Using a Pedagogic Corpus and Sketch Engine Tools to Analyze a Literary Text in an Arabic Language Classroom

DOI: https://doi.org/10.33806/ijaes.v25i2.941

Hafid Maachi
Mohammed V University in Rabat, Morocco

Hakima Khamar
Mohammed V University in Rabat, Morocco

Lutfi Omar Abubkr
Mohamed bin Zayed University for Humanities, UAE

Received: 15.11.2024 Accepted: 9.3.2025 Published: 2.6.2025

Abstract: Corpora represent an essential pedagogical tool in the field of language education, contributing effectively to the enhancement of language teaching methodologies and techniques, the improvement of competencies for teachers, and the development of linguistic and cognitive skills for students. However, Arabic language teaching methods completely lack the use of this tool in classrooms, which falls within the broader approaches of using technology in language pedagogy. The study highlights the significance of corpora in language teaching and learning. Moreover, it addresses the challenges associated with using corpora for teaching Arabic in classrooms. The research demonstrates how to construct a pedagogic corpus based on the literary novel "The Thief and the Dogs," which is included in the second-year baccalaureate curriculum in Morocco. In this study, we will annotate the corpus with parts of speech and implement a series of activities and exercises designed within the framework of Arabic language instruction, utilizing the Sketch Engine tools. The study highlights the central role of corpora in achieving excellent educational outcomes and provides recommendations for effectively integrating it into Arabic language teaching.

Keywords: Arabic language teaching, corpora, language education, pedagogic corpus, pedagogical tool, teaching methodologies and techniques

1. Introduction

The world has witnessed significant changes in the last decade driven by the rapid progression of science, technology, and digitalization. Numerous technological tools of recent times have proven themselves indispensable pedagogical resources in the field of education. One of the most important tools in this category is corpora, which have gained importance in contributing much to applied linguistics and brought great opportunities, especially for several domains such as language teaching and learning. Many researchers concur that corpora have transformed language education (Leech 1997; Braun 2005; Römer 2006, 2008; Conrad and Levelle 2008; Boulton 2010, 2011). According to Römer (2011: 205), the effect of corpora on linguistic research as a whole, in particular second language teaching

and learning has been revolutionary. In the English Language Teaching (ELT) domain, Boulton (2017), for instance, conducted a meta-analysis of 88 empirical studies that have quantified the impacts of learning via corpora. The overall synthesis of these studies provides strong evidence for this instructional approach, its effectiveness, and its success in the area of language acquisition and pedagogy.

In the book's introduction "How to Use Corpora in Language Teaching", Sinclair (2004: 2) highlighted the essential role of corpora in education, highlighting their significance as both a valuable and indispensable tool within the educational landscape. Despite the significant potential that corpora offer for education, particularly in language teaching, the Arabic language still faces a considerable lack of research on utilizing these corpora in classroom teaching compared to English. Al-Sulaiti and Atwell (2006: 136) emphasized this gap, noting that experimental linguistic studies in English-speaking countries have achieved substantial progress due to the availability of extensive corpora. In contrast, linguistic research in Arabic relies heavily on limited linguistic data. Although Arabic corpora exist, they remain insufficient to support large-scale experimental linguistic research.

Given the complex nature of the Arabic language with its complex structure and precise grammatical rules, there was a need to improve innovative teaching methods and techniques, which are better than those classroom methods. Corpora represent a useful tool in the teacher's hands that could supply realistic models of language uses aligned with educational contexts. Through these corpora, learners can understand the linguistic rules and structures commonly used in educational settings and develop the ability to better understand the content and apply the grammatical rules in real-life classroom situations. This approach will not only improve educational outcomes but also encourage students to actively participate and promote interactive and independent learning in Arabic.

The first section of this research investigates and highlights the significance of corpus linguistics in language teaching and learning. Moreover, it addresses some challenges associated with using corpora for teaching Arabic in classrooms. The second section delineates the methodology for constructing a pedagogic corpus derived from a literary novel integrated into the curriculum for second-year baccalaureate students in Morocco. It encompasses a comprehensive overview of the processes involved, ranging from data collection to part-of-speech tagging. The third section describes the implementation of specific educational activities and exercises for teaching Arabic, using Sketch Engine tools to analyze the educational corpus. The fourth section presents the findings, which are then discussed in depth in the fifth and final section.

2. Literature review

2.1 Corpus linguistics and language teaching and learning

The application of corpus linguistics to the field of language teaching and learning began in the late 1980s and early 1990s. From that time, academics and researchers recognized the great potential that was inherent in this methodological approach for enhancing and benefiting language pedagogies in general (Flowerdew 2009: 327).

Leech (1997: 9) pointed out that corpus-based approaches began to become prominent from the mid-1980s onwards, as the study of corpora became an integral part of the wider field of linguistics and more so in the areas of language teaching and learning. The application of corpus linguistics in language teaching has generated much interest in Western tertiary institutions, with numerous research projects benefiting from this method (McEnery 2006: 97). This phenomenon is not limited to the scope of the English language; it is also applicable to the pedagogy of other languages, such as Arabic. According to Conrad and Levelle (2008: 539), most studies related to the use of corpora in language teaching focus primarily on the English language. Nonetheless, the methods and tools used in these studies can be adapted for use with other languages.

Corpora can substantially influence language teaching across a range of areas, as outlined in Figure 1 below. According to Johansson (2009), corpora offer significant and varied contributions to the field of language teaching. Corpora serve as a valuable data resource for teachers, students, and researchers in second language acquisition, language lexicographers, educational material designers, and others. For example, corpus-based analysis can reveal critical issues across various aspects of language use by learners, and these findings can enable material designers to develop improved teaching resources that consider students' strengths and weaknesses, potentially leading to more effective language acquisition (Nesselhauf 2004). Hunston (2002:189) indicates that revising curriculum design based on corpora contributes to enhancing the quality of teaching and adapting it to the needs of contemporary students. Consequently, using corpora allows teachers to develop interactive and innovative teaching tools. It also promotes students' autonomy in exploring and discovering linguistic knowledge, enabling them to apply linguistic concepts independently and expand their research skills (Römer, 2006: 124).

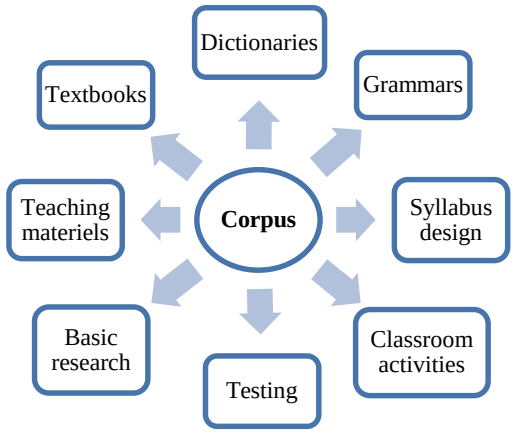

Figure 1. Uses of corpora of relevance for language teaching (Johansson 2009).

Corpora contribute significantly to language teaching, influencing two main areas. The first involves using corpus data to select appropriate vocabulary and develop curricula and teaching materials. The second area is the use of corpus data as a direct instructional tool. These two areas have been referred to as "indirect applications" and "direct applications" of corpora (McEnery and Xiao 2011; Römer 2011). Figure 2 illustrates that direct applications of corpora in language teaching and learning involve both students and teachers who engage with corpora and concordance tools through hands-on activities. In contrast, indirect applications pertain to researchers, material designers, and curriculum developers. According to Römer (2010: 19), direct applications of corpora focus on addressing "how" to teach, while indirect applications aim to answer questions regarding "what" and "when" to teach. Applications of pedagogical corpora can be classified based on their impact, including "direct" applications that have a clear effect on students and teachers in language teaching and learning, and "indirect" applications that assist researchers and educational material designers in developing teaching materials and curricula, as shown in Figure 2 (Römer 2008: 113).

Corpora have had a strong influence on the pedagogical approach in language teaching, both in the curricular content and the methodologies of instructions used (McEnery 2006: 10). They constitute rich data of authentic linguistic resources that permit learners to work with real-life linguistic examples that they are likely to encounter outside the classroom. Moreover, corpora offer learners a vast amount of examples concentrating on specific linguistic aspects (for example, rare vocabulary or grammatical structures), which allow them to grasp these aspects, as well as their contextual applications and intertextual connections (Reppen 2010: 35). Sinclair (2004: 297) highlighted the many advantages of corpus use for both teachers and learners, adding that activity in this area not only increases knowledge about linguistics but also makes the learning process more enjoyable and entertaining.

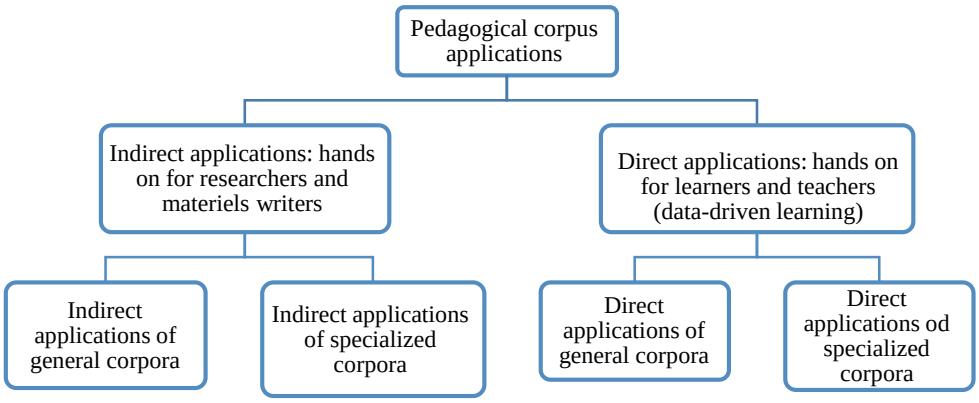

Figure 2. The use of corpora in language learning and teaching (Römer 2010: 19).

Direct applications of corpora and corpus analysis tools in the classroom support a range of theories and concepts related to language teaching and

acquisition, particularly concerning learner autonomy, the use of authentic and real-life texts, learner-computer and peer interactions, as well as the direct teaching of linguistic features and patterns (Friginal 2018:12). Fuster-Márquez and Gregory-Signes (2018: 166) argue that in direct approaches, students explore corpus data either through activities designed by teachers or independently by directly accessing the available data.

2.2 Corpora-based and technologically enhanced approaches to language education

In recent decades, various educational approaches have emerged, utilizing a range of technological tools within the framework of "Computer-Assisted Language Learning" (CALL). Among these technologies, corpora are widely used by instructors to teach languages in classroom settings (Ma et al. 2022: 461). Conrad and Levelle (2008: 542) noted that in the realm of CALL, and particularly through "Data-Driven Learning" (DDL), students especially advanced learners can benefit from corpora. This approach allows them to analyze and independently simulate language, enhancing their linguistic abilities and encouraging exploratory learning. Additionally, corpus-based activities practiced by students fall within the scope of DDL (Götz and Granger 2024:15).

The CALL field, which encompasses the use of technology in language instruction, has witnessed significant developments. Its tools offer numerous advantages to students aiming to acquire language skills independently, outside the confines of traditional classrooms (Götz and Granger 2024:7-8). Educational applications within corpus linguistics have gained increasing attention among language instructors, as they form part of CALL (Abdel Latif 2021: 1). Tim Johns is credited with coining the term "Data-Driven Learning" (DDL) in 1990, marking the beginning of a new approach to language teaching and learning. He introduced an innovative perspective, making students active participants in discovering language through their exploration of linguistic data (Johns 1991:2). DDL is used in classrooms employing CALL to help students explore targeted linguistic patterns, with activities and exercises based on corpora (Levchenko 2017: 31).

2.3 Challenges of using corpora in teaching Arabic in the classroom

Despite the growing interest in corpora in Arab countries, there is a significant lack of Arabic research addressing their application in teaching and learning Arabic within classroom settings. The few available studies focus mainly on the role of corpora in teaching Arabic as a foreign or second language. For instance, Zaki (2017) proposed a practical approach for using corpus software in teaching Arabic as a foreign language, aligned with advancements in corpus-based pedagogy. Whitcomb and Alansary (2017) discussed various methods for incorporating corpora in teaching and learning Arabic as a foreign language. Abdel Latif (2021) investigated training Gulf-region teachers of English as a foreign language in corpus skills, collecting data to gauge their views and explore their intentions to use corpora in future teaching. Al-Sulaiti and Atwell (2006) covered the design of a contemporary Arabic corpus, outlining how to create and use corpora in linguistic

research and teaching Arabic to non-native speakers. Alfaifi and Atwell (2012) examined learner corpora, reviewing two Arabic learner corpora and highlighting their potential for language learning and teaching.

Regarding corpus creation and tools, Alotaibi (2017) introduced an ambitious project to develop a large linguistic corpus aimed at providing valuable resources for training translators and teaching Arabic. Almujaiwel and Al-Thubaity (2016) reviewed a new language processing tool (Arabic Corpus Processing Tools ACPTs Version 4.6) used for corpus analysis, outlining key functions of the tool that could be leveraged to support language teaching and learning.

Arab countries are currently experiencing a notable lack of investment in the application of corpora for the teaching of Arabic. Al-Sulaiti and Atwell (2006: 147) observed that corpora usage is predominantly restricted to the English Department at Kuwait University, where it is mainly utilized for translation and lexicography instruction. Nevertheless, recent studies have started to underscore the importance of incorporating corpora into Arabic language education.

For instance, Bacha and Khachan (2023) analyzed essays written by Arabic-speaking students, employing tools like Lextutor and VoyantTools to evaluate word frequency and lexical density. Their findings illustrated the potential of corpora to enhance writing assessment. In a similar vein, Alfuraih and El-Jasser (2024) developed the University Learner Translator Corpus (ULTC), a parallel corpus of 75 million words, aimed at overcoming data collection challenges and advancing translation education.

Furthermore, Al-Sabbagh (2023) highlighted the urgent need for the development of parallel corpora for Arabic dialects to improve neural machine translation. Relying exclusively on Standard Arabic corpora can lead to negative transfer, compromising translation quality and limiting the understanding of dialect-specific characteristics.

A major challenge in teaching Arabic using corpus-based methods is the lack of corpora specifically designed for educational purposes. Most existing Arabic corpora were created for research and are unsuitable for classroom use. Additionally, the costs of licenses and analysis tools can be prohibitive. While teachers can access a few free corpora, these options are limited (Zaki 2017:517). Mansour (2013:81) noted that linguistic research in Arab countries rarely uses corpora due to their scarcity compared to English language resources.

3. Creation of a pedagogic corpus

3.1 Criteria for designing pedagogic corpus

Pedagogic Corpus includes language used in the educational environment, such as textbooks or verbal interactions among students in the classroom. This language is crucial for the learner, as it is encountered frequently and contributes to the learning process. It is also used to develop teaching skills for teachers (Bennett 2010: 13-14). The criteria for designing corpora vary depending on their objectives and the linguistic needs the corpus aims to address (Williams 2002: 46). The steps involved in creating a linguistic corpus include the design of the corpus, text collection, corpus annotation, and the compilation and storage of metadata related to the texts,

in addition to including linguistic tagging (McEnery and Hardie 2012: 241). Aston (2001: 37) pointed out that using a small, specialized corpus has advantages over a larger corpus in language pedagogy due to its ease of organization, analysis, and retrieval of information, as well as the clarity of the educational objective behind it.

It can be said that corpora designed for educational purposes should be more cohesive than traditional corpora. They should, as much as possible, complement the curriculum to facilitate their incorporation into the educational context (Braun 2007: 308). Lee (2011: 167) noted that using textbook corpora in the classroom aligns well with the context of language education. According to Reppen (2010: 36), there are three different ways in which teachers can present practical activities to students using corpora. First, teachers can bring materials extracted from corpus searches and have students work with these materials prepared by the teacher. Second, teachers can use some available online linguistic corpora. Third, teachers can bring ready-made corpora or create specialized corpora for the classroom (e.g., a corpus derived from readings or students' papers) and have students interact with these corpora. In this study, we adopted the third method, which involves creating an educational corpus derived from a literary novel used in an educational syllabus.

3.2 Creating the pedagogic corpus

We selected the novel "The Thief and the Dogs" (اللص والكلاب) written by Naguib Mahfouz to use as a pedagogic corpus. This novel is included in the curriculum for second-year baccalaureate students in the fields of Literature and Humanities in Morocco. The novel contains a variety of texts, blending narration and description, reflecting the diversity of its linguistic structures.

There are several methods to convert the novel, whether in paper or digital format, into a machine-readable format. These methods include:

1. Typing the book into the computer using specialized software;
2. Using an Optical Character Recognition (OCR) scanner, which scans the printed text and converts it into a digital plain text format;
3. Using a Speech-to-Text system;
4. Converting PDF files into editable text through PDF editing software.

After converting the novel into an electronic format, the text is then revised and cleaned of any excess content or imperfections that might affect the analysis and statistics. In the next phase, the electronic format is converted into a plain text file (.txt) and encoded in Unicode UTF8 or Unicode UTF16.

3.3 Including metadata and linguistic annotation in the pedagogic corpus

After preparing the pedagogic corpus in a machine-readable format, it is enriched with metadata for classification and organization. This includes specifying the temporal and spatial context of the text, identifying the author, the title of the novel or book, the publication date, and other relevant information. The next stage involves adding linguistic annotation, which refers to the process of adding linguistic tags or labels to the components of the text to identify its linguistic structure and various meanings. Part-of-speech (POS) tagging is one of the most common and widely used types of annotation, as it is closely related to many corpus

linguistics studies and supports other tagging processes, such as lemmatization, syntactic parsing, and semantic annotation. This type of annotation involves assigning a tag or label to each word token, which helps to determine the minimal part of speech of the word (Gries and Berez 2017: 383).

We annotated the pedagogic corpus with parts of speech using the Stanford Arabic Part-of-Speech Tagger, which was designed and developed by the Natural Language Processing Group at Stanford University (https://nlp.stanford.edu/software/arabic-tagger.shtml). This tool employs a set of tags derived from the Penn Arabic Treebank (ATB) to describe and identify the different parts of speech in Arabic texts, providing a detailed and accurate analysis of the Arabic language (The Stanford Natural Language Processing Group). Figure 3 shows a screenshot from our pedagogic corpus, tagged with parts of speech using an Extensible Markup Language (XML) format.

Figure 3. An annotated text from the pedagogic corpus tagged with parts of speech in XML format

4. Applying a set of educational activities based on the pedagogic corpus

4.1 Overview of Sketch Engine and its functions

There are a number of programs and tools that support the Arabic language, including Sketch Engine, aConCorde, AntConc, WordSmith, Ghawwas, and IntelliText Corpus Queries (Alfaifi and Atwell 2016: 348). These tools are designed for text corpus searches, except for Sketch Engine and IntelliText, which rely on web-based queries. The evaluation criteria for these tools include the ability to read Arabic text files in Unicode formats (UTF8 or UTF16), display diacritical marks, adjust the accuracy of Arabic text display from right to left, unify diacritics or the hamza, provide a user interface in Arabic, and enable users to open or upload personal collections in Arabic (Almujaiwel and Al-Thubaity 2016: 1-2).

In this study, we chose Sketch Engine, an internet-based platform (Kilgarriff et al. 2004), this program allows users to work with 600 sub-corpora in over 90 languages, including Arabic. Users can take advantage of this program by working with available corpora or uploading their own linguistic corpora created by

researchers themselves. Among the users of Sketch Engine are linguists, lexicographers, translators, terminologists, teachers, students, historians, and other researchers (https://www.sketchengine.eu.). Sketch Engine is widely used in English language teaching and has also been used intermittently in teaching other languages such as Chinese, Japanese, and Arabic (Kilgarriff et al 2014: 16).

4.2 Applying educational activities and exercises using Sketch Engine tools

We have designed five types of educational activities that are carefully aligned with the content of the pedagogic corpus and tailored to the student's proficiency levels. The use of corpora contributes to providing an accurate representation of the language that students will encounter, making it a valuable tool that enables teachers to plan educational content, develop teaching materials, and create effective instructional strategies, including enhancing the design of educational activities (O'Keeffe et al 2007; Reppen 2010).

For this reason, we designed the activities to leverage corpora as a central resource, ensuring that students engage with authentic language use while enhancing their linguistic and analytical skills through carefully planned educational tasks. These activities were successfully implemented using the advanced functionalities of Sketch Engine tools. Then, we presented the results of the analysis for each activity. The activities are outlined as follows:

Activity 1: Word frequency analysis

1. **Objective**: analyzing the frequency of occurrence of primary and secondary characters
2. **Tool used**: Wordlist.
3. **Instructions**:

a. Students are instructed to search in Figure 4 for the names of characters and count the frequency of each character's occurrence.

b. Students compare the most frequently occurring characters in the corpus and infer their primary and secondary roles.

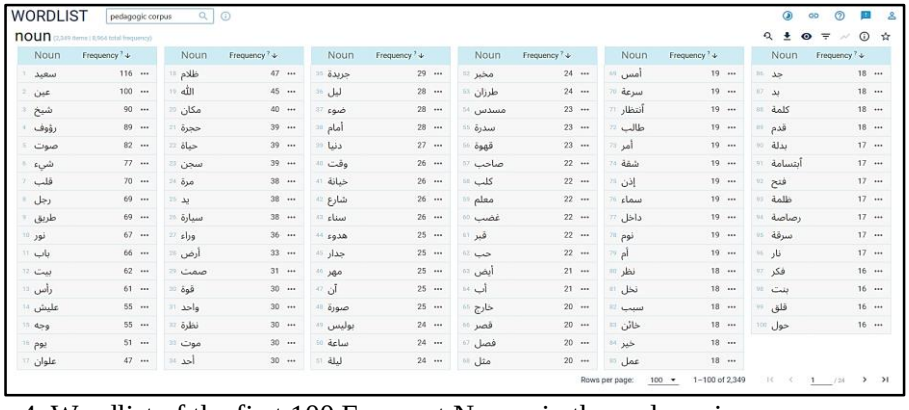

Figure 4. Wordlist of the first 100 Frequent Names in the pedagogic corpus.

Activity 2: Verb tense analysis

1. **Objective**: analyzing the use of verb tenses in various contexts and their impact on the structure of the novel.
2. **Tool used**: Wordlist and Concordancer.
3. **Instructions**:

a. Students extract the first 10 verbs along with their frequency rates in the corpus from Figure 5.

b. Using the Concordancer tool, students identify the tenses of the verb "*قتل*" (to kill) and classify them according to tense (past, present, future) based on Figure 6.

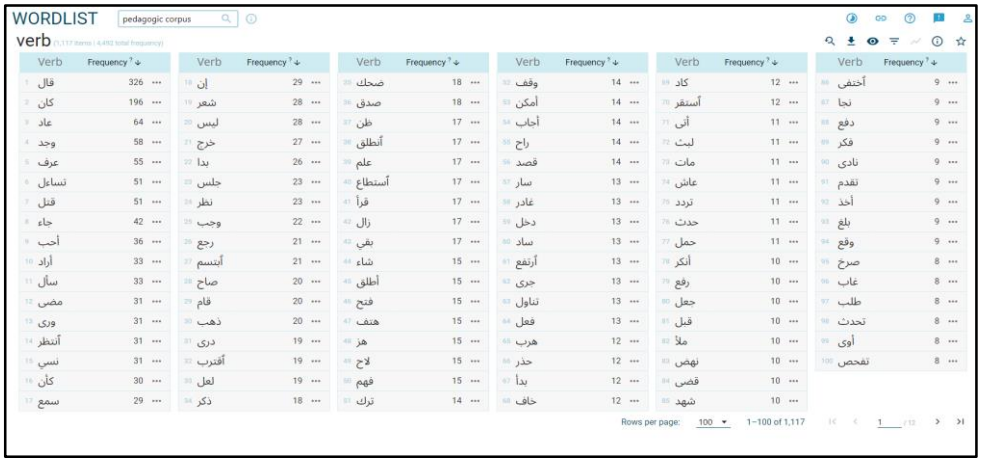

Figure 5. Wordlist of the first 100 frequent verbs in the pedagogic corpus.

Figure 6. Concordances of verb "*قتل*" (to kill) in the pedagogic corpus.

Activity 3: Collocation analysis

1. **Objective**: studying the collocations of the word "صوت" (sound).
2. **Tool used:** Word Sketch.
3. **Instruction**: Students extract the collocates of the word "silence" using Figures 7 and 8.

Figure 7. Word Sketch of the word "صوت" (sound).

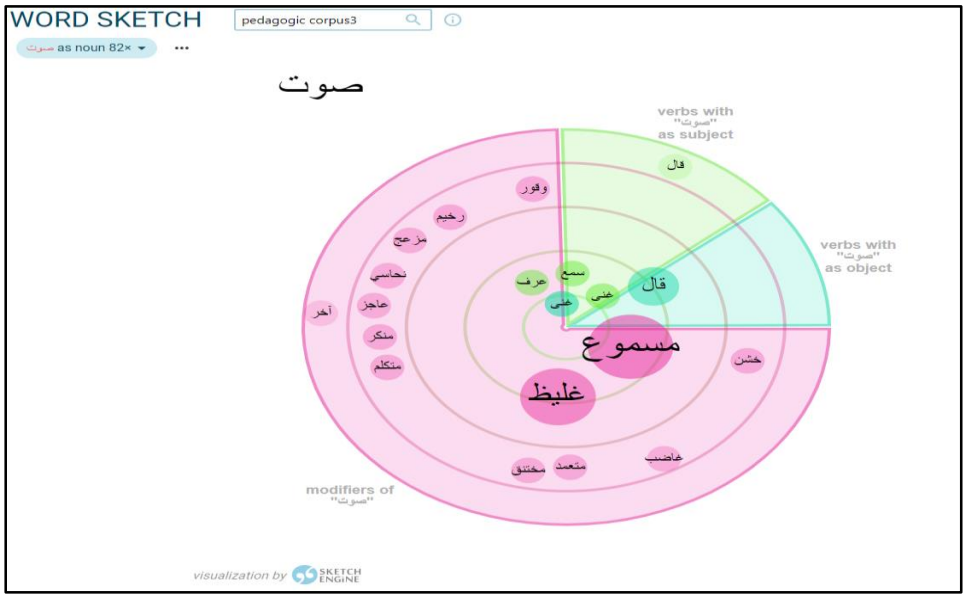

Figure 8. Word Sketch of the word "صوت" (sound).

Activity 4: Analysis of figurative language in the novel

1. **Objective**: Identify figurative language such as similes and highlight their role in enhancing the literary meanings within the novel, as well as to understand their impact on the text.
2. **Tool Used**: Wordlist.
3. **Instruction**: Students use Figure 9 to explore an example of a simile with the comparative particle "like" (كـ), then determine whether the dominant emotion conveyed is positive or negative.

Figure 9. Wordlist of words Beginning with the Simile Particle "كـ" (Like).

Activity 5: Deducing grammar

1. **Objective**: analyzing contexts using concordance to deduce grammatical rules.
2. **Tool Used**: Concordance, with words tagged for parts of speech.
3. **Instruction**: Students are asked to identify the "parts of speech" of the words as shown in Figure 10, and deduce them in Arabic.

Figure 10. Concordances of the word "*السجن*" (prison).

5. Results

The implementation of the designed educational activities, through analyzing the pedagogic corpus using Sketch Engine tools, demonstrated notable results, with outputs as follows:

Activity 1: this activity is designed to enable students to recognize both the principal and secondary characters in the story by checking their frequency of occurrence in the corpus. The students use the Wordlist tool to view the data as illustrated in Figure 04. Characters with high occurrences are assumed to be principal characters, while those with low occurrences are secondary characters.

Table 1 displays a general frequency analysis of the two groups. This kind of analysis will help students understand the character roles better and their associated effects on the development of action in the story.

Table 1. Frequencies of characters 'names in the pedagogic corpus.

Characters 'names	Frequencies
سعيد	**116**
الشيخ	**90**
رؤوف	**89**
نور	**67**
عليش	**55**
علوان	**47**
سناء	**26**
طرزان	**24**
بوليس	**24**
معلم	**22**

Activity 2: students will engage in an exploration of common verbs through an analysis of their frequency within the corpus through the use of the Wordlist tool. Figure 05 displays the top 100 most common verbs in a tabulated form, ranked according to their frequency of use. Of these, Table 2 highlights the top 10 verbs with the highest frequency of occurrence, which are particularly useful in gaining insight into their importance within the texts.

Through a close analysis of these frequencies, the students will look into the different contexts where these verbs are used across the corpus, which would deepen their understanding of the language in use. They will also use the Concordance tool to make a deep analysis of the verb "قتل." They will make a categorization of its occurrences according to tense: past, present, and future, as shown in Figure 06. This will not only polish their verb recognition but also enrich their overall comprehension of how verbs function within different temporal contexts.

Table 2. Distribution of 10 frequent verbs in the corpus.

Verb	Frequencies
Said "قال"	**326**
Was "كان"	**196**
Returned "عاد"	**64**
Found "وجد"	**58**
Knew "عرف"	**55**
Wondered "تساءل"	**51**
Killed "قتل" –	**51**
Wanted "أراد"	**42**
Loved "أحب"	**36**
Proceeded "مضى"	**31**

Activity 03: using the word sketch tool, students will be able to identify the collocates of the word "صوت" (sound), and its grammatical patterns, and display them as a word cloud, as demonstrated in Figures 7 and 8. Table 3 shows the grammatical patterns of the word's collocates of "صوت" (sound).

Table 3. Grammatical patterns of the collocates of the words of "صوت" (sound).

Pattern	Verbs	Modifiers
Verbs with "صوت" *as* subject	Sang *"غنى"* Heard *"سمع"* Knew *"عرف"* Said *"قال"*	
Verbs with "صوت" *as* object	Sang *"غنى"* Said *"قال"*	
Modifiers of "صوت"		Audible "*مسموع*" Rough "*غليظ*" Dignified "*وقور*" Resplendent (or melodious, depending on context)"*رخيم*" Arrogant "*متكبر*" Harsh "*خشن*" Angry "*غاضب*" Choked *"مختنق"*

Activity 04: students will delve into the world of figurative language, specifically the similes using the comparative particle "like" کے, as in Figure 15. They will look at several similes to see how they manage to paint vivid pictures and express emotions. Figure 15 indicates that the students will separate the phrases into those expressing positive emotions and those showing negative emotions. Then they will chart the frequency of each, noting finally that *negative emotions* have a higher rate of occurrence. This will help them better understand the use of figurative language in the expression of emotions and that the tone in a piece of writing greatly varies according to the use of similes.

Activity 5: students engage in a more thorough linguistic component analysis by identifying parts of speech for every word in three given sentences. They will

examine the sentences carefully and label each word as a noun, verb, adjective, adverb, or whatever other grammatical classification is appropriate. In order to make the students' work easier to understand, the part-of-speech tags will be shown in Figure 10. Along with their analysis, the students *will translate these tags into Arabic,* increasing their knowledge of how English and Arabic languages interrelate. This activity will help them solidify their understanding of grammatical structures and improve their analytical skills.

6. Discussion

Corpora hold considerable importance in many areas of applied linguistics, especially in the area of language teaching, where they are both a teaching tool and a rich source of linguistic material. Many countries, particularly those where English is the mother tongue, have exploited these corpora to develop teaching methods and approaches by integrating them into the teaching and learning systems of the English language in classrooms. In this way, corpora have become tools used by both teachers and students in carrying out many activities and exercises based on them.

The application of educational activities based on the pedagogical corpus, using Sketch Engine, yields tangible results, as this approach contributes to the development of Arabic language learning competencies, such as text analysis, word analysis in context, grammar learning, identifying linguistic patterns, exploring syntactic structures, and studying collocations, among other competencies. The five educational activities designed focus on different aspects of the novel, as previously detailed in the earlier sections. Through the first activity, students will become familiar with the primary and secondary roles of the characters in the novel, providing them with a preliminary understanding of its contents. The second activity will help students gain a deep understanding of verbs and acquire the ability to use them flexibly in different contexts, especially when classified by tense and analyzing the sentences that include them through the contextual search tool. This will enable students to construct correct and varied sentences, enhancing their skills in both written and oral expression. The third activity highlighted the importance of using tools such as contextual search and word sketch tools to explore collocations and syntactic structures related to keywords in the text. This contributes to enhancing students' understanding of recurring linguistic and syntactic patterns, helping them interpret the relationship between words and meanings in various contexts. In the fourth activity, students will gain a deep understanding of rhetorical figures such as metaphors and similes, learning how to use them to enhance literary meanings. The contextual search tool aided in exploring specific examples of metaphors and similes in the text, thus improving their ability to analyze literary texts and comprehend the rhetorical impact of texts. The fifth and final activity provides students with exercises to study verb conjugations, analyze conditional sentences, and understand temporal and spatial adverbs, thereby contributing to the development of their writing and linguistic analysis skills. Additionally, the use of word tags and structured query language helps students improve their ability to search and extract linguistic information from texts accurately and effectively.

Despite the significant importance of corpora in the field of language teaching and learning, Arabic suffers from a notable lack of such resources, which negatively impacts the quality of linguistic research and limits learning and development opportunities in this area. In this context, Mansour (2013: 85) proposes the creation of a national corpus for the Arabic language, which would provide teachers and students with the opportunity to apply a variety of activities and exercises that enhance linguistic and cognitive competencies. Reppen (2010: 44) also emphasized that corpora are a rich source of materials that reflect the actual and real use of language, with which students interact outside the classroom. The integration of these corpora into the educational process is seen as a model for incorporating authentic materials into the classroom, a once significant challenge.

Integrating corpora into Arabic language classrooms yields significant benefits for both teachers and students. For instructors, using corpora enhances pedagogical methodologies by facilitating lesson preparation, content explanation, the design of educational activities and exercises, and the development of assessments and evaluations. Nevertheless, certain challenges may impede teachers' willingness to adopt corpora, as their implementation may be perceived as technical, despite its fundamentally educational nature (Lee 2011: 176). To mitigate these challenges, comprehensive training for teachers is essential. Empirical studies indicate that teachers demonstrate a heightened readiness to incorporate corpora in the classroom following specialized training in this domain (Chen et al. 2019; Abdel Latif 2021), which necessitates the acquisition of a distinct skill set commonly referred to as "corpus pedagogy." This skill set empowers teachers to effectively leverage corpora to improve student performance (Ma, Tang and Lin 2022: 2733).

In terms of teachers' comprehension of corpora, Kreyer, Shacub and Güldenring (2016: 395) identifies several critical dimensions, including an understanding of foundational concepts such as the definition of a corpus, its various types, and its practical applications. This comprehension also encompasses research and analytical competencies, which involve selecting appropriate corpora and employing software tools for data analysis and interpretation. Moreover, this knowledge integrates critical thinking skills, focusing on the interpretation of results, the connection of findings to linguistic theories, the evaluation of data reliability, and the analysis of potential biases. Finally, the pedagogical application of corpora is of paramount importance, as it entails designing innovative educational activities that utilize corpora and the assessment of their effectiveness in meeting educational objectives.

For students, Sinclair (1997: 38) argues that teaching-based corpora allow learners to develop their understanding of what language is about; promote creativity, and give them the skills needed to overcome the obstacles posed by natural English, which results in a beneficial impact on different dimensions of linguistic competence. As for students using corpora, Frankenberg-Garcia (2012: 50) argues that we do not need to have a lot of knowledge on the matter; in fact, simply giving our students some orientation can gradually empower them so they

will start exploring how to use corpora and their tools by themselves without the teacher helping all the time.

7. Recommendations

Our study concludes with a few important recommendations:

1. It is essential to strengthen partnerships between the ministry responsible for education and experts in various fields, including education, corpus linguistics, computational linguistics, Arabic language processing, and language pedagogy. This collaboration should primarily focus on the design and creation of educational corpora, the development of analytical tools, and their effective integration into the educational process.
2. Organize special training courses for teachers to enable them to use corpora effectively in the classroom.
3. There is a great need to encourage more research and studies on applying corpora in teaching and learning Arabic. Such research should be steered towards developing innovative, efficient teaching practices that constitute a strong scientific base for this field.
4. Providing classrooms with computers or interactive whiteboards (smart boards) and linking them to a computer that will enable the corpus-based approach to teaching Arabic.

8. Conclusion

Given the modern developments in language pedagogy and the current dependence on the use of modern technological tools, the incorporation of corpora in Arabic language classrooms presents a major challenge for the improvement of educational quality. This will particularly attempt to improve practices in teaching the Arabic language in the Arab world, raise teacher performance, and develop both linguistic and cognitive competencies among students. It will also strengthen students' abilities in communication and analysis. This methodology will relate the acquisition of the Arabic language to modern linguistic contexts, by nurturing a cadre of learners who possess language competencies corresponding to the requirements of the digital age.

However, the biggest challenge remains the development of specifically designed Arabic corpora that align with learning curricula and syllabi, therefore increasing their utility in teaching environments. There is also an overwhelming need for the advancement of software tools that support the Arabic language, allowing it to make full use of its special characteristics.

Acknowledgments

We would like to express our gratitude to the Center of Excellence in Arabic Language at Mohammed Bin Zayed University for Humanities for supporting academic and research projects in this field. Their efforts to promote research are inspiring and help advance linguistic studies.

Hafid Maachi – Corresponding Author
PhD Student, affiliated with the Research Laboratory: Arabic Teaching and Applied Linguistics Studies
Mohammed V University in Rabat, Rabat, Morocco
ORCID Number: 0009-0000-9909-9706
Email: hafid.maachi@um5r.ac.ma

Hakima Khamar
Professor of Higher Education, Department of Arabic Language.
Mohammed V University in Rabat, Rabat, Morocco.
ORCID Number: 0009-0007-1082-0933
Email: hakima.khamar@flsh.um5.ac.ma

Lutfi Omar Abubkr
Associate Professor, Department of Arabic Language.
Mohamed bin Zayed University for Humanities, Abu Dhabi, United Arab Emirates.
ORCID Number: 0000-0002-5055-9895
Email: lutfi.bubkr@mbzuh.ac.ae

References

Abdel Latif, Muhammad M. (2021). 'Corpus literacy instruction in language teacher education: Investigating Arab EFL student teachers' immediate beliefs and long-term practices.'*The Journal of the European Association for Computer Assisted Language Learning*, 33(1): 34-48. https://doi.org/10.1017/S0958344020000129.

Alfaifi, Abdullah and Eric S. Atwell. (2012). 'Arabic learner corpora (ALC): A taxonomy of coding errors'. *International Computing Conference in Arabic*, Cairo, Egypt.

Alfuraih, Reem and Noha El-Jasser. (2024). 'Exploitation and evaluation of an Arabic-English composite learner translator corpus'. *International Journal of Arabic-English Studies*, 24(1): 155–172. https://doi.org/10.33806/ijaes.v24i1.552

Almujaiwel, Sultan and Abdulmohsen Al-Thubaity. (2016). 'Arabic corpus processing tools for corpus linguistics and language teaching'. *The Globalization of Second Language Acquisition and Teacher Education Conference*, Fukuoka.

Alotaibi, Hind M. (2017). 'Arabic-English parallel corpus: A new resource for translation training and language teaching'. *Arab World English Journal*,8(3): 319-337.

Al-Sabbagh, Rania. (2024). 'The negative transfer effect on the neural machine translation of Egyptian Arabic adjuncts into English: The case of Google translate'. *International Journal of Arabic-English Studies*, 24(1): 95–118. https://doi.org/10.33806/ijaes.v24i1.560

Al-Sulaiti, Latifa and Eric S. Atwell. (2006). 'The design of a corpus of contemporary Arabic'. *International Journal of Corpus Linguistics*, 11(2): 135-171.

Aston, Guy. (2001). 'Learning with corpora: An overview.' In Guy Aston (ed.), *Learning with Corpora*, 7-45. Houston: Athelstan. https://digital.casalini.it/8849117019 - Casalini id: 2250267

Bennett, Gena. (2010). *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers.* Michigan: The University of Michigan Press. https://doi.org/10.3998/mpub.371534

Boulton, Alex. (2010). 'Data-driven learning: Taking the computer out of the equation'. *Language Learning*, 60(3):534-572.

Boulton, Alex. (2011). '*Data-driven learning: The perpetual enigma*.' In G. Gilquin, S. Granger and F. Meunier (eds.), *Exploring the Landscape of Learner English*. John Benjamins.

Boulton, Alex. (2017). 'Corpora in language teaching and learning. *Language Teaching*, 50(4): 483-506. https://doi.org/10.1017/S0261444817000167

Braun, Sabine. (2005). 'Corpus-based teaching materials: Pedagogical applications of corpora'. *Linguistic Insights: Studies in Language and Communication*, 22: 53-69.

Braun, Sabine. (2007). 'Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora'. *The Journal of the European Association for Computer Assisted Language Learning*, 19(3): 307-328. https://doi.org/10.1017/S0958344007000535

Kreyer, R., S. Schaub and B. Güldenring. (2016). 'Towards corpus literacy in foreign language teacher education: Using corpora to examine the variability of reporting verbs in English'. *Angewandte Linguistik in Schule und Hochschule*, 1: 391-415.

Chen, Meilin, John Flowerdew and Laurence Anthony. (2019). 'Introducing in-service English language teachers to data-driven learning for academic writing'. *System*, 87: 102-148. https://doi.org/10.1016/j.system.2019.102148

Conrad, Susan and Kimberly R. LeVelle. (2008). 'Corpus linguistics and second language instruction.' *The Handbook of Educational Linguistics*, 1: 539-556. https://doi.org/10.1002/9780470694138.ch38

Flowerdew, **John.** (2009). 'Corpora in language teaching.' In Michael Long and Catherine Doughty (eds.), *The Handbook of Language Teaching*, 327-350. New Jersey: Wiley-Blackwell. https://doi.org/10.1002/9781444315783.ch19

Frankenberg-Garcia, Ana. (2012). 'Integrating corpora with everyday language teaching.' In James Thomas and Alex Boulton (eds.), *Input, Process and Product: Developments in Teaching and Language Corpora*, 36-53. Brno: Masaryk University Press.

Friginal, Eric. (2018). *Corpus Linguistics for English Teachers: New Tools, Online Resources, and Classroom Activities.* London: Routledge.

Fuster-Márquez, Miguel and Carmen Gregori-Signes. (2018). 'Learning from learners: A non-standard direct approach to the teaching of writing skills in EFL in a university context'. *Innovation in Language Learning and Teaching*, 12(2): 164-176. https://doi.org/10.1080/17501229.2016.1142549

Götz, Sandra and Sylviane Granger. (2024). 'Learner corpus research for pedagogical purposes: An overview and some research perspectives.' *International Journal of Learner Corpus Research*, 10(1): 1-38. https://doi.org/10.1075/ijlcr.00039.got

Hunston, Susan. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524773.

Johansson, Stig. (2009). '*Some thoughts on corpora and second language acquisition*.' In Karin Aijmer (ed.), *Corpora and Language Teaching*, 33-44. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.33.05joh

Johns, Tim. (1991). 'Should you be persuaded: Two examples of data-driven learning'. *English Linguistic Research Journal*, 4: 1-16.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz and David Tugwell. (2004). *The Sketch Engine*. In Geoffrey Williams and Sandra Vessier (eds.), *Proceedings of the 11th EURALEX International Congress*, 105-115. Lorient: Université de Bretagne-Sud.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz and David Tugwell. (2014). 'The Sketch Engine: Ten years on'. *Lexicography*, 1(1): 7-36.

Lee, Shinwoong. (2011). 'Challenges of using corpora in language teaching and learning: Implications for secondary education'. *Linguistic Research*, 28(1): 159-178.

Leech, Geoffrey. (1997). Teaching and language corpora: A convergence'. In Anne Wichmann, Steven Fligelstone, Tony McEnery and Gerry Knowles (eds.), *Teaching and Language Corpora*, (1–23). London: Longman.

Levchenko, Victor. (2017). 'Use of corpus-based classroom activities in developing academic awareness in doctoral students.' *The New Educational Review*, 48: 28-40.

Ma, Qing, Jinlan Tang and Shanru Lin. (2022). 'The development of corpus-based language pedagogy for TESOL teachers: A two-step training approach facilitated by online collaboration'. *Computer Assisted Language Learning*, 35(9): 2731-2760. https://doi.org/10.1080/09588221.2021.1895225

Ma, Qing, Yuan Eric R., Cheung Lisa and Jian Yang. (2022). 'Teacher paths for developing corpus-based language pedagogy: A case study'. *Computer Assisted Language Learning*, 37(3): 461–492. https://doi.org/10.1080/09588221.2022.2040537

Mansour, Mahmoud A. (2013). 'The absence of Arabic corpus linguistics: A call for creating an Arabic national corpus'. *International Journal of Humanities and Social Science*, 3(12): 81-90.

McEnery, Tony. (2006). *Corpus-Based Language Studies: An Advanced Resource Book.* New York: Routledge.

McEnery, Tony and Andrew Hardie. (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, Tony and Xiao Richard. (2011). 'What corpora can offer in language teaching and learning'. In Eli Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning*, 364-380. New York: Routledge.

Nesselhauf, Nadja. (2004). '*Learner corpora and their potential for language teaching.*' In John Sinclair (ed.), *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.12.11nes

Nola Bacha, Nahla and Victor Khachan. (2023). 'A Corpus-based lexical evaluation of L1 Arabic learners' English literary essays.' *International Journal of Arabic-English Studies*, 23(2): 415–442. https://doi.org/10.33806/ijaes.v23i2.470

O'Keeffe, Anne, Michael McCarthy and Ronald Carter. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511497650

Reppen, Randi. (2010). *Using Corpora in the Language Classroom*. Cambridge: Cambridge University Press

Römer, Ute. (2006). 'Pedagogical applications of corpora: Some reflections on the current scope and a wish list for future developments.' *Zeitschrift für Anglistik und Amerikanistik*, 54(2): 121-134. https://doi.org/10.1515/zaa-2006-0204

Römer, Ute. (2008). 'Corpora and language teaching.' InAnke Lüdeling and Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*, 112-130. Berlin: Mouton de Gruyter.

Römer, Ute. (2010). 'Using general and specialized corpora in English language teaching: Past, present, and future.' In M. C. Campoy-Cubillo, B. Bellés-Fortuño, and L. Gea-Valor (eds.), *Corpus-Based Approaches to English Language Teaching*, 18–35. London: Continuum.

Römer, Ute. (2011). 'Corpus Research Applications in Second Language Teaching'. *Annual Review of Applied Linguistics*, 31:205–225. https://doi.org/10.1017/S0267190511000055

Sinclair, John. (1997). 'Corpus evidence in language description.' In Anne Wichmann, Steven Fligelstone, Tony McEnery and Gerry Knowles (eds.), *Teaching and Language Corpora*, 27–39. London: Longman.

Sinclair, John. (2004). 'Introduction'. In John Sinclair (ed.), *How to Use Corpora in Language Teaching*, 1-10. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.12.02sin

Whitcomb, Laura and Sameh Alansary. (2017). 'Using linguistic corpora in Arabic foreign language teaching and learning'. In Kassim Wahba, Liz England and Zeinab A. Taha (eds.), *Handbook for Arabic Language Teaching Professionals in the 21st Century*, 219-231. New York: Routledge. https://doi.org/10.4324/9781315676111

Williams, Geoffrey. (2002). 'In search of representativity in specialised corpora: Categorisation through collocation'. *International Journal of Corpus Linguistics*, 7(1): 43-64. https://doi.org/10.1075/ijcl.7.1.03wil

Zaki, Mai. (2017). 'Corpus-based teaching in the Arabic classroom: Theoretical and practical perspectives.' *International Journal of Applied Linguistics*, 27: 514–541. doi: 10.1111/ijal.12159