, samt namn på historiska, politiska eller andra offentliga personer som nämns i verken.

Språk

[Arabiska](#)

Tidsperiod(er) som undersökts

2008 – 2024

Dataformat / datastruktur

[Text](#)

Resurstyp

Korpus

Tänkt användning

Språkteknologiskt datorprogram, Mänsklig användning

Text corpus

- Antal språk
 - Enspråkig
- Språk
 - Arabiska (ara)
- Modalitet
 - Skriftspråk
- Storlek
 - Ord: 80.5 million
 - Filer: 1,745
- Källa
 - <http://www.hindawi.org>
- Länk till andra media
 - Text: <https://www.hindawi.org>

Geografisk utbredning

Geografisk plats: [Nordafrika](#), [Mellanöstern](#)

Ansvarig institution/enhet

Institutionen för språk och litteraturer

Forskningsområde

[Språkteknologi \(språkvetenskaplig databehandling\)](#) (Standard för svensk indelning av forskningsämnen 2011)

[Studier av enskilda språk](#) (Standard för svensk indelning av forskningsämnen 2011)

Nyckelord

[Korpuslingvistik](#), [Korpusundersökning](#), [Arabiska](#), [Arabiska alfabetet](#)

Tillgänglighetsnivå

Åtkomst till data via SND

Data är fritt tillgängliga

Användning av data

[Att tänka på vid användning av data som delas via SND](#)

Licens

[CC BY 4.0](#)

Versioner

Version 1. 2024-12-11

Kontakt för frågor om data

Andreas Hallberg

andreas.hallberg@sprak.gu.se

Denna resurs har följande relationer

Sammanställer [Hindawi](#)

CLARIN Virtual Collection Registry

[Lägg till i samling](#)

En virtuell samling är kopplad till ett specifikt forskningsändamål och innehåller länkar till dataresurser i olika digitala arkiv. Samlingen är lätt att skapa, få åtkomst till och citera.

Read more about virtual collections on the [CLARIN website](#).

Ladda ner metadata

[DataCite](#)

[MetaShare](#)

[MetaShare-CMDI](#)

[DDI 2.5](#)

[DDI 3.3](#)

[DCAT-AP-SE 2.0](#)

[JSON-LD](#)

[PDF](#)

[Citation \(CSL\)](#)

[Filöversikt \(CSV\)](#)

Publicerad: 2024-12-11