

Importance of Language Corpora in the Teaching and Learning of Arabic: A Descriptive, Analytical, and Evaluative Study of Selected Models

Dr Amel Ourabah¹

Zineb Boudaoud²

Meriem Guettia³

Amrouche Fatima⁴

¹University of Algiers 2- Algeria

^{2,3,4}University of Tipaza – Algeria^{2,3,4}

¹amel.ourabah@univ-alger2.dz

²z-boudaoud@hotmail.com

³guettia.meriem@univ-tipaza.dz

⁴amrouche.fatima@univ-tipaza.dz

Received:11/09/2025

Accepted: 02/02/2026

Published: 20/04/2026

Abstract

This article highlights the importance of language corpora and their role in the teaching and learning of Arabic across linguistic and literary domains and across various modules. It does so by examining selected representative models through description, analysis, and evaluation to assess their effectiveness and efficiency in meeting, first, the requirements of the Arabic language and, second, the needs of learners. This study attempts to address the following research question: Do specialised language corpora in the Arabic language and literature provide added scholarly value for university students and learners? Through this investigation, the study aims to affirm the role of such corpora in enhancing the success of the teaching and learning process.

Keywords: Language corpora; Arabic language; role; teaching and learning; description, analysis, and evaluation.

Introduction:

Language corpora are of considerable importance and received notable attention from early Arab linguists because of the need to compile all aspects of Arabic, including its vocabulary, into a single corpus. Early Arab linguists established a linguistic corpus by formulating grammatical rules based on the principle of oral transmission, as well as by documenting eloquent vocabulary and texts drawn from the Noble Quran, the Prophetic Hadith, and the speech of the Arabs in both poetry and prose. This was undertaken to preserve the Arabic language from error, corruption of speech, and the decline of eloquence and clarity. In doing so, they relied on the principles of audition, collection, and induction, gathering linguistic material from Arab tribes renowned for their eloquence and purity of expression. This rendered their work systematic and coherent through the scientific material they compiled and organised into a series of works, now referred to as “corpora,” such as *Kitāb al-‘Ayn* by al-

Khalīl ibn Aḥmad al-Farāhīdī, *Lisān al- 'Arab* by Ibn Manẓūr, *Gharīb al-Quran* by Ibn Qutaybah, and others.

The need for such corpora persists to the present day; indeed, it has become increasingly necessary in light of the notable developments the language is undergoing in the age of the internet and artificial intelligence. Consequently, interest in their use in the teaching and learning of Arabic has increased for both native and nonnative learners. Moreover, a focus on corpus linguistics, which specialists regard as a modern research methodology for language instruction across various academic levels, has emerged. It provides users with high accuracy, rapid analysis, and the ability to handle large volumes of data quickly.

In recent years, language corpora have attracted considerable attention owing to scientific and technological advancements, as well as the increasing use of computers for diverse research, cognitive, and scientific purposes. Specialists and researchers in corpus linguistics have likewise emphasised the value of their use in educational contexts, given the facilitation they offer in both teaching and learning.

From this standpoint, this article seeks to address the following research problem:

Do specialised language corpora in the Arabic language and literature provide added scholarly value for university students and learners of Arabic?

This study aims to demonstrate their effectiveness in ensuring the success of the teaching and learning process by describing and analysing selected corpus models, followed by their evaluation.

1. Concept of the Language Corpus:

The term “language corpus” is a compound consisting of two elements: “corpus” and “linguistic”. The term “corpus” is derived from the root *d-w-n*, which denotes a large collection of texts compiled on the basis of a particular language in use; it may pertain to various fields and disciplines and hence the designation “language corpus.”¹

Language is defined as “sounds by which each people express their purposes,”² serving as a means of communication, interaction, and mutual understanding among individuals who share a common linguistic affiliation.

Hence, specialists define it as “a collection of linguistic units that have been gathered and organised according to clear design criteria for the purpose of being used as a sample of language.”³ Another definition describes it as “a vast quantity of machine-readable texts, whether in written or spoken form, through which the use of language in natural contexts is

¹ Ibn Manzur, Jamal al Din. *Lisān al 'Arab*. Beirut: Dar Sadir, 3rd ed., 1414 AH, vol. 17, 1461.

² Ibn Jinni. *al Khasa'is*. Vol. 1, 35.

³ Al Fifi, Abd Allah ibn Yahya. "Wasm al Mudawwanat al Lughawiyya: al Maḥmūm wa al Majalat." *Mu'ta Journal for Research and Studies, Humanities and Social Sciences Series* 38, no. 1 (2023): 278.

manifested.” Thus, a language corpus is a large body of linguistic texts collected from diverse sources such as ,books, journals, radio and television broadcasts, bulletins, and periodicals , then organised, indexed, and stored electronically to enable users to exploit them for various purposes, including citation, the training of language-processing software, and other computational applications.⁴

2. Types of Language Corpora and Their Importance:

Arabic language corpora are divided into several types, some of which are determined by the field of knowledge or the target group, and others by the intended purpose of their compilation. These types include the following:⁵

- experimental language corpora
- research language corpora
- General language corpora
- specialised language corpora
- historical language corpora
- monitor corpora
- educational language corpora
- learner corpora
- pedagogical or instructional language corpora

They are also utilised across a range of knowledge domains, including the following:⁶

- linguistics
- Language teaching and learning
- sociolinguistics
- lexicography
- translation
- the study of intellectual trends
- information science
- bioinformatics and forensic science, among others.⁷

3. Criteria for the Design of Language Corpora:

The construction of language corpora is based on a set of criteria that constitute essential conditions for their development, namely, the following:

⁴ Qawasimiyya, Warda. “Dawr Mudawwanat al Lisaniyyat al Hasubiyya fi Khidmat al Lugha al ‘Arabiyya: Tahlil wa Tatbiq ‘ala al Mutun al Lughawiyya.” *Majallat Ma‘alim* 18, no. 1 (First Semester 2025): 160.

⁵ al Dukruri, Ayman. *al Mudawwanat al Lughawiyya wa Dawruha fi Mu‘alajat al Nusus al ‘Arabiyya*. Riyadh: King Salman Global Academy for Arabic Language, 1445 AH, 57 and following.

⁶ al Dukruri, Ayman. *al Mudawwanat al Lughawiyya wa Dawruha fi Mu‘alajat al Nusus al ‘Arabiyya*. Riyadh: King Salman Global Academy for Arabic Language, 1445 AH, 57 and following.

⁷ Ibid., 38 and following.

- identifying the target users whose texts will be included in the corpus
- the type of linguistic material
- the temporal and spatial scope of the corpus
- the size of the corpus
- The methodology of data collection
- the tagging mechanism and the type of annotation

Taken together, these conditions ,according to Skinner ,constitute a successful corpus capable of achieving its intended objectives.⁸

4. The Importance of Language Corpora

Language corpora, both general and specialised, are of considerable importance to their users. Their significance is reflected in the following:

- They serve as tools for linguistic statistical research⁹
- They contribute to the construction of dictionaries
- , assist in the development of language skills
- and are utilised in various applications, such as morphology, lexicon, syntax, semantics, pragmatics, and discourse analysis¹⁰
- They have an impact on linguistic attainment among university students in general and among learners in particular

5. Learner Corpus:

Before presenting selected models of corpora available in the Arabic language and literature, it is necessary to first provide a brief overview of learner corpora. Regardless of academic level, these are among the most widely used types and are reliant upon by both native and nonnative Arabic speakers. They differ from other corpora in the purposes for which learners employ them and in the extent to which their design aligns with learners' needs.

From this perspective, learner corpora adopt a specific methodology that differs somewhat from other language corpora because of the particular nature of this group. They are characterised by the following:

- the source of data
- their computational nature, in that they are stored electronically in a manner that allows texts to be machine-readable

⁸ Sinclair, John. "Corpus and Text: Basic Principles." In Wynne, ed., 3.

⁹ Sha'nabi, Nur al Din; 'Amamra, Kamal; and 'Idda, Bashir. "Ahamiyyat al Mudawwanat al Lughawiyya fi al Bahth al Ihsa'i al Lughawi: Mudawwanat al Lugh al 'Arabiyya Namudhajan." *Majallat Jusur al Ma'rifa li al Ta'limiyya wa al Dirasat al Lughawiyya wa al Adabiyya* 11, no. 1 (31 March 2025). University of Hassiba Ben Bouali, Chlef, 85–97.

¹⁰ See al Shahri, Ali Gharamah. "Lisaniyyat al Mudawwanat: Ta'rif wa Majalat Ifada." *al Majalla al 'Arabiyya li al 'Ulum wa Nashr al Abhath* 8, no. 3 (30 September 2022): 81 and following.

- their authenticity, as they are produced in natural, nonartificial contexts
- their purpose, which is clearly defined as the acquisition and teaching of language¹¹

For example, the corpus of Arabic learners provides valuable services to students through the information it offers, including the following:¹²

- corpus composition: 76 per cent handwritten texts, 17 per cent computer-written texts, and 7 per cent audio recordings
- student gender: 67 per cent male and 33 per cent female

It contains diverse texts from various contexts and includes 282,732 words across 1,585 written and spoken texts produced by 932 learners of Arabic in the Kingdom of Saudi Arabia. It is also supported by a team of specialised professors and researchers who ensure the provision of information and facilitate its comprehension and acquisition.

- Data collection: corpus data were collected from 2012–2013
- Number of texts: The corpus comprises a total of 1,585 texts
- user types: 53 per cent native speakers of Arabic and 47 per cent nonnative speakers
- Academic level of the target group: 80 per cent preuniversity and 20 per cent university level
- number of languages: The corpus includes a large number of languages, including Arabic, which comprises 790 texts and 150,938 words

This clearly demonstrates the effort exerted by those responsible for the corpus and the objectives they have set to advance teaching and learning.

6. An Analytical Study of Selected Models:

Given the vast number of corpora, this study presents selected models of language corpora available for teaching and learning at the university level to demonstrate their role and the importance of relying on them to develop intellectual and scholarly skills.

6.1. Arabic Poetry Corpus:

This corpus consists of poetic encyclopaedias dedicated to the presentation of Arabic poetry. It includes thousands of poems spanning from the pre-Islamic period to the modern era. The platform provides statistics on user engagement and utilisation, ranked as follows:

- Abbasid period: 2,147
- modern period: 1,488

¹¹ Al Fifi, Abd Allah; Mahmud Ismail Salih; Aqil Hamid al Shammari; and Sultan Nasir al Majyul. *al Mudawwanat al Lughawiyya al 'Arabiyya: Bina'uha wa Tara'iq al Istifada Minha*. Edited by Salih ibn Fahd al Usaymi. 1st ed. Riyadh, 1436 AH/2015, 97 and following.

¹² See <https://www.arabiclearncorpus.com/home-ar>

- Andalusian period: 1,108
- Umayyad period: 1,025
- pre-Islamic poetry: 896
- Mamluk period: 804
- Ottoman period: 348
- early Islamic period: 283
- Mu‘allaqāt: 10¹³

This clearly indicates that users of this corpus prefer the Abbasid period, owing to its richness and eloquence in poetry. The corpus provides students with comprehensive access to Arabic poetry.

6.2. The Syntax and Morphology Corpus

Within the grammar module, several language corpora are available, each designed for a specific category of learners. Among these, the corpus “Ta‘allum al-‘Arabiyya” was selected. This corpus presents grammar lessons based on the *Ājurrūmiyya*, a well-known grammatical text familiar to all students specialising in Arabic. It offers grammatical content at an elementary-intermediate level across thirty-three lessons.

The material begins with an introduction to the importance of grammar, followed by a series of independent lessons, including the following:

[speech—inflection—markers of the nominative—markers of the accusative—markers of the genitive—markers of the jussive—verbs—jussive of the imperfect—subject—agent substitute—subject and predicate—*kāna* and its sisters—*inna* and its sisters—*ẓanna* and its sisters—indefinite and definite—adjective—coordination—emphasis—apposition—direct object—absolute object—adverbial—circumstantial accusative—specification—exception—*lā* of absolute negation—vocative—object of purpose—object of accompaniment—genitives—construct state]

The Arabic grammar in this corpus is presented in two formats:

- written
- audiovisual, in the form of video

These two modes enable learners to select the option that best suits their comprehension and learning.

By way of illustration, the lesson on the subject (*fā‘il*) defines it as a noun indicating the doer of an action or one characterised by it, for example, *jā’a Zayd* (“Zayd came”)/*ikhḍarrat al-arḍ* (“the land became green”). It then presents and explains its grammatical markers, noting that the nominative case in nouns is indicated by the ḍamma, wāw, and alif, as in *fāza Muḥammad*

¹³ See <https://poetry.coiod.com/>

(“Muḥammad succeeded”), *najāḥa al-mujtahidūn* (“the diligent students succeeded”), and *ta’āwana al-muwazzifān* (“the two employees cooperated”). Its position in the sentence is also explained: it follows the verb and precedes the object, as illustrated by the example *kataba al-ṣaḥafī taqrīran* (“the journalist wrote a report”). The forms of the subject are then presented, indicating that it may occur as an explicit noun, as in *jā’a Zayd*, or as an implicit noun (a pronoun), as in *qara’ tukiṭāban* (“I read a book”).

Notably, the written corpus provides an English equivalent for each grammar lesson, thereby facilitating comprehension for nonnative Arabic speakers, including the translation of all written content into English.¹⁴

6.3. Prosody Corpus:

This corpus is concerned with the science of prosody and the metres of Arabic poetry. In this context, several corpora may be identified, including the corpus of Ibrāhīm Ḥasan Muḥammad, which takes the form of a series of prosody lessons. It provides students with all the necessary information about prosody, including its concept, origin, causes of emergence, and purpose. It also introduces the metres of Arabic poetry by their names and patterns, beginning with *al-sarī’*, *al-khafīf*, *al-rajaz*, *al-basīṭ*, *al-ṭawīl*, *al-wāfir*, *al-mutaqārib*, *al-kāmil*, *al-munsariḥ*, *al-mujtathth*, *al-ramal*, *al-hazaj*, *al-madīd*, *al-muḍāri’*, *al-muqṭaḍab*, *al-muḥdath*, and *al-khabab*.

At the bottom of the corpus, the titles of the metres are presented individually for those wishing to explore each metre in detail by selecting its name. Accessing the detailed presentation of a given metre provides a comprehensive overview, including its definition, metrical pattern, governing rules, forms, and types, followed by poetic examples analysed through prosodic scansion, along with indications of the modifications (*ziḥāfāt*) and irregularities (*‘ilal*) that may affect it.¹⁵

In addition, the corpus of AbūQudāma al-Miṣrī presents comprehensive summaries of prosody and rhyme in the form of independent lessons. An example is the fourth lesson, devoted to *ziḥāfāt*, in which the author begins with an introduction emphasising the importance of acquiring knowledge of prosody for every student before engaging in the study of Arabic poetic metres. He then refers to the ten metrical feet and the changes that may occur to them, whether through the omission, addition, or modification of letters. This is followed by a definition of *ziḥāfāt*, accompanied by illustrative examples, and an explanation of their characteristics and types, supported by examples clarifying each type in turn. The lesson concludes with notes and observations beneficial to students in the literature, providing them with valuable knowledge in this field.¹⁶

¹⁴ See <https://learning.aljazeera.net/ar/generallanguage->

¹⁵ See <https://3arod.wordpress.com/>

¹⁶ See https://abuqudamah.com/talkheesat_shafeyah_4

Another corpus is that of Abū Zārī al-Madanī, in which prosody lessons are presented in successive stages:

- The first stage consists of returning to one of the most important works in prosody, *Mizān al-Dhahab* by al-Sayyid al-Hāshimī, through careful reading, followed by the completion of the required exercises and the full scansion of poetic verses
- The second stage involves the memorisation of the didactic poem *Mujaddid al-‘Awāfī min Rasm al-‘Arūd wa al-Qawāfī* by ‘Abd Allāh al-‘Alawī al-Shinqīṭī, as well as the didactic poem of al-Ṣabbān and the text of *al-Kāfiyya*¹⁷

6.4. The Semantics Corpus:

This corpus, developed by Dr Monji al-‘Amrī, is intended for third-year students in the Department of Arabic Language at the Faculty of Arts and Humanities in Kairouan and at the Higher Institute of Applied Studies in Humanities in Gafsa. It aims to assist students in understanding issues in semantics and engaging with its questions through interaction with their enquiries. It also provides links to download a wide range of electronic linguistic references. Among the references it offers are *Language and Logic* by Ḥassān al-Bāhī; *An Introduction to Cognitive Grammar*; *In Cognitive Semantics*; *An Introduction to Modern Semantics*; *The Structure of the Arabic Language*; and *Cognitive Linguistic Theories*, among others, in addition to offering lessons in semantics.¹⁸

Another corpus is that of Dr Hisham Sa‘d al-Dīn, which is devoted to semantics and discourse analysis. It provides an extensive list of books on language, all of which are beneficial for students and learners, including *Dalā’il al-I’jāz*; *Language* by Vendryes; *Lectures and Applications in Semantics*; *Semantics in Arabic Heritage and Linguistic Study*; *Grammar and Meaning*; *Text, Discourse, and Procedure*; and *Discourse and Argumentation*.¹⁹

Conclusion:

On the basis of the foregoing discussion of language corpora in general and educational corpora in particular, as illustrated through the selected models, the following conclusions may be drawn:

- Educational language corpora are important and valuable within the educational system, as they provide students with numerous services, including vocabulary, grammar, and texts, which greatly help learners, particularly Arabic learners, overcome the difficulties they face. Owing to the precise language they offer, they enable students to learn Arabic autonomously through participation in information exchange and interaction with corpus content, involving knowledgesharing, discussion, observation, and inference, whether the content is written or audiovisual.

¹⁷ See https://abuzare.blogspot.com/2015/03/blog-post_12.html

¹⁸ See <http://semantique3.blogspot.com/>

¹⁹ See <http://saadeddine-semandisc.blogspot.com/>

- Educational language corpora are beneficial to students and provide them with all that they need to know and learn, whether they are native or nonnative speakers. Their use and reliance on them are advantageous for learning and the development of scholarly and intellectual skills.

Hence, the selected models demonstrate that such corpora effectively meet the needs of students and learners. Moreover, it remains incumbent upon learners to select the type of corpus that best fulfils their requirements in terms of comprehension, acquisition, and benefit.

Bibliography

al Dukruri, Ayman. *al Mudawwanat al Lughawiyya wa Dawruha fi Mu'alajat al Nusus al 'Arabiyya*. Riyadh: King Salman Global Academy for Arabic Language, 1445 AH.

Al Fifi, Abd Allah ibn Yahya. "Wasm al Mudawwanat al Lughawiyya: al Mafhum wa al Majalat." *Mu'ta Journal for Research and Studies*, Humanities and Social Sciences Series 38, no. 1 2023.

Al Fifi, Abd Allah; Mahmud Ismail Salih; Aqil Hamid al Shammari; and Sultan Nasir al Majyul. *al Mudawwanat al Lughawiyya al 'Arabiyya: Bina'uha wa Tara'iq al Istifada Minha*. Edited by Salih ibn Fahd al Usaymi. 1st ed. Riyadh, 1436 AH/2015.

Al Shahri, Ali Gharamah. "Lisaniyyat al Mudawwanat: Ta'rif wa Majalat Ifada." *al Majalla al 'Arabiyya li al 'Ulum wa Nashr al Abhath* 8, no. 3 (30 September 2022).

Ibn Jinni. *al Khasa'is*. Edited by Muhammad Ali al Najjar. Cairo: Egyptian General Book Authority.

Ibn Manzur, Jamal al Din. *Lisān al 'Arab*. 3rd ed. Beirut: Dar Sadir, 1414 AH.

Qawasimiyya, Warda. "Dawr Mudawwanat al Lisaniyyat al Hasubiyya fi Khidmat al Lughā al 'Arabiyya: Tahlil wa Tatbiq 'ala al Mutun al Lughawiyya." *Majallat Ma'alim* 18, no. 1 (First Semester 2025).

Sinclair, John. "Corpus and Text: Basic Principles." In Wynne, ed.

Websites

<https://3arod.wordpress.com/>

https://abuqudamah.com/talkheesat_shafeyah_4

https://abuzare.blogspot.com/2015/03/blog-post_12.html

<https://learning.aljazeera.net/ar/generallanguage->

<https://poetry.coiod.com/>

<https://www.arabiclearnercorpus.com/home-ar>

<http://semantique3.blogspot.com/>

<http://saadeddine-semandisc.blogspot.com/>