International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Using Corpora in the Saudi Arabian Classroom

Ibrahim Ali Alasiri, Mohammed Farhan Alfaifi, Nawaf Saud M. Alhali, Sari Aljohani, Jaber Salman Alraythi and Grami Mohammad A Grami *

King Abdulaziz University, Saudi Arabia.

International Journal of Science and Research Archive, 2026, 18(02), 271-276

Publication history: Received on 27 December 2025; revised on 02 February 2026; accepted on 04 February 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.18.2.0216

Abstract

This paper argues that integrating corpus-based approach into language education and classrooms through direct and indirect applications can enhance the learners' linguistic competence. By using data-driven learning tools and other methods and apply them effectively. Additionally, it also highlights the various benefits in incorporating corpus-based into the classrooms and bridging the gap between theoretical and practical applications.

Keywords: Corpora; Saudi Classoom; EFL; Education; Technology

1. Introduction

1.1. The Advantages of Corpora

Language corpora can give us a unique view of how people use language. Instead of relying on theocraticals, corpora draw from real life conversations, articles, books, etc. this makes corpora useful in analyzing the structure and function of language. Whether it is spoken or written, the data in a corpus provides solid evidence of how language works in different situations and settings.

One of the main advantages of using corpora is the ability to study both spoken and written text at the same time. Each of these forms has their own functions. Spoken language includes tone, rhythm, and stops. Written language on the other hand includes things like punctuation and organization. Nevertheless, they often compliment each other. By comparing them, we can get a better understanding of how people communicate. As Dash and Arulmozi (2018) note, these two forms should be treated as equally important in any serious language study.

Another major advantage of corpora is that it helps us see the deeper layers of language. Things like implied meaning, context determining a word's usage, and how tone can reflect someone's emotions or intentions. These features are usually hard to notice on their own just by reading a sentence once. But when we have thousands of examples through corpus, patterns begin to emerge, and it becomes easier for us to notice them. As mentioned in the course material (History, Features, and Typology of Language Corpora, *Advantages of a Corpus*.), understanding these hidden aspects requires close and careful analysis, and corpora make that possible.

Susan Hunston (2012) also points out that corpora have had a big impact on how we build grammar books and dictionaries. These days, many learners use dictionaries that are based on corpus data, meaning that they reflect how people actually speak and write, not just how they're supposed to. This is an amazing feature, especially for those who want practical, real-life examples, like language learners and teachers.

* Corresponding author: Grami Mohammad A Grami

Corpora also make it easier for us to study writing style. If you look at the works of a single author across multiple texts, you can see certain habits, like sentence structure or favorite phrases. This can be useful in some fields like forensic linguistics or literary analysis, where identifying someone's writing fingerprint (writing style) can be important.

Corpora are a big asset in education. For teachers, they can be used to find reliable examples of grammar, vocabulary, and even different cultural expressions. Meanwhile, students can use them to see how language is used in real situations, resulting in a learning method more efficient than memorizing rules. It makes language far more engaging in a way textbooks often couldn't achieve.

Back in the day, building corpora was expensive and time consuming, but not anymore. With internet access and basic tools, it is now easier to collect a large amount of language data from websites and online sources. Dash and Arulmozi (2018) note that while well-established languages like English already have plenty of corpora, more recent efforts are starting to build similar resources for less-represented languages, including many spoken in India. This will help to ensure promising development for more inclusive linguistic research.

By making language patterns visible, corpora allow us to question assumptions, spot changes, and appreciate how dynamic and adaptive human communication really is. Furthermore, as more languages gain digital representation through corpora, the potential to teach, understand, and preserve them, grows even stronger.

In conclusion, corpora offer a wide range of advantages. They provide authentic data, deep analysis, and help connect theory and real usage. Whether you are a linguist, teacher, or just someone who is curious about languages, corpora offer a better insight that traditional methods cannot bring to the table.

2. Phrases and Collocations in Corpora

Grasping how words naturally team up collocations and other established set phrases is absolutely key to truly mastering a language (Sinclair, 1991; McEnery & Hardie, 2012; O'Keeffe & McCarthy, 2022). This crucial skill underpins fluent, accurate, and genuinely natural communication, which is the very hallmark of authentic discourse. However, many second language learners consistently struggle with collocations, primarily due to their vast number, their often nuanced contextual usage, and the persistent interference from their native tongue (Nesselhauf, 2005; Du et al., 2022; Li, 2017).

Large digital text collections, widely known as corpora, offer demonstrably invaluable resources for thoroughly exploring and effectively teaching these crucial word partnerships (Li, 2017; McEnery & Hardie, 2012; Sinclair, 2004). Pioneering studies consistently highlight how corpora provide direct access to authentic language, offering rich empirical insights into real world usage by proficient speakers (Sinclair, 1991, 2004; O'Keeffe & McCarthy, 2022). This evidence based analytical approach is vital for uncovering complex collocational patterns that traditional teaching methods or simple intuition might otherwise overlook.

Direct engagement with corpora enables detailed, systematic observation of word co-occurrence frequency, specific context, and subtle semantic nuances (Li, 2017; Sinclair, 1991; Nesselhauf, 2005). Analyzing concordance lines, for instance, clearly reveals common collocates and the typical grammatical structures of phrases (O'Keeffe & McCarthy, 2022; McEnery & Hardie, 2012; Sinclair, 2004).

Investigating particularly challenging combinations, such as specific verb-noun or verb-preposition pairs that often perplex learners (Du et al., 2022; Nesselhauf, 2005; Li, 2017), becomes more systematic with corpus data. Extensive research confirms that hands-on corpus work significantly boosts learners' collocational accuracy and their overall phraseological awareness (Li, 2017; Nesselhauf, 2005; Du et al., 2022), fostering a deeper, more robust understanding of natural wordings. Corpora are thus powerful tools for linguistic inquiry and effective language pedagogy.

3. Applications and Implications of Corpora

The application of corpus linguistics extends across a broad range of research disciplines that are looking for ways to study language, by providing a large collection of naturally occurring language data. Dash and Ramamoorthy (2019) "language corpora have become primary resources for numerous branches of application-oriented and description-based linguistics" (p 209). The application of corpora varies depending on the field they are used in and how they are analyzed, by providing means to analyze the linguistic characteristics of a language such as phonetics, morphology and semantics of a language. Also, the data it provides can be researched in relation to lexicography and translation studies

as well as machine learning. Moreover, depending on the data it can give both how a language is produced in different social groups and how language comprehension varies between dialects or regions. These features are why corpora is considered a multidimensional, multifunctional, and multidirectional tool to study the language (Dash & Arulmozi 2018). This section discusses some of corpora applications in various fields and their implications for research and analysis.

The growth of corpora reflects from its applications, which are used in mainstream linguistics, applied linguistics and language technology, serving as an empirical evidence-based data foundation to those fields research and studies (Dash & Arulmozi 2018). In mainstream linguistics, corpora provide authentic data that reflects real-world usage of a language, these data are used in lexical studies, grammar writing and speech analysis. In these fields, linguists can look beyond constructed examples and observe actual language patterns using corpora which support the construction of descriptive models and the formulation of linguistic theories (Dash & Ramamoorthy 2019). In lexicology and lexical semantics, application of corpora helps uncover word meaning, collocations, and semantic shifts. While in syntax and pragmatics they reveal contextual and structural patterns that deepen the understanding of language function and use. Corpora allow researchers to verify their observations, pinpoint new linguistic phenomena and outline emerging rules that empowers their studies claims to which are based on a naturally occurring language data, thus gives accuracy and relevance of their analysis across both theoretical and practical domines.

Language corpora applications are seeing a surge of usage in applied linguistics in the 21st century, including sociolinguistics, psycholinguistics, stylistics and translation studies. Corpora often provide metadata such as age, gender and social status to support sociolinguistic and stylistic analysis (Dash & Arulmozi 2018). In critical discourse analysis, small custom-built or large datasets help providing linguistic evidence to uncover ideologies and power relations in texts, while corpora tools such as concordancers assist in identifying collocational patterns for deeper interpretations (Baker, 2012). Corpus stylistic applies quantitative methods to the analysis of literary texts to identify creative language use and stylistic trends, while varieties of regional language dialects can be studied through corpora. Corpus data contributes to word-sense disambiguation in lexicography they are essential because of the real examples that are extracted from naturally occurring language to create accurate and up-to-date dictionary entries (Dash & Ramamoorthy 2019). Moreover, parallel and aligned corpora are use in translation studies for identifying translation equivalents and building terminological resources. In language acquisition research and English language teaching, corpora can track the pattern in which one can acquire the language, and it is used to create authentic materials that suit the learner from the language real usage in every day.

Language corpus applications are used and became essential in language technology and computational linguistics, applications such as machine translation, text annotation, speech processing and parsing. These applications are aiding to complete tasks such as word sense disambiguation, spelling correction and text-to-speech conversion systems (Dash & Arulmozi 2018). Parallel and translation corpora play a big role in multilingual processing and machine learning. Machine learning and cross-lingual studies utilize parallel, comparable and translation corpora, which became essential in some studies (Dash & Arulmozi 2018). Web text corpora (WTC) and computer-mediated communication (CMC) corpora face challenges such as normalization, annotation and copyright infringement, to ensure data quality it needs careful assembly of the data to avoid any complications and to ensure its validity by providing and accurate data to the study (Dash & Arulmozi 2018, Baker, 2012).

Beyond the scoop of their various applications, corpora extend to have an important theoretical and methodological implications for the study of language. According to Dash and Arulmozi the use of corpora has significantly impacted the way linguistic research is conducted, adapting an empirical, data-driven approach, unlike intuition-based methods, which allows researchers to confirm or challenge earlier theories with observable evidence through the analysis of naturally occurring language data (Dash & Arulmozi 2018). More than a text collection, a corpus uses tools such as concordancers and collocators to analyse a structured dataset, which help reveal patterns that are too subtle or widespread for manual analysis. This has given deeper insight into studies of lexical semantics, sociolinguistics and stylistics. Dash and Arulmozi added that misunderstanding the nature of a corpus can lead to flawed studies, highlighting the need for a researcher to be able to identify a corpus database from other language datasets (Dash & Arulmozi 2018).

Language corpora growth and application significantly reinforced fields of research such as lexicology, semantics, sociolinguistics, psycholinguistics and stylistics trends, making them more data-driven and insightful. It allowed researchers to rely on corpus data in their research to uncover the answers to linguistic questions that traditional methods can not address. However, there are limitations. Large corpora often lack context, omitting visuals, intonation and layout features crucial to meaning (Baker, 2012). Also, an overemphasis on frequency can also obscure the significance of rare but meaningful occurrences (Baker, 2012). Moreover, lexicographers may struggle with data

overload, while some corpora underrepresent genres like poetry and dialogue or lack balance (Dash & Ramamoorthy 2019, Baker, 2012). Despite these challenges, ongoing studies in the refinement of corpus design aim to fix such issues and provide more stable tools for its applications in language studies.

4. Using Corpora in the Classroom

Corpora—collections of real-life language data—can be powerful tools in language teaching when integrated into classroom activities. Their use enhances students' exposure to authentic language and encourages critical, data-driven learning.

4.1. Types of Use

Leech (1997) identifies three modes of direct corpus use in education:
- Teaching About Corpora: Instructing students in corpus linguistics as an academic subject.
- Teaching to Exploit Corpora: Training students to independently use corpus tools.
- Exploiting Corpora to Teach: Using corpora to teach language concepts such as grammar and vocabulary.
These methods are especially suited to advanced learners in tertiary education due to their complexity and need for guided analysis.

4.2. Data-Driven Learning (DDL)

Johns (1991) introduced data-driven learning, which encourages learners to explore authentic language data and derive rules independently. The process emphasizes three core stages:

- Observation – Examining concordance lines.
- Classification – Identifying patterns and structures.
- Generalization – Forming linguistic rules based on data.

DDL shifts the learning paradigm from teacher-centered explanation to learner-centered discovery, thus promoting autonomy and critical thinking.

4.3. The “Three I’s” Approach

Carter and McCarthy (1995) proposed a complementary method to DDL:
- Illustration – Presenting real corpus data.
- Interaction – Engaging learners in collaborative analysis.
- Induction – Encouraging students to infer grammatical or lexical rules.
This method suits exploratory teaching and aligns with the inductive reasoning promoted in DDL.

4.4. Classroom Applications

Teachers can apply corpora in multiple ways:
- Analyzing collocations and lexical bundles.
- Comparing learner data to native-speaker corpora.
- Exploring grammatical patterns (e.g., tense, aspect, modals).
- Using local learner corpora for error analysis and peer feedback.
Tools like AntConc or web-based corpus platforms support these activities.

4.5. Practical Classroom Applications

Here are several ways to integrate corpora into language classes:

4.5.1. Teaching Vocabulary and Collocations

Use concordance lines to show how words like “make” or “take” combine with other words (e.g., “make a decision”).

Teach common **collocations** and **phraseology** for fluency.

4.5.2. Grammar Discovery

Analyze how grammar structures (e.g., passive voice, tense, conditionals) are used in real speech and writing.

Compare usage across different registers (spoken vs. written).

4.5.3. Error Analysis

Use **learner corpora** to identify common learner errors.

Have students correct authentic examples from their peers or learner corpora.

4.5.4. Genre Awareness

Use corpora to explore how different genres (e.g., emails, reports, essays) vary in language use.

Raise awareness of **register**, **formality**, and **discourse structures**.

4.6. Role of the Teacher

Even in student-centered DDL, **teachers play a vital role**:

Guide the learning process.

Scaffold activities for learners of varying levels.

Prevent misinterpretation of corpus data (Sinclair, 2004).

Using corpora in the classroom shifts language learning from passive absorption to active discovery. When properly supported, it fosters autonomy, deeper linguistic insight, and real-world language awareness. Though challenges exist, especially around training and resource access, corpora offer a powerful, evidence-based approach to modern language education.

5. Conclusion

Incorporating corpora in language studies offers many benefits as it provides learners with accurate data and examples that can help enhance their linguistic competence and awareness. It also helps researchers to identify the common patterns, and word frequency and to develop a more accurate analysis on how words are used in various linguistic contexts can be an effective way to teach language learners. While a powerful tool, corpora have limitations. These include the lack of contextual aspect of language use that can influence meaning, they cannot fully capture the dynamic and ever-changing nature of the language due to other social, historical factors as language changes over time which corpora may not accurately describe. In spite of, with further studies and the ongoing development and as corpus resources expand educators and language learners will be able to use it more effectively and foster their linguistic competence.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Dash, N. S., & Arulmozi, S. (2018). *History, features, and typology of language corpora*. Springer Singapore.

[2] Hunston, S. (2012). Applications of corpora in applied linguistics. In *Corpora in Applied Linguistics* (pp. 96–136). Cambridge University Press.

[3] Du, X., Afzaal, M., & Fadda, H. A. (2022). Collocation use in EFL learners' writing across multiple language proficiencies: A corpus-driven study. *Frontiers in Psychology, 13*, Article 752134. https://doi.org/10.3389/fpsyg.2022.752134

[4] Li, S. (2017). Using corpora to develop learners' collocational competence. *Language Learning & Technology, 21*(3), 153–171. https://scholarspace.manoa.hawaii.edu/bitstream/10125/44625/1/21_03_li.pdf

[5] McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

[6] Nesselhauf, N. (2005). *Collocations in a learner corpus* (Vol. 14). John Benjamins Publishing. https://doi.org/10.1075/scl.14

[7] O'Keeffe, A., & McCarthy, M. J. (2022). *The Routledge handbook of corpus linguistics* (2nd ed.). Routledge. https://doi.org/10.4324/9780367076399

[8] Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

[9] Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. https://doi.org/10.4324/9780203594070

[10] Dash, N. S., & Arulmozi, S. (2018). *History, features, and typology of language corpora*. Springer Singapore.

[11] Dash, N. S., & Ramamoorthy, L. (2019). *Utility and application of language Corpora*. Springer Singapore.

[12] Baker, P. (2012). Contemporary Corpus Linguistics. Continuum.

[13] Carter, R., & McCarthy, M. (1995). Grammar and the spoken language. Applied Linguistics, 16(2), 141–158.

[14] Johns, T. (1991). "Should you be persuaded": Two samples of data-driven learning materials. In Classroom Concordancing, ELR Journal 4, University of Birmingham.

[15] Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), Teaching and Language Corpora (pp. 1–23). London: Longman.

[16] McEnery, T., & Xiao, R. (2010). What corpora can offer in language teaching and learning. In E. Hinkel (Ed.), Handbook of Research in Second Language Teaching and Learning: Volume II (pp. 364–375). New York: Routledge.

[17] Sinclair, J. (2004). How to Use Corpora in Language Teaching. Amsterdam: John Benjamins.