applied sciences

MDPI

Article

The Saudi Novel Corpus: Design and Compilation

Tareq Alfraidi [1,*], Mohammad A. R. Abdeen [2], Ahmed Yatimi [3], Reyadh Alluhaibi [4] and Abdulmohsen Al-Thubaity [5]

[1] Department of Linguistics, Islamic University of Madinah, Madinah 42351, Saudi Arabia
[2] Department of Computer Science, Islamic University of Madinah, Madinah 42351, Saudi Arabia; mabdeen@iu.edu.sa
[3] Department of Literature and Rhetoric, Islamic University of Madinah, Madinah 42351, Saudi Arabia; alyatimi@iu.edu.sa
[4] Department of Computer Science, Taibah University, Madinah 41477, Saudi Arabia; rluhaibi@taibahu.edu.sa
[5] King Abdulaziz City for Science and Technology, Riyadh 12354, Saudi Arabia; aalthubaity@kacst.edu.sa
* Correspondence: t.alfraidi@iu.edu.sa

Abstract: Arabic has recently received significant attention from corpus compilers. This situation has led to the creation of many Arabic corpora that cover various genres, most notably the newswire genre. Yet, Arabic novels, and specifically those authored by Saudi writers, lack the sufficient digital datasets that would enhance corpus linguistic and stylistic studies of these works. Thus, Arabic lags behind English and other European languages in this context. In this paper, we present the Saudi Novels Corpus, built to be a valuable resource for linguistic and stylistic research communities. We specifically present the procedures we followed and the decisions we made in creating the corpus. We describe and clarify the design criteria, data collection methods, process of annotation, and encoding. In addition, we present preliminary results that emerged from the analysis of the corpus content. We consider the work described in this paper as initial steps to bridge the existing gap between corpus linguistics and Arabic literary texts. Further work is planned to improve the quality of the corpus by adding advanced features.

Keywords: corpora; corpus linguistics; Arabic; Saudi novels

check for updates

Citation: Alfraidi, T.; Abdeen, M.A.R.; Yatimi, A.; Alluhaibi, R.; Al-Thubaity, A. The Saudi Novel Corpus: Design and Compilation. *Appl. Sci.* **2022**, *12*, 6648. https://doi.org/10.3390/app12136648

Academic Editor: Kuei-Hu Chang

Received: 20 May 2022
Accepted: 27 June 2022
Published: 30 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Due to the evolution of the digital age, corpus linguistics has found its way in research communities and has proven valuable in language study in various ways in recent decades. Today, corpus linguistics has now established itself as a distinct discipline. The term corpus linguistics typically refers to two practices. It can refer to the practice of creating a (large) dataset that mirrors or represents the use of a language or one of its varieties. The term can also refer to the use of such a dataset to analyze and investigate various aspects of the language(s) under consideration [1,2]. This paper focuses on the first practice.

Corpora, as "collection[s] of texts … stored in an electronic database" [3], represent valuable resources for language documentation, Natural Language Process application, and linguistic analysis. Indeed, the use of corpus analysis is considered to be "the most powerful empirical approach for analyzing the patterns of language use" [4]. Although this approach is commonly used to analyze the linguistic components of language, such as lexis, grammar, or semantics, it has recently emerged as useful across interdisciplinary areas of research due to its successful interaction with different disciplines [5]. These include sociolinguistics [6,7], discourse analysis [8], translation [9], literature [10], forensic studies [11], language teaching [12], and social media [13]. The intersection between corpus linguistics and literature recently formed its own interdisciplinary area of research called 'Corpus Stylistics' [4,14]. In this area, the use of the techniques and tools of corpus linguistics in the study of literary texts, particularly novels, has proven fruitful. This successful achievement appears to be a result of this application's capacity to facilitate

the relatively objective examination of large data, which is computationally processed, in a systematic way [15]. Examples of existing studies include [16–18]. Although such an analysis practice has become popular among the research community of corpus linguistics and stylistics in English and some other languages, it has been ignored in the context of the Arabic language. This situation is likely due to the absence of a specialized corpus of texts that represent Arabic literature or its subgenres (e.g., novels), as shown in Section 2.3. Texts from the news genre, on the other hand, have been the primary focus of most Arabic corpus compilers.

The central goal that the authors of this paper wish to achieve is to fill such a gap and contribute to the field of Arabic corpus stylistics, a neglected area of research. One crucial step that must be taken is to build a corpus that facilitates the studies of Arabic stylistics. Hence, this paper describes the processes we have taken and our decisions to create the Saudi Novel Corpus. Specifically, it presents the decisions we made regarding the design criteria of the corpus and the subsequent steps we followed to construct it.

One might ask: Why do we need a specific corpus for Saudi novels? Apart from the gap highlighted above, we are motivated to compile such a corpus because of the following. Novels are one of the strongest features of the literary movement in Saudi Arabia, due to their artistic excellence, an abundance of production, and occupation of a prominent position at the cultural, media, and academic levels, both locally and in the broader Arab world [19]. One aspect of this prominent position is the dramatic increase in the number of Saudi novels in recent years: It is reported that approximately 1571 Saudi novels have been published over the last two decades [19]. This intense presence of the Saudi novel has drawn the attention of many researchers of Arabic literature and criticism to conduct numerous studies on different aspects of Saudi novels. According to [19], around 217 studies (printed as books) analyze the style of Saudi novels and their aesthetics and artistic and linguistic features. In addition, there are approximately 162 unpublished MA and Ph.D. theses conducted at different Saudi universities targeting the same scope [20]. These numbers signal the importance of Saudi novels and their position in the Arabic literary landscape.

However, one fundamental shortcoming stems from these studies: not one of them applies corpus linguistics methods as far as we know. In other words, there is an apparent absence of the quantitative approach and corpus methodology in the aforementioned studies, demonstrating the inability to explore different aspects of large data from Saudi novels. This is likely to have emerged due to the unattainability of a computerized dataset of Saudi novels. Therefore, compiling such a corpus of Saudi novels will give scholars the opportunity to enhance their studies and will introduce them, through the use of corpus linguistics techniques, to valuable results that are not expected when relying only on human intuition.

The rest of the present paper is organized as follows: Section 2 provides a background of the interface between corpus linguistics and the field of literature and stylistics in a broader context and presents some global practices. This is followed by sub-sections that review both the current non-Arabic literary corpora and the Arabic corpora, showing the gap in the relation between corpus linguistics and Saudi novels. Section 3 is devoted to the design criteria of the corpus, and Section 4 focuses on the methodological steps we followed to create it. Section 5 discusses the issue of copyright and the measures we applied to keep our work in line with regulations. Section 6 demonstrates some statistical visualization of different aspects of the content of the corpus and provides some experimental results. Finally, we conclude the paper with some remarks and future suggested improvements.

2. Background and Related Works

2.1. Corpus Linguistics and Literature

In the literature, the study of literary texts is normally referred to by the term 'stylistics,' which is perceived as a "field of empirical inquiry, in which the insights or techniques of linguistic theory are used to analyze literary texts" [21]. Thus, this field is linguistically

oriented, as linguistic models and tools play a major part in the research linked to it. In addition, stylistics' primary goal is to connect the formal linguistic features of a text to the interpretation of meaning encoded within it [15,17]. Recently, stylistics has been successfully combined with corpus linguistics to facilitate the analysis of literary texts [4,22]. According to [21], the empirical character of stylistics and corpus linguistics has had a significant role in the mutual relationship established between them. Consequently, such interaction has created the area of 'corpus stylistics', which can be defined as the study of the style of literary texts using corpus linguistics methods to aid the analysis of textual meanings with empirical evidence [15,23].

One of the essential features of corpus stylistics is the application of computationally sophisticated methods to examine a large dataset (corpus) of literary texts. This application helps achieve a large number of statistical observations that would be difficult to detect in a manual study [15].

The growing popularity of corpus stylistics has motivated corpus linguists and stylisticians (i) to create several corpora, including literary texts and specifically novels [24,25], and (ii) to analyze these texts through the application of corpus stylistic techniques. Examples of these techniques include the investigation of keywords [14,16], concordance [17,18], and cluster [22]. A special focus on novels is warranted, as they constitute one of the most popular written subgenres of literature in many languages, and they are consumed by many readers. One potential reason for this tendency is that novel is seen as an art form capable of embracing intellectual, moral, social, and political issues and presenting them in appealing ways that are, in many cases, parallel to reality. This may also explain why novels are popular research targets among stylisticians and literary critics.

In the context of English and other European languages, several corpora containing texts from novels exist to facilitate the use of corpus methods and tools (i.e., applying the corpus stylistic approach). More details about such corpora are given in Section 2.2.

As a result of the emergence of corpus stylistics, a considerable number of studies on novels have been conducted and have revealed the fruitfulness of the methods utilized in this field. In [16], for example, has analyzed the linguistic style of Joseph Conrad's *Heart of Darkness*. He specifically investigated the keywords, collocations, and word/phrase distributions. In her seminal study, [17] conducted an intensive corpus analysis of Jane Austen's novels along with those of some of her contemporaries. Fischer-Starcke in [17] identified keywords and investigated concordance lines, enabling her to observe dominant topics and analyze some semantic fields (e.g., textuality). Similarly, Wijitsopon in [14] also applied corpus-driven techniques to analyze six of Austen's novels. This approach allowed him to identify lexical items and linguistic patterns that characterize the style of Austen's fictional writing. He then emphasized that using the corpus approach in his analysis was beneficial because the results he obtained supported and refuted some scholars' intuitive observations on Austen's works. Using the concordance function available in CLiC software, Mahlberg in [18] investigated the language of Dickens's novels. They observed some meaningful patterns, which they reported to be mostly associated with text-internal variations. In a recent study, Nais in [26] used the corpus method to analyze the style of Henry James's *The Portrait of a Lady* and Edith Wharton's *The Age of Innocence* to determine patterns of metaphorical instances of two words: 'seeing' and 'vision'.

In the context of other European languages, Mostafa in [27] has analyzed a corpus of fictional Spanish comprised of 19 texts written by three Spanish writers. The focus of the study is on the use of the word arabe. Using certain corpus techniques, including word clusters and concordance, the author identified several stylistic differences between the writers. Kubis in [28] used the corpus method to study the character networks of Polish novels. He set two-time frames, the 19th- and 20th-centuries, to observe whether stylistic changes occurred during those timeframes.

The above review is not intended to provide a detailed critical review of studies applying the corpus stylistic approach. Rather, it aims to show several major contributions of these studies to the analysis of novels. Such studies have provided the field of stylistics

with new insights that were not previously possible without the use of data derived from the corpora of novels.

2.2. Non-Arabic Corpora

Currently, there are many European language corpora that were created for different purposes and developed to serve as either general/reference or specialized corpora. In the former, the developer's primary aim is to compile a very large corpus containing a variety of genres and domains of the language under consideration. In contrast, in the latter, corpus developers target a specific genre (e.g., newspaper, literature, academic writing) or context (e.g., translation, language teaching) [3]. As this research focuses on literature, we shed light on the construction of specialized corpora containing literary texts, and novels in particular.

Overall, the existing corpora of literature can be divided into two major categories: monolingual and bi-/multilingual [29]. The former includes only texts from only one language, while the latter consists of texts from two or more languages (ibid). The existing monolingual corpora can be further divided into two sub-categories. The first category includes those corpora that contain texts produced by only one author. The Corpus of George MacDonald's Fiction is an excellent example of this type [23]. This corpus contains 41 texts (mainly novels) written by the Scottish author George MacDonald and published between 1858 and 1897. The corpus is comprised of approximately 4.5 million words tagged with POS classes. Another example of this sub-category is the DNov corpus, which contains the novels of the well-known English writer Charles Dickens [18]. This corpus was built through the CLiC Dickens Project, a collaboration between the University of Birmingham and the University of Nottingham. It contains 15 novels which collectively comprise around 3.8 million words. The DNov texts were also tagged along different elements, including quotation, suspension, and sentence. This corpus is integrated into a web interface designed to facilitate the efficient use of the corpus, available via: https://clic.bham.ac.uk, accessed on 12 April 2022.

Another corpus of the same category but from another European language is the Contemporaneous Spanish Novels Corpus, created in 2005, available via: http://catalog.elra.info/en-us/repository/browse/ELRA-W0041/, accessed on 12 April 2022. It contains 11 novels (638,099 words) written by Inmaculada Ferrer-Vidal Turull. However, access to this corpus is not free; therefore, searching its content is restricted. A final example of this sub-category is the Corpus of August Strindberg's novels, created in 2017 to represent the Swedish writer August Strindberg, available via: https://spraakbanken.gu.se/en/resources/strindbergromaner, accessed on 12 April 2022. The corpus contains around 4.3 million words. Notably, this corpus is not devoted to novels alone, as texts from drama form a portion of it. The corpus content can be searched via the concordance tool 'Korp' [30]. Although these corpora are beneficial for analyzing the features of the linguistic style of a particular writer, the results obtained from them cannot be generalized and extended to the style of other novels in the language under consideration.

The second sub-category of monolingual corpora includes those that comprise a range of novels authored by different writers. This includes the Corpus of Eighteenth-Century Prose Fiction [31]. This corpus contains 143 literary texts (e.g., novels, short stories) written by 89 writers, comprising around 9.7 million words. A distinctive feature of this corpus is the chronological presentation of the included data, which are split into three periods. This feature enables diachronic investigations of the data to observe any possible linguistic and stylistic development that may have occurred. A similar approach has been applied to the construction of the Corpus of the Canon of Western Literature. The developer has divided the data into four chronological periods that reflect different literary movements [24]. However, this corpus is much larger than the first example given above (around 72 million words), and it is not built only from novels, as poems and plays form a portion of the data. In the CLiC Dickens Project mentioned above, the developers have recently extended their contribution beyond collecting only Dickens novels. They created

the 19th Century Reference Corpus [10] to serve as a reference corpus for contextualizing the findings obtained from DNov, i.e., to trace any possible link between the style of 19th century writers and that of Dickens. The corpus contains 29 novels produced by prominent British writers, amounting to around 4.5 million words. Four other non-English corpora can also be classified under this sub-category, namely: the Great Corpus of French and French-speaking Literatures (available via: https://tinyurl.com/2p974fvf, accessed on 12 April 2022), the Frantext corpus: French literature of the 18th–20th century (available via: https://tinyurl.com/yfphdwbw, accessed on 12 April 2022), the Ukrainian Novel Corpus (available via: https://tinyurl.com/2a6t49z2, accessed on 12 April 2022), and the Corpus of Estonian Fiction (available via: https://tinyurl.com/2tbdw539, accessed on 12 April 2022).

As for bi-/multilingual, the most common sub-type in this category is known as parallel corpora, which contain texts from one (source) language and the equivalent translations into one or more other languages [29]. There are several examples that can be given here. The Corpus of Romanian–English Literature was created in 2016 and contains texts from novels, drama, and other literary books, available via: https://tinyurl.com/9vcprdrf, accessed on 12 April 2022. The original in English contains 87,681 words, whereas the translation in Romanian contains 88,498 words. This size may be regarded as rather small, however. The following two complementary corpora exemplify larger parallel corpora: ParFin 2016, the Finnish–Russian Parallel Corpus of Literary Texts [32], and the Russian–Finnish Parallel Corpus of Literary Texts, available via: https://tinyurl.com/2thn7dcb, accessed on 12 April 2022. The former contains literary texts from Finnish and their translations into Russian, with a word count of 2,044,172. On the other hand, the latter contains literary texts from Russian and their translations into Finnish, with a word count of 5,900,000. Both corpora are integrated into the web interface tool 'Korp', available via: https://tinyurl.com/27wbhkk5, accessed on 12 April 2022. Another interesting example of the parallel corpora of literary texts is the MULTEXT-East '1984' Annotated Corpus [33]. This corpus contains the novel *1984* (by British author George Orwell) in its original language English, along with its translation into 11 languages with a word count of 1,064,424. The data included is morpho-syntactically tagged and lemmatized.

Our review of non-Arabic literary corpora is not intended to cover all founding, relevant corpora comprehensively. Instead, it aims to present the significant contributions that corpus linguistics has made to the field of literature through the practice of corpora construction. Nevertheless, although targeting the same scope, these corpora differ in terms of their target language, time span, size, and the popularity of the writers. As a result, some of these corpora are suitable for diachronic stylistic studies, others are ideal for analyzing the style of certain writer(s) or language(s), and still others are useful for comparing different writers' styles. This leads us to the question: What is the status of the existing Arabic corpora? We provide the answer in the following sub-section.

2.3. Arabic Corpora

Currently, many Arabic corpora exist to serve different purposes, each varying in size, content, availability, and method of annotation. These corpora are either specialized or general. At the outset, we can assert that the majority of the texts in the existing specialized Arabic are derived from online newspapers [34], including the Al-Haya Corpus (available via: https://catalogue.elra.info/en-us/repository/browse/ELRA-W0030/, accessed on 12 April 2022), Al-Watan Corpus (available via: https://sourceforge.net/projects/arabiccorpus/files, accessed on 12 April 2022), Akhbar El-Khaleeg (available via: https://sourceforge.net/projects/arabiccorpus/files, accessed on 12 April 2022), Arabic Newswire Corpus (available via: https://catalog.ldc.upenn.edu/LDC2001T55, accessed on 12 April 2022), Al-Nahar (available via: http://catalog.elda.org/en-us/repository/browse/ELRA-W0027/, accessed on 12 April 2022), Arabic Gigaword Corpus (available via: https://catalog.ldc.upenn.edu/LDC2011T11, accessed on 12 April 2022), Abu El-Khair Corpus (available via: http://www.abuelkhair.net/index.php/en/arabic/abu-el-khair-corpus, accessed on 12 April 2022), Bangor Arabic Annotated Corpus (available

via: https://thesai.org/Publications/ViewPaper?Volume=9&Issue=11&Code=IJACSA&SerialNo=20, accessed on 12 April 2022), and ANTCorpus (available via: https://antcorpus.github.io/, accessed on 12 April 2022). Other specialized corpora were designed to represent particular genres and contexts of Arabic, excluding Arabic novels. Examples of these corpora include the Quranic Arabic Corpus (available via: https://corpus.quran.com/, accessed on 12 April 2022), KACST Error Corrected Corpus (available via: https://ieeexplore.ieee.org/abstract/document/6193415, accessed on 12 April 2022), Arabic Learner Corpus (available via: https://www.arabiclearnercorpus.com/, accessed on 12 April 2022), Tunisian Arabic Corpus (available via: http://www.tunisiya.org/, accessed on 12 April 2022), Curras (a corpus of Palestinian dialect, available via: https://portal.sina.birzeit.edu/curras/, accessed on 12 April 2022), SUAR (a corpus of Saudi dialect, available via: https://www.sciencedirect.com/science/article/pii/S187705091832163X, accessed on 12 April 2022), and AraSenCorpus (a corpus of Arabic tweets, available via: https://github.com/yemen2016/AraSenCorpus, accessed on 12 April 2022). More details about these corpora and others are reported in [35–37].

Likewise, several general corpora of Arabic, which contain a variety of text types and genres and are built for a number of purposes, ignore Arabic (including Saudi) novels altogether, and exclude them from their data. Examples of these corpora are the Corpus of Contemporary Arabic [35] (The developer of this corpus states that the corpus contains around 21,958 words from short stories. This genre, however, is regarded as different from novels.), Open Source Arabic Corpus [38], KALIMAT [39] and arTenTen [40], and KACST Text Classification Corpus [41]. On the other hand, several general Arabic corpora do include novels in their data (e.g., the International Arabic Corpus (ICA, available via: http://www.bibalex.org/ica/en/About.aspx, accessed on 12 April 2022), ArabiCorpus (available via: https://arabicorpus.byu.edu/search.php, accessed on 12 April 2022), and the corpus of King Abdulaziz City for Science and Technology (KACST, available via: https://corpus.kacst.edu.sa, accessed on 12 April 2022). However, a closer examination of such corpora reveals multiple limitations, summarized in the following points:

1. Most of the Arabic general corpora do not provide details about the regional information of the included novels. In other words, they do not explicitly indicate the geographical distributions of the data for each genre and subgenre represented. Thus, these corpora may or may not contain texts from Saudi novels since their users are not able to discern from which Arab countries the data originate. As a result, it is impossible to rely on these corpora to conduct a corpus study of Saudi novels. This limitation is true of the Corpus Linguae Arabicae [42], ICA [43], NEMLAR Written Corpus [44], and Jordon Comprehensive Contemporary Arabic Corpus [45].
2. Although the Arabic literature genre comprises part of the data of some general corpora, it is not clear what subgenres of Arabic literature are included in those corpora, i.e., we do not know whether Arabic or Saudi novels are included. An example of this issue is seen in the University of Jordon Arabic Corpus [46]. Although the developers of this corpus explicitly state that Arabic literature constitutes only 7% of the total of the corpus (around half a million words), no information is provided about the subgenres of Arabic literature. This drawback is also true for the NEMLAR Written Corpus [44], Corpus of Historical Arabic [34], and Corpus Linguistic Tools for Historical Semantics in Arabic [47].
3. Some corpora use texts' publication location as a principal criterion to regionalize the text. In other words, the texts are attributed to the country in which the publisher is located, not to that of the writer's nationality. For example, when a particular novel is authored by a Saudi writer but published by a company situated outside Saudi Arabia (e.g., Lebanon), the text is ascribed to that country rather than to Saudi Arabia. Such an application is problematic since it results in some of the data produced by Saudi novelists appearing erroneously as non-Saudi data to users. A clear example of this practice is found in the KACST corpus. The author [36] explicitly states that 'text location' is one of the main criteria for categorizing the texts. This criterion indicates

where the text was published, while the writer's nationality is entirely ignored. Hence, this corpus is not a reliable resource for a corpus-linguistic and stylistic study on Saudi novels. The same is true for the ICA [48].

4. A few corpora specify the names of authors whose novels are included. As a result, writers' nationalities are easily identifiable. However, a careful examination of the metadata of the novels reveals that the dataset representing novels by Saudi authors is rather small. For instance, ArabiCorpus, which contains 32 Arabic novels with 1,026,171 words, explicitly provides the title and author's name for each Arabic novel included in the corpus; hence, we can easily determine their regional distributions. However, a closer look at authors' names identifies that only one Saudi novel is considered, whereas most of the novels are written by Egyptian novelists. Therefore, this corpus is too limited to examine Saudi novels.
5. Our survey has identified two corpora of data derived from Saudi novels. However, the developers of both corpora targeted only web novels written in different Saudi dialects. The first corpus, *Gumar*, is a corpus of Gulf dialects [49]. Saudi novels comprise around 60% of the total data included in this corpus. The second corpus, *Rewayatech*: Saudi Web Novels Dataset [50], was created to represent novels written using Saudi dialects. The data source for both corpora is online forums. Hence, the included novels are not edited, printed, nor published by a company. In addition, the authors of the novels included in the two corpora remain anonymous. This kind of data is entirely different from the data we intend to collect and analyze. Our corpus targets data from printed novels written by known authors using mainly Modern Standard Arabic (MSA). Therefore, web novels that are written using different Saudi dialects do not form a source for our corpus because they are not normally considered by scholars of Arabic literature as legitimate and do not represent an elevated use of Arabic. Consequently, they are typically neglected in linguistic and stylistic studies, i.e., researchers and critics of Saudi literature do not typically target those novels. Therefore, we decided to focus on the novels mainly written using MSA and published in print. Furthermore, a technical reason for avoiding novels written using Saudi dialects is that most NLP Arabic tools, including Part of Speech (POS) taggers, have been modeled to perform perfectly with MSA [51]. Hence, handling and preprocessing the collected texts will be more easily manageable.

Considering these limitations, it is clear that the Arabic language lags behind in the development of an Arabic literary corpus in general and Saudi novels in particular. This is because the developers of Arabic corpora almost entirely overlook this genre. Thus, the current Arabic corpora (both general and specialized) cannot be exploited to address specific research questions related to Saudi novels' linguistic and stylistic features. This may explain why the use of corpus stylistics is absent from the study of Saudi novels. Hence, creating a specialized corpus for Saudi novels is highly demanded.

3. Design Criteria

The present project aims to create a specialized and tagged corpus that reflects the language used in Saudi novels. This corpus will be available for research purposes (e.g., linguistics, stylistic research). To achieve this goal, the following design criteria, inspired mainly by [36,52], were considered:

1. Scope
   The scope of this corpus encompasses Saudi novels, as specified in the title. This is the core distinguishing criterion, separating it from other existing Arabic corpora. Therefore, we considered as the corpus population only those novels written by Saudi writers, regardless of where they were born, where they permanently live/lived, or where the novel was published. Hence, when a particular novel is produced by a Saudi writer but published outside Saudi Arabia, it is regarded as compatible with this criterion. In order to confidently meet this criterion, we relied mainly on the informative bibliographical study conducted by [53]. This study indexes only Saudi novelists and their contributions. Finally, it should be noted that the language used in Saudi novels is mainly MSA, a relatively unified variety of Arabic that is used across Arab countries and represents the official usage [54]. The use of some Arabic dialects can occur, although it is not a common practice.
2. Time Span
   The intended plan is to cover the period from 1930, when the first Saudi novel was published, until 2019. We focused specifically on the date of a novel's first publication, even if it is republished at later dates as well. This information is acquired via various sources including bibliographies, studies on Saudi novels, and personal communications with novelists and experts in the field. The application of this criterion is valuable because it will enable corpus users to conduct diachronic analysis and, hence, observe any historical changes that might have occurred in the language of Saudi novels. Furthermore, we decided to divide the period from 1930 to 2019 into decades to make diachronic analysis more feasible.
3. Sampling
   The full text of each novel collected is included in the corpus. This approach has a significant advantage, as it increases the likelihood of detecting the most linguistic characteristics [36]. Had the corpus only included a subset of each text (e.g., 5000 words), it would have excluded a great deal of important features, especially when targeting a particular genre (see: [55]).
4. Gender
   The corpus includes both male and female writers. Taking such a criterion into account is valuable, since it allows researchers to conduct comparative analyses between the linguistic and stylistic features of the language used by each gender.
5. Size
   The initial size of the corpus we intended to achieve is, at minimum, around 2,000,000 words. We set this size for two reasons: (i) to align with the time and fund limit given by the funder to the project team to accomplish the task, and (ii) to account for the difficulty of gaining and preprocessing the texts in a short time. In other words, most of the novels are not available in a machine-readable format, which will most likely require a lengthy procedure and a large sum of money for the human resources needed to obtain, edit, and preprocess the corpus data. However, we believe that this size represents a good deal of Saudi novels' linguistic and stylistic features, bearing in mind that we aimed to include texts that varied in terms of writer name, time period, and gender. Moreover, the size of the corpus can always be increased in subsequent phases.

Finally, we would like to shed light on two intertwined notions related to corpora compilation. These are representativeness and balance. Representativeness refers to a status of a corpus when it contains a range of text types of the language or its variety. Hence, it usually includes different genres, periods, genders, ages, etc. Balance refers to maintaining the relative sizes of each type and subtype included in the corpus to achieve adequate representativeness of the language (i.e., applying a sampling frame) [3,29]. However, it should be borne in mind that there is no agreement among corpus linguists on the ideal application of these two notions during the process of corpus compilation [55]. In addition, it is almost impossible to create a fully representative corpus that accurately reflects the

language under consideration [36]. One exception to this is, for example, when corpora compilers target the works of a particular author, complete representativeness can then be achieved [55].

In our case, based on the above mentioned criteria, we can argue that the corpus achieves a reasonable extent of the data representativeness and balance since it is new in its scope (Saudi novel), covers a wide time span (90 years), contains writers from both genders, and includes the full text of each novel. The only factor that may limit the corpus representativeness and balance is the size of the corpus at present. However, as we mentioned above, the project's forthcoming phases will help overcome this limitation. More details on the data collection process and its limitations are given in Section 4.

4. Corpus Construction: Processes and Steps

After identifying the design criteria of the corpus, we move on to describe the practical procedures of the corpus construction. It is to be noted that any corpus has to be developed through rigorous, consecutive processes and steps. Hence, this section illustrates the phases of constructing the Saudi Novel Corpus. Figure 1 visualizes these steps.

Figure 1. Methodological steps of the corpus construction.

4.1. Data Collection

The data collection procedure is a critical issue corpus compilers must face, and it is typically the starting point of corpus construction. The achievement of this step usually depends on whether data is available in a format that allows a convenient and speedy process. Most of our collected data were not available on the web in machine-readable formats, so the data could not be captured by any crawling tool. Instead, the data were either available as image-only PDF copies or were not available electronically at all. Hence, we manually collected the related data from two sources: (i) free download web libraries or (ii) public/personal libraries or bookstores. If obtained from the second source, an additional step had to be taken to scan each page of the physical novel.

It must be noted that we applied an opportunistic approach to collect the data. Such an approach means that the corpus "make[s] no pretension to adhere to a rigorous sampling frame" [29]. We were forced to follow this data collection method because we faced a major obstacle from the initial stages of the project. We encountered problems gaining the texts and converting them to a machine-readable format, i.e., not all the novels we obtained were perfectly converted via the Optical Character Recognition (OCR) tool. This problem does not seem specific to our case, as it is reported to be a common challenge in the construction of novels' corpora in general [23]. Therefore, the texts included were subject to availability. In other words, we included only the texts that we easily gained and managed to acquire in the appropriate format for further computational processes. As a result, no subjective preference (i.e., a focus on particular writers) or fixed sampling frame was applied with regard to the selection of the included novels. This method was beneficial to our case, and it enabled us to manage our time within the project period given. We were able to collect as much of the relevant data as we could in a reasonable time frame. However, following the strategy applied by [56], we continued monitoring while collecting the data to catch any possible demographic and temporal gaps that could occur. Then, when they appeared, we attempted as much as possible to fill them by searching for related novels.

4.2. Optical Character Recognition (OCR)

We used an OCR tool to convert text images saved in PDF documents into a machine-readable format to facilitate text processing. The tool we used, FineReader, is commercial software that supports 198 languages including Arabic. It recognizes multi-fonts and different image resolutions and detects differences between documents (The tool is produced by the ABBYY Company and can bought be via the following link: https://pdf.abbyy.com/blog/focusing-on-pdf, accessed on 12 April 2022). More details about the specifications of this tool can be found via the following link: https://pdf.abbyy.com/media/2446/brochure-finereaderpdf-full-feature-list-en.pdf, accessed on 12 April 2022. Although this tool converted the texts to an adequate quality, the outputs were not free of errors. Therefore, we had to thoroughly review the entire converted text word by word, compare it with the original version, and correct the errors that we encountered. This step, although very time-consuming, was crucial, as it assured us that the data entered in the corpus was correct. The revision was conducted in two rounds: Round 1 was completed by experienced graduate students working part-time as copy editors for the Publishing Department at Islamic University of Madinah. Their responsibility in the department is to produce the final copy of the books in the required standard before printing. Round 2 was completed by the PI of the project, who checked the outputs of Round 1.

4.3. Data Cleaning

After digitizing the data and obtaining it in an editable and machine-readable format, a data cleaning task was performed both manually and automatically. In the manual stage, we removed the unwanted elements from each text. Specifically, we removed the following:

1. Images presented on the front and back covers;
2. Table of contents;
3. Edition notice containing copyright, publication information, printing history, edition date, cataloging information, and ISBN;
4. Dedication;
5. Preface;
6. Footnotes;
7. Authors' biography.

By removing these elements, we were confident that the data entered in each text file represented only the Saudi novel itself.

In the automatic cleaning, we used *Almoshatheb Alarabi* tool (The tool can be freely downloaded via the following link: https://sourceforge.net/projects/ghawwasv4/, accessed on 12 April 2022) to acquire the texts in the proper format. The following cleaning options available in the tool were chosen to unify the structure of the texts included in the corpus and to allow the POS tagger to perform efficiently:

1. Removing extra spaces;
2. Removing diacritics, because they are not used frequently in the novels;
3. Splitting numbers and symbols from words;
4. Removing Kasheeda (elongated letter); this is a stretch added to connected Arabic letters to lengthen them in various contexts, such as adding emphasis, providing legibility, and making text line justification [57];
5. Breaking the texts into sentences to allow a speed tagging process.

Figure 2 represents a sample view of the output of the automatic cleaning generated by *Almoshatheb Alarabi*.

(خيط أول) أحلام ... أعيتها رحلة البحث عن الحرية وسط تقاليد صارمة نصبت كتمثال الحرية منذ عشرات السنين ، كانت تحارب طواحين الهواء كما فعل دون كيشوت، حاربت الشمس ووقفت ضد شروقها حاولت إخراس أمواج البحر وأن يضل الليل طريقه إلى دروب المدينة ... أصمت أذنيها عن سماع تغريد الطيور ودوران الطبيعة من حولها، وجلها تنادي بإحناء الهامات لتهدأ العاصفة ... والعاصفة لا تهدأ أبدا بل تمور وتمور لتبعثر الآمال وتنثر السحب خيوطا في الرمال فتحلق الحرية بعيدا كطير يطير بنصف جناح. ما هي الحرية؟ أتساءل عن معنى تلك الكلمة الساحرة الرائعة الحارقة ... أنا المكبلة بالأغلال وقيود لا ترى وقضبان تحيطني من كل الجهات ...

(First series) Ahlam... She had become fatigued in her vigorous quest for freedom amid strict traditions upstanding like the statue of Liberty tens of years ago. She used to fight the windmills as Don Quixote did. She fought the sun and barred it from rising. she tried to vanquish roar in the waves of the sea, and the night to miscue its root to the paths of the city... The wave shut her ears tight from hearing the birds chirping and nature rotating around her. Furthermore, all of them call for the bow of the heads to calm the storm... Furthermore, the storm never calms down rather spreads and disseminates hopes, and clouds disperse threads in the sand, so freedom flies like a bird that flies with half a wing. What is freedom? I am marveled at the meaning of such gorgeous and magical word... I am handcuffed with shackle, constrained with invisible chains and surrounded from all angles...

Figure 2. Sample view of the output of the automatic cleaning generated by *Almoshatheb Alarabi* tool and its translation.

4.4. Word Annotation

Annotation is the process of inserting additional information into corpus data [3]. In the current corpus, we aimed at annotating each word with POS tags. This means that each word in the corpus is assigned to its most appropriate word class. This kind of annotation is hoped to provide researchers with additional grammatical information about the content, increase the usability of the corpus and improve the text analysis [58].

In this step, the words of the corpus were annotated with POS tags using the free Arabic Linguistics Pipeline (ALP) tool [59] (The tool can be accessed via the following link: http://arabicnlp.pro/alp/index-ar.php, accessed on 12 April 2022). This tool was chosen objectively based on the results of a study by [60] that compared five well-known Arabic POS taggers (Arabic Stanford, CAMeL, Farasa, MADAMIRA, and ALP) using a sample from the current corpus. ALP performed better along all three considered metrics (Recall, Precision, and F-score). In addition, the developers of this tool reported that it reached an accuracy of over 97% in testing. Moreover, one of the main advantages of ALP is that it performs three computational tasks in one single process: word segmentation, POS tagging, and NER. Thus, words do not need to be segmented first then tagged second [59]. Another distinctive feature of this tool, unlike other Arabic taggers, is that it contains 58 tags that label the words of the texts with more detailed grammatical features. For example, the tool contains thirteen tags for nouns, nine tags for adjectives, and five for verbs. Using ALP, the current corpus was provided with more detailed grammatical features, which we hope will enrich analyses conducted by linguistic and stylistic researchers. Figure 3 presents a sample of the tagged words in the corpus.

كان/PSTV النهر/D+SMN في/P هذا/DM المكان/D+SMN ينزلق/PRSV بهدوء/P+SMN
ليشكل/PRP+PRSV قوساً/SMN كبيراً/SMAJ يحدد/PRSV المعالم/D+PIN بوضوح/P+SMN
حاد/PX: SMAJ ففي/C+P ناحية/SFN تقوم/PX.PRSV داخل/SMN سور/SMN
مهدم/PX.SMAJ طاحونة/SFN قديمة/SFAJ متآكلة/SFAJ البناء/PX.D+SMN
بحجارتها/P+SFN+PRO الكبيرة/D+SFAJ التي/REL ربما/AV كانت/PSTV يوماً/SMN
حجارة/PIN معبد/SMN أو/C قلعة/PX.SFN أو/C تشبه/PRSV بيتاً/SMN كبيراً/SMAJ
مهجوراً/PX. SMAJ وتتوسط/C+PRSV السور/D+SMN مجموعة/SFN من/P أشجار/PIN
الجوز/PX.D+SMN تلقي/PRSV ظلالاً/PIN كثيفة/PIAJ على/P مساحات/PFN واسعة/PFAJ
PX./ تصل/PRSV قريباً/AV من/P الطريق/PX./D+SMN

Figure 3. Sample view of the tagged texts using the ALP tool.

4.5. File Naming and Saving

In this final step, the corpus texts were saved in plain text format in UTF-8 encoding. The corpus files were stored in two versions: tagged and untagged/raw. Each file was also given a unique ID for further processing.

4.6. Metadata

Metadata, a valuable feature of any corpus, refers to the overarching data that describes the data included in the corpus [2]. Metadata contains information about external features of the texts, including the author's name, gender, date of publication, type of genre, subgenre, etc. This information is useful as it "aids the investigation of the data in the corpus" [29]. In other words, metadata allows corpus users to conduct their studies using a range of different aspects of the texts and to address several related questions (ibid.). Without this high-level information, the texts would be disconnected from their external context [61]. Thus, the usability of the corpus would be limited if metadata were not included.

The metadata assigned to each text in the current corpus is as follows: title, writer's name and gender, number of words and types, publisher, year of the first appearance of the novel, and edition publication year. The final two pieces of metadata are different from one another: the former refers to the year the novel was first published, while the latter refers to the year of the edition from which we took the data. Sometimes, the first appearance year is the same as the edition year. Considering the former is important because it represents the language usage of that year and enables users to investigate the novels from a diachronic perspective.

The corpus metadata are kept in a separate database in the form of an Excel document. This method of data storage is preferable, as it facilitates comparative analyses between different types of texts based on the metadata information (See: [61]).

5. The Issue of Copyright

One of the obstacles that corpus compilers commonly encounter is the issue of copyright related to the texts to be included in the corpus. In many cases, a crucial step is to obtain permission from publishers, allowing corpus users to download the full content of the texts, preview the texts without downloading them, and/or allow the use of the corpus for commercial purposes. For the time being, we applied the following measures, which align with those applied by the compiler of the KACST corpus [36]:

1. Texts will not be distributed.
2. Texts will not be available for download.
3. Texts will not be available for preview. One exception will be applied here: when researching via the web interface we intend to program, corpus users will be able to see a brief snippet of text related to their target search word.

4. Bibliographical information related to each text will be provided to protect the copyrights.
5. The corpus will not be used for commercial purposes.

These measures bring the production of the current corpus in line with intellectual property regulations in Saudi Arabia (See: https://tinyurl.com/yckxs5rt, accessed on 12 April 2022).

6. The Corpus Statistics and Preliminary Results

6.1. General Statistics

In this subsection, we present some general statistics that emerge from analyzing the content of the corpus.

Table 1 shows the basic statistics of the corpus. The first point that should be made here is that the corpus size has exceeded the original target number we set before proceeding with the work. Prior to constructing the corpus, we, as stated in Section 3, initially set 2,000,000 words as a target size. Fortunately, we managed to exceed this number and reached just over 3,000,000 words.

Table 1. The basic statistics of the corpus.

Item	Total Number
No. of novels	53
No. of words	3,134,074
No. of types (unique words)	214,736
No. of writers	24 (15 male, 9 female)
Covered period	1930–2019

Table 2 presents the distributions of the number of novel texts and words over the standardized period (decades) and gender. The table shows that the data distributions of the novels across the periods specified are not even. It reveals that most of the data come from the novels published in the last two decades. This represents 72% of the total number. This imbalanced distribution is caused by the problem of the data availability issue that we encountered during the collection process, while finding the novels published in recent times was quite attenable, those published in earlier periods were difficult to obtain as many of them are most likely out of stock and not republished or made available electronically. Moreover, the last two decades have witnessed a dramatic increase in the authorship movement in Saudi novels compared to the preceding decades, as reported by [19]. Similar reasons can be linked to the imbalanced distribution concerning the gender of the writers, where the novels written by males outweigh those written by females, as presented in Table 2.

Another feature of the corpus data is that the 53 novels were written by different 24 Saudi writers. However, those writers do not have equal numbers included in the corpus. For example, Abdulrraḥmān Munīf and ‘Abdu Khāl have six novels each, Ghāzī Alquṣaybī have five novels, while laylā Aljuhanī and Sa’ad Alghuraybī have one novel each. This imbalance between the writers’ novels has also resulted from the attainability issue. i.e., we could not obtain every single novel written by each Saudi writer. This inequality might be less caused by the fact that some writers have produced less novels than the others (one or two).

Table 2. Detailed statistics on the content of the corpus.

Period	No. of Texts		No. of Words	Size Percentage
	Male	Female		
1930–1939	1	0	12,077	0.4
1940–1949	0	0	0	0
1950–1959	1	0	59,039	1.9
1960–1969	0	0	0	0
1970–1979	4	0	146,259	4.6
1980–1989	2	1	101,066	3.2
1990–1999	4	2	555,954	17.8
2000–2009	13	7	1,135,247	36.2
2010–2019	13	5	1,124,432	35.9
Total	38	15	3,134,074	100
	53			

6.2. Experimental Results

In this subsection, we performed further linguistic analysis to explore more internal features of the corpus and reported the results. We mainly focused on investigating the frequencies of the words and the POS tags and classes. Table 3 lists the frequencies of n-grams at three levels:

- Unigram (one word);
- Bigram (sequence of two words);
- Trigram (sequence of three words).

Table 3. The top ten uni-gram, bigram, and trigram in the Saudi Novel Corpus.

Rank	Unigram	Trans.	Bigram	Trans.	TriGram	Trans.
1.	<في>	in	<بعد أن>	after that	<في تلك اللحظة>	at that moment
2.	<من>	from	<قبل أن>	before that	<لا بد من>	must
3.	<أن>	that	<كل شيء>	every thing	<من غير أن>	other than that
4.	<على>	on	<لم يكن>	it was not	<لا يمكن أن>	cannot be that
5.	<لا>	not	<في هذه>	in this	<في كل مكان>	in every place
6.	<إلى>	to	<من قبل>	from before	<على الرغم من>	in spite of
7.	<ما>	not	<في هذا>	in this	<من دون أن>	without
8.	<لم>	did not	<لا بد>	must	<لا أريد أن>	I do not want
9.	<كان>	was	<إلا أن>	except that	<في اليوم التالي>	next day
10.	<التي>	who (Feminine relative pronoun)	<أريد أن>	I want that	<أكثر من مرة>	more than ones

We calculated the results after removing all punctuation from the texts. The table displays that the most frequent ten unigrams are solely function words, such as the prepositions في (in) and من (from). Although the top ten bigrams and trigrams are mostly function words, they also include among them a few content words, such as أريد (want), اللحظة (moment), and اليوم التالي (next day). This tendency indicates the dominance of function

words over content words in the corpus. The first most frequent content word is ranked in position 51, which is the word الله (The Arabic term that refers to God in the religion of Islam), followed by the word شيء (thing), ranked in position 56.

To further analyze these results, we compared them with the results merged from the analysis of the Arabic Newspapers Corpus 2012, which its data was collected from different Arabic news websites and contains around 2,500,000 [62]. Table 4 shows the results according to the same three levels as in Table 3.

Table 4. The top ten unigram, bigram, and trigram in the Arabic Newspapers Corpus 2012.

Rank	Unigram	Trans.	Bi–Gram	Trans.	Tri–Gram	Trans.
1.	<في>	in	<من خلال>	through	<صلى الله عليه>	May God bless him
2.	<من>	from	<إلى أن>	until	<الله عليه وسلم>	God bless him
3.	<على>	on	<الله عليه>	God be upon him	<الله صلى الله>	God bless God
4.	<أن>	that	<في هذا>	in this	<رسول الله صلى>	The Prophet, may God bless
5.	<إلى>	to	<صلى الله>	God bless	<النبي صلى الله>	The Prophet, may God bless
6.	<التي>	who	<أكثر من>	more than	<على الرغم من>	in spite of
7.	<عن>	about	<من أجل>	in order to	<رضي الله عنه>	May Allah be pleased with him
8.	<ما>	did not	<في هذه>	in it	<في الوقت الذي>	at the time that
9.	<الذي>	who	<من قبل>	from before	<مشيرا إلى أن>	indicating that
10.	<لا>	not	<عليه وسلم>	peace be upon him	<الله عليه وسلم>	God bless him

From Tables 3 and 4 above we notice that there is almost identicality between the Saudi Novel Corpus and the Arabic Newspaper corpus in the unigram case, especially in the top five cases. However, the Saudi Novel Corpus shows uniqueness in the use of the verb كان (was). This practice may have arisen as a result of the genre style that novel typically has, as it is common for the novel's writers or the characters to narrate past events.

The divergence from identicality shows more in the bigram since the text genre starts to show effects. Only three incidences where there were similarities of bigrams though for different ranks. As an example, the bigram في هذا (in this) is repeated in the two cases but as number seven in the Saudi Novel Corpus and number four in the Arabic Newspapers Corpus 2012.

On the other hand, the similarity is significantly reduced to only one occurrence for the trigram. We notice that the trigram على الرغم من, meaning 'in spite of,' is the only similarity in this case. Furthermore, the prepositional adverb في تلك اللحظة (at that moment), which identifies a point of time, is the most commonly used trigrams in the Saudi Novels Corpus. This tendency indicates a genre effect in the data presented. Similarly, the genre of the text shows very much in the trigram list of the Arabic Newspaper Corpus since the religious articles constitute a portion of the corpus where the phrase صلى الله عليه وسلم (Peace be upon him) occurs so often. Such a comparison may shed light on some of the distinguishing linguistic features of the Saudi Novel Corpus. However, more comparisons with other Arabic genres are essential to confirm such distinctiveness.

As for the frequency of the tags, Table 5 shows the top ten most frequent tags in the corpus. It demonstrates that the tag SMN, which stands for a singular masculine noun, is at the top of the list, followed by the tag P, which refers to a preposition, while the tag SMAJ, which refers to a singular masculine adjective, comes at the bottom of the list.

Table 5. The top ten tags used in the corpus.

Rank	Tag	Meaning
1.	SMN	Singular masculine noun
2.	P	Prepositions
3.	PRO	Pronouns
4.	D	Definite article
5.	PRSV	Present verb (active)
6.	C	Conjunctions
7.	PSTV	Past verb (active)
8.	SFN	Singular feminine noun
9.	PIN	Plural irregular noun
10.	SMAJ	Singular masculine adjective

To compare the frequencies of the POS categories used in the corpus, we decided to reduce all the 58 tags provided by the tagger ALP into the eight main POS categories: Noun, Verb, Preposition, Adjective, Adverb, Particle, Article, and Conjunction. This procedure was conducted by mapping each tag to one of these categories. This mapping method was only used for comparison purposes since it enabled us to explore how much the main POS categories were utilized in the corpus and to undertake the comparison efficiently. We assessed the exploitation of the POS categories in the corpus at three levels: uni-tag (individual tag /individual POS category), bi-tag (the sequence of two tags), and tri-tag (the sequence of three tags). Table 6 presents the outcomes that emerged from the comparison.

Table 6 illustrates that the most frequent uni-tags (i.e., POS classes) in the corpus are Noun, followed by Verb and Preposition, respectively. Adverb remains at the bottom. It also shows that the most frequent bi-tags are 'Noun, Noun,' followed by 'Article, Noun.' The sequences 'Noun, Verb' and 'Noun, Conjunction' are the least frequent bi-tags among the top ten. The table also suggests that 'Noun, Article, Noun' and 'Preposition, Noun, Noun' are the most frequent POS sequences for the tri-tag. These results collectively indicate that the POS category 'Noun' is dominantly used in the corpus compared to other categories. Finally, the significance of these results would be admitted when compared with a corpus from a different genre, provided the ALP tool is used to tag the words of that corpus.

Table 6. The most frequent uni-tags, bi-tags and tri-tags.

Rank	Uni-Tag	Bi-Tag	Tri-Tag
1.	Noun	Noun, Noun	Noun, Article, Noun
2.	Verb	Article, Noun	Preposition, Noun, Noun
3.	Preposition	Preposition, Noun	Noun, Preposition, Noun
4.	Article	Verb, Noun	Verb, Preposition, Noun
5.	Particle	Noun, Article	Noun, Noun, Noun
6.	Conjunction	Noun, Preposition	Preposition, Article, Noun
7.	Adjective	Particle, Verb	Verb, Noun, Noun
8.	Adverb	Verb, Preposition	Noun, Article, Adjective
9.	——	Noun, Verb	Noun, Noun, Preposition
10.	——	Noun, Conjunction	Particle, Verb, Noun

7. Conclusions and Future Work

Although there has been a growing interest in the construction of Arabic corpora, there is still more demand for constructing specialized corpora that contain data collected from Arabic literary texts. In this paper, we introduced the Saudi Novels Corpus. Currently, the corpus contains around 3,000,000 tagged words collected from 53 novels that are written by different writers (males and females) and cover a span of time periods (from 1930 to 2019).

In addition, we automatically annotated each word in the corpus with POS tags to advance the research in its content. We finally reported some preliminary results that have revealed the diachronic distributions and various linguistic features of the content of the corpus.

The results of our work showed an increase in the number of novels in the time period of the study. A significant increase in the number of novels has been observed in the last two decades. The results also showed an increase in the contents (number of words) as well as the number of authors from both genders. We also noticed an imbalance in the number of novels written by each listed writer. Some writers have many novels in the ensemble (up to six novels), while others have one or two novels.

The linguistic analysis of our corpus and a comparison with that of the Arabic News Corpus 2012 showed that the unigram in both corpora is almost identical. However, in the case of diagram and trigram, there have been significant variations and few similarities. The reduction in the number of similar bigrams and trigrams as compared to the unigram is expected as the text genre started to show its effect in the last two cases.

We are positive that the contributions of this work help advance the study of corpus stylistics in Arabic, facilitate the analysis of Saudi fictional works, and pave the way to addressing several questions related to the language used in Saudi novels. For example, when made available, the corpus can be useful for exploring the linguistic and stylistic features of Saudi novels through the application of different corpus linguistic techniques, such as identifying the keywords and lexical bundles and investigating concordance lines.

We consider the works reported in this paper as preliminary steps toward bridging the existing gap between corpus linguistics and the study of Arabic literary texts. Hence, more tasks will be undertaken to improve the quality of the corpus with some advanced features. We specifically aim to expand the size of the corpus to include additional relevant data, which is currently under process. Meanwhile, we will work on filling the gaps in the data distribution. Moreover, we plan to add more features to the data, such as word lemmatization and direct speech annotation. Finally, we intend to build a web search engine that provides some functions to facilitate research queries in the corpus.

Author Contributions: Conceptualization, M.A.R.A. and A.A.-T.; Formal analysis, R.A.; Funding acquisition, T.A.; Investigation, T.A. and A.Y.; Methodology, T.A. and R.A.; Project administration, T.A.; Resources, M.A.R.A. and A.Y.; Software, R.A.; Writing—original draft, T.A.; Writing—review & editing, M.A.R.A. and A.A.-T. All authors have read and agreed to the published version of the manuscript.

Funding: The research was funded by the Deanship of Scientific Research at Islamic university of Madinah. Grant number (581).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors wish to thank Abed Alhakim Freihat (University of Trent, Italy) for his scientific support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kennedy, G. *An Introduction to Corpus Linguistics*; Routledge: Abingdon, UK, 1998.
2. Kübler, S.; Zinsmeister, H. *Corpus Linguistics and Linguistically Annotated Corpora*; Bloomsbury Publishing: London, UK, 2015.
3. Baker, P. *Glossary of Corpus Linguistics*; Edinburgh University Press: Edinburgh, UK, 2006.
4. Biber, D. Corpus linguistics and the study of literature: Back to the future? *Sci. Study Lit.* **2011**, *1*, 15–23.
5. Mahlberg, M. Corpus stylistics: bridging the gap between linguistic and literary studies. *Text Discourse Corpora Theory Anal.* **2007**, *8*, 219–246.
6. Baker, P. *Sociolinguistics and Corpus Linguistics*; Edinburgh University Press: Edinburgh, UK, 2010.
7. O'Sullivan, J. *Corpus Linguistics and the Analysis of Sociolinguistic Change: Language Variety and Ideology in Advertising*; Routledge: Abingdon, UK, 2019.
8. Ancarno, C. Corpus-assisted discourse studies. In *The Cambridge Handbook of Discourse Studies*; Cambridge University Press: Cambridge, UK, 2020.
9. Mikhailov, M.; Cooper, R. *Corpus Linguistics for Translation and Contrastive Studies: A Guide for Research*; Routledge: Abingdon, UK, 2016.
10. Mahlberg, M.; Wiegand, V.; Stockwell, P.; Hennessey, A. Speech-bundles in the 19th-century English novel. *Lang. Lit.* **2019**, *28*, 326–353. [CrossRef]
11. Wright, D. Corpus approaches to forensic linguistics. In *The Routledge Handbook of Forensic Linguistics*; Coulthard, M., May, A., Sousa-Silva, R., Eds.; Routledge: London, UK, 2021; pp. 611–627.
12. Jones, M.; Durrant, P. What can a corpus tell us about vocabulary teaching materials. In *The Routledge Handbook of Corpus Linguistics*; Routledge: Abingdon, UK, 2010; pp. 387–400.
13. Al-Laith, A.; Shahbaz, M.; Alaskar, H.F.; Rehmat, A. Arasencorpus: A semi-supervised approach for sentiment annotation of a large arabic text corpus. *Appl. Sci.* **2021**, *11*, 2434. [CrossRef]
14. Wijitsopon, R. A corpus-based study of the style in Jane Austen's novels. *Manusya J. Humanit.* **2013**, *16*, 41–64. [CrossRef]
15. Mahlberg, M.; Biber, D.; Reppen, R. Literary style and literary texts. In *The Cambridge Handbook of English Corpus Linguistics*; Cambridge University Press: Cambridge, UK, 2015; pp. 346–361.
16. Stubbs, M. Conrad in the computer: Examples of quantitative stylistic methods. *Lang. Lit.* **2005**, *14*, 5–24. [CrossRef]
17. Fischer-Starcke, B. *Corpus Linguistics in Literary Analysis: Jane Austen and Her Contemporaries*; Bloomsbury Publishing: London, UK, 2010.
18. Mahlberg, M.; Stockwell, P.; Joode, J.d.; Smith, C.; O'Donnell, M.B. CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics. *Corpora* **2016**, *11*, 433–463. [CrossRef]
19. Al-Yūsuf, K. Ḥarakat al-Ta'līf wa-al-Nashr al-Adabī fī al-Mamlakah al-'Arabīyah al-Sa'ūdīyah khilāl 'ishrīn 'āman 2000–2020. 2021, *accept*.
20. Al-Ḥaydarī, A. Dalīl al-rasā'il al-Jāmi'īyah fī al-adab wa al-naqd fī al-Mamlakah al-'Arabīyah al-Sa'ūdīyah: 1966–2021- Taḥlyl wa byblywqrāfyā. 2022, *accept*.
21. Wynne, M. Stylistics: Corpus approaches. In *Encyclopedia of Language & Linguistics;* Elsevier: Amsterdam, The Netherlands, 2006; pp. 223–226.
22. Mahlberg, M. Corpus linguistics and the study of nineteenth-century fiction. *J. Vic. Cult.* **2010**, *15*, 292–298. [CrossRef]
23. Maiwald, P. Exploring a Corpus of George MacDonald's Fiction. *North Wind. J. Georg. Macdonald Stud.* **2011**, *30*, 5.
24. Green, C. Introducing the Corpus of the Canon of Western Literature: A corpus for culturomics and stylistics. *Lang. Lit.* **2017**, *26*, 282–299. [CrossRef]
25. Bornet, C.; Kaplan, F. A simple set of rules for characters and place recognition in French novels. *Front. Digit. Humanit.* **2017**, *4*, 6. [CrossRef]
26. Nais, L. "A style which defies convention, tradition, homogeneity, prudence, and sometimes even syntax": Henry James's The Portrait of a Lady and Edith Wharton's The Age of Innocence. *Int. J. Lit. Linguist.* **2020**, *9*, 25. [CrossRef]
27. Mostafa, M.M.; Nebot, N.R. A Corpus-based Computational Stylometric Analysis of the Word "Árabe" in Three Spanish Generación Del 98 Writers. *J. Lang. Teach. Res.* **2018**, *9*, 928–938. [CrossRef]
28. Kubis, M. Quantitative analysis of character networks in Polish 19th-and 20th-century novels. *Digit. Scholarsh. Humanit.* **2021**, *36*, ii175–ii181. [CrossRef]
29. McEnery, T.; Hardie, A. *Corpus Linguistics: Method, Theory and Practice*; Cambridge University Press: Cambridge, UK, 2011.
30. Borin, L.; Forsberg, M.; Roxendal, J. Korp—the corpus infrastructure of Språkbanken. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 21–27 May 2012; pp. 474–478.
31. Gemeinböck, I. Containing chaos: compiling a corpus of eighteenth century prose fiction. In Proceedings of the Annual Conference of the Poetics and Linguistics Association (PALA), Online, 7–12 August 2016; Volume 29.
32. Bartis, I. FinnishRussian/Russian-Finnish Parallel Corpus of Literary Texts. Kielipankki. 2017; Volume 4. Available online: http://urn.fi/urn:nbn:fi:lb-20140730173 (accessed on 12 April 2022).
33. Erjavec, T. MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Lang. Resour. Eval.* **2012**, *46*, 131–142. [CrossRef]
34. Belinkov, Y.; Magidow, A.; Barrón-Cedeño, A.; Shmidman, A.; Romanov, M. Studying the history of the Arabic language: Language technology and a large-scale historical corpus. *arXiv* **2018**, arXiv:1809.03891.

35. Al-Sulaiti, L.; Atwell, E.S. The design of a corpus of contemporary Arabic. *Int. J. Corpus Linguist.* **2006**, *11*, 135–171. [CrossRef]
36. Al-Thubaity, A.O. A 700M+ Arabic corpus: KACST Arabic corpus design and construction. *Lang. Resour. Eval.* **2015**, *49*, 721–751. [CrossRef]
37. El-Khair, I.A. Abu el-khair corpus: A modern standard arabic corpus. *Int. J. Recent Trends Eng. Res.* **2016**, *2*, 11.
38. Saad, M.K.; Ashour, W.M. Osac: Open source arabic corpora. In Proceedings of the 6th ArchEng International Symposiums (EEECS), Opatija, Croatia, 12–15 December 2010; Volume 10.
39. El-Haj, M.; Koulali, R. KALIMAT a multipurpose Arabic Corpus. In Proceedings of the Second Workshop on Arabic Corpus Linguistics (WACL-2), Lancaster, UK, 22 January 2013; pp. 22–25.
40. Arts, T.; Belinkov, Y.; Habash, N.; Kilgarriff, A.; Suchomel, V. arTenTen: Arabic corpus and word sketches. *J. King Saud-Univ.-Comput. Inf. Sci.* **2014**, *26*, 357–371. [CrossRef]
41. Khorsheed, M.S.; Al-Thubaity, A.O. Comparative evaluation of text classification techniques using a large diverse Arabic dataset. *Lang. Resour. Eval.* **2013**, *47*, 513–538. [CrossRef]
42. Zemánek, P. CLARA (Corpus Linguae Arabicae): An Overview. In Proceedings of the ACL/EACL Workshop on Arabic Language, Toulouse, France, 6 July 2001.
43. Alansary, S.; Nagi, M.; Adly, N. Building an International Corpus of Arabic (ICA): Progress of compilation stage. In Proceedings of the 7th International Conference on Language Engineering, Cairo, Egypt, 5–6 December 2007; pp. 5–6.
44. Yaseen, M.; Attia, M.; Maegaard, B.; Choukri, K.; Paulsson, N.; Haamid, S.; Krauwer, S.; Bendahman, C.; Fersøe, H.; Rashwan, M.; et al. Building annotated written and spoken Arabic LRs in NEMLAR project. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, 22–28 May 2006.
45. Sawalha, M.; Alshargi, F.; Alshdaifat, A.; Yagi, S.; Qudah, M.A. Construction and annotation of the Jordan comprehensive contemporary Arabic corpus (JCCA). In Proceedings of the Fourth Arabic Natural Language Processing Workshop, Florence, Italy, 1 August 2019; pp. 148–157.
46. Hammo, B.; Al-Shargi, F.; Yagi, S.; Obeid, N. Developing tools for Arabic corpus for researchers. In Proceedings of the Second Workshop on Arabic corpus Linguistics (WACL-2), Lancaster, UK, 22 January 2013.
47. Ismail, O.; Yagi, S.; Hammo, B. Corpus Linguistic Tools for Historical Semantics in Arabic. *Int. J.-Arab.-Engl. Stud. (IJAES)* **2014**, *15*, 135–152.
48. Alansary, S.; Nagi, M. The international corpus of Arabic: Compilation, analysis and evaluation. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), Doha, Qatar, 25 October 2014; pp. 8–17.
49. Khalifa, S.; Habash, N.; Abdulrahim, D.; Hassan, S. A large scale corpus of Gulf Arabic. *arXiv* **2016**, arXiv:1609.02960.
50. Addawood, A.; Alzeer, D. Rewayatech: Saudi Web Novels Dataset. *Preprints* **2020**, 2020080628. [CrossRef]
51. Alkhazi, I.S.; Teahan, W.J. BAAC: Bangor Arabic Annotated Corpus. *Mach. Transl.* **2018**, *22*, 23. [CrossRef]
52. Sinclair, J. *Corpus, Concordance, Collocation (3. Impr)*; Oxford University Press: New York, NY, USA, 1995.
53. Al-Yūsuf, K. Mu'jam al-ibdā' al-Adabī fī al-Mamlakah al-'Arabīyah al-Sa'ūdīyah—al-riwāyah, madkhal tārīkhī, dirāsah bibliyūjrāfīyah bibliyūmitrīyah. 2010, *accept*.
54. Al Suwaiyan, L.A. Diglossia in the Arabic language. *Int. J. Lang. Linguist.* **2018**, *5*, 228–238. [CrossRef]
55. Nelson, M. Building a written corpus. In *The Routledge Handbook of Corpus Linguistics*; Routledge: Abingdon, UK, 2010; pp. 53–65.
56. Love, R.; Dembry, C.; Hardie, A.; Brezina, V.; McEnery, T. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *Int. J. Corpus Linguist.* **2017**, *22*, 319–344. [CrossRef]
57. Benatia, M.J.E.; Elyaakoubi, M.; Lazrek, A. Arabic text justification. In Proceedings of the TUG 2006 Conference, Marrakesh, Morocco, 25–28 September 2006; Volume 27, pp. 137–146.
58. McEnery, T.; Wilson, A. *Corpus Linguistics*; Edinburgh University Press: Edinburgh, UK, 2008.
59. Freihat, A.A.; Bella, G.; Mubarak, H.; Giunchiglia, F. A single-model approach for Arabic segmentation, POS tagging, and named entity recognition. In Proceedings of the 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), Algiers, Algeria, 25–26 Apirl 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–8.
60. Alluhaibi, R.; Alfraidi, T.; Abdeen, M.A.; Yatimi, A. A Comparative Study of Arabic Part of Speech Taggers Using Literary Text Samples from Saudi Novels. *Information* **2021**, *12*, 523. [CrossRef]
61. Adolphs, S.; Knight, D. Building a spoken corpus. In *The Routledge Handbook of Corpus Linguistics*; Routledge: Abingdon, UK, 2010, pp. 38–52.
62. Al-Thubaity, A.; Khan, M.; Al-Mazrua, M.; Al-Mousa, M. New language resources for arabic: corpus containing more than two million words and a corpus processing tool. In Proceedings of the 2013 International Conference on Asian Language Processing, Nagoya, Japan, 14–19 October 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 67–70.