Hindawi
Mobile Information Systems
Volume 2022, Article ID 3105790, 8 pages
https://doi.org/10.1155/2022/3105790

Research Article

Online Corpus Construction of English Text Collection, Data Cleaning, and Similarity Analysis

Huanyu Wang

Tangshan Normal University, Tangshan 063000, China

Correspondence should be addressed to Huanyu Wang; wanghuanyu@tstc.edu.cn

Received 7 July 2022; Revised 10 August 2022; Accepted 1 September 2022; Published 15 September 2022

Academic Editor: Yajuan Tang

Corpora are applied to analyze and study the characteristics of the target language. In language education, corpora are playing an increasingly essential role due to their large capacity, authenticity, rapid and accurate retrieval, as well as quick and easy statistics. At present, a great number of universities are trying to apply the textbook corpus to English teaching. However, most of the existing corpora face the issue of poor sharing. In addition, these corpora may be limited to a specific textbook, which leads to the lack of wide coverage of the retrieval and analysis results. As a result, it is quite necessary to develop a set of English corpora that is highly relevant, well shared, and easy to use by fully integrating existing teaching resources according to the characteristics of English subjects in universities. In recent years, the use of corpus-assisted English language teaching has gained widespread attention and exploration as computers have become more and more popular. After all, a corpus-based teaching model can effectively eliminate the various drawbacks of traditional vocabulary teaching. In fact, the corpus has a large amount of authentic corpus. The authenticity and practicality of the corpus facilitate students' mastery and use of English vocabulary in real contexts. What is more, the new model of corpus-assisted English vocabulary teaching can greatly increase independent learning and cooperative activities, so that students can increase their internal motivation for learning. This study begins with a brief introduction to the concept and characteristics of corpora. To be specific, the advantages of the corpus application in foreign language teaching are explained. At the same time, this research further analyzes the shortcomings of the existing corpus in university English education from the perspective of the current development and application of English corpora as well as clarifies the importance of building a corpus of university English teaching materials. After that, the system's operating environment and main development techniques are determined according to the specific requirements of the corpus for university English textbooks. In other words, the overall design and detailed design of the corpus and its management system were then carried out on the basis of the chosen technology platform. In addition, the structure of the tables in the database is analyzed and the basic components and operation procedures of the system are introduced. Furthermore, the functional modules of the system are designed. At the same time, the automatic word and sentence separation methods of the original corpus, the corpus entry process, the cross-distance search of the corpus, and the statistical analysis of the search results are discussed in detail. In conclusion, this study is based on English text collection and data cleaning techniques to build an online corpus.

1. Introduction

In modern society, English has always been a crucial tool of expression for international political and economic communication [1, 2]. For a long time to come, English will continue to dominate the world's languages. In this context, the study of English, especially the semantic and feature analysis of English literature, can provide essential information to support the learning and research of non-native English speakers [3]. However, there is a paucity of computer-supported software for phraseological characterization of linguistic corpora in China. As a result, the quality of information-based linguistic research results is poor [4]. In other words, prominent English linguists in China have to use foreign linguistic software to conduct their research and analysis. These programs are cumbersome to use, require payment, and require special access [5]. In addition, there are also many problems with incompatible formats as well as

poor integration between different related software. Thus, these issues greatly increase the cost of research, prolong the research cycle, and significantly reduce the efficiency of research [6]. At the same time, vocabulary is the basic unit of language and an important component of language knowledge. What is more, it is the basis for the acquisition of language in the areas of listening, speaking, reading, and writing [7]. Therefore, vocabulary is the carrier of semantic expressions and can play a role in transmitting the information. As a result, vocabulary acquisition is the primary threshold for learning and mastering a language [8]. In recent years, vocabulary teaching strategies in foreign language teaching have developed considerably as national education authorities and educators have paid more attention to vocabulary teaching and research [9].

Information technology is quite a useful and intelligent technology for today's society and a necessary tool for future development [10, 11]. At the same time, information technology is an important tool for the sustainable development of enterprises or research institutions in universities. Compared to traditional methods, the involvement of information technology can greatly facilitate the research and construction of various disciplines [12]. At the same time, due to the innovation of research methods, it is possible to improve the research tools and methods and make them more precise [13]. As a result, research on information technology can significantly optimize people's lives. The automatic collection and modular analysis of information in the course of the research and the efficiency and accuracy of data processing are unmatched by traditional research methods [14]. A corpus is a collection of texts, either written or spoken, usually stored in a computer database [15]. Written corpora are usually derived from books, newspapers, and journals. Therefore, they can be accessed electronically by scanning or downloading them directly [16]. Spoken corpora are derived from real-life uses of language, such as telephone conversations, business meetings, and television programs [17]. In summary, by studying a corpus of a particular language or group of languages, one can identify important features of the target language. This is of great importance for analyzing and studying language usage and patterns, writing dictionaries and lexicons, and teaching language [18]. With the development of computer technology, the integration of the corpus and computer technology allows the corpus to grow in size. In addition, this integration can make the application more efficient and promote the corpus to be more widely used [19]. A corpus is a study of language use within the objective world itself. It uses a set of quantitative methods and computer technology to study real language usage data. As a result, the development of corpora is a great contribution to the research of natural language processing as well as applied linguistics [20].

Nowadays, the corpus is developing rapidly. Different kinds of corpora have been created to meet different needs, so the classification of corpora is complicated [21]. From a professional point of view, corpora can be divided into the general corpus and specialized corpora. The general corpus is characterized by its large capacity, a wide range of topics, and high usefulness [22]. Commonly used general-purpose corpora include the American National Corpus and the American Contemporary English Corpus [23]. Specialized corpora are reflections of a particular area of language. For example, if you are studying business negotiations and business agreements in English, you will not be able to find a general corpus that is relatively specialized. Instead, a specialized corpus of business English should be used. In addition, the corpus can be divided into a parallel corpus and an analogous corpus, as shown in Figure 1. A parallel corpus is a bilingual or multilingual corpus composed of the original text and its corresponding translated text [24]. To be specific, a parallel corpus can be divided into a one-way parallel corpus, a two-way parallel corpus, and a multiway parallel corpus according to the direction of translation. An analogical corpus refers to two or more corpora composed of different forms or variants of the same language [25]. As a result, analogical corpora can be divided into monolingual analogical corpora and multilingual analogical corpora.

One of the main advantages of the corpus, when combined with computer technology, is its large volume of information. In addition, the online corpus allows for quick searches and statistics on relevant grammatical features as needed [26]. For foreign language teaching, corpora also provide a rich source of materials. In the early days, corpora were an important source for printed texts, such as word dictionaries, grammar books, and textbooks. Nowadays, almost all grammar books and word dictionaries that are widely used in foreign language teaching use corpora in the process of writing [27]. The application of multimedia information technology in teaching and learning, as well as the rapid development of computer technology and network technology, has led to an increasing use of corpora in teaching. Also, the use of corpora in classroom teaching mainly includes the learning of word meanings and grammar, the understanding of common word combinations, the differentiation of synonyms and near-synonyms, and the analysis of the language environment in which the words are adopted [28]. The university English textbook corpus is a teaching information resource formed by combining a large amount of English textbook content, organized in a certain way, with relevant search and analysis software. This kind of corpus can organize various kinds of materials according to the characteristics of human English teaching in order to meet the needs of teachers and students for university English teaching information [29]. This corpus not only facilitates teachers' teaching and lesson preparation but also promotes students' learning initiatives. However, most of the existing textbook corpora are stand-alone. In other words, they only provide electronic text files of relevant textbooks. As a result, users need to copy these files to their own computers and use some common search software to retrieve and analyze them. The disadvantage of this approach is that it is poorly shared and cannot be used anywhere and anytime. Although some corpora are also available online, they are usually only available for a specific textbook, which leads to a lack of rich linguistic information resources and a lack of extensive coverage of search and analysis results.

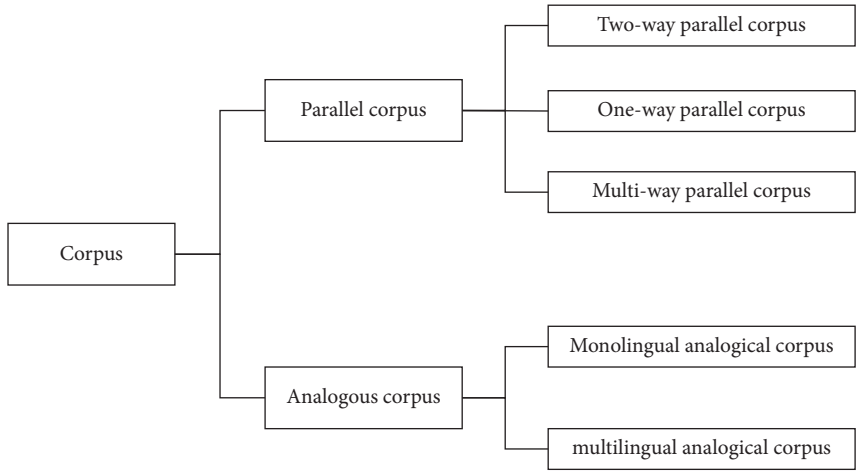

Figure 1: A parallel corpus and an analogous corpus.

The linguistic corpus has been developed for decades and has made great strides in the construction of the English corpus and the study of common features of language [30]. This is where conventional language teaching and research can be anchored, and thus the fundamentals of corpus linguistics can be secured. Also, there is a great demand for specialized language databases and corresponding software development and research. With regard to the software for language feature research, the research results are still lacking in four aspects [31]. Firstly, there is a lack of support for the type of corpus of scientific and technical academic discourse. Secondly, there is a single method for frequency-based word form search and calculation. In addition, the phraseological research approach is not relatively effective or comprehensive enough. What is more, the computer technology applied is quite old and computationally inefficient. To be specific, the software used in the current corpus linguistics research does not effectively integrate the basic functions of corpus linguistics research. For example, these basic functions mainly include word list creation, word collocation calculation and extraction, frequency-based block extraction, topic extraction of text, as well as lexical assignment.

In order to address the current situation and problems of the linguistic corpus and linguistic feature research, this study will make use of the features of the existing software for language retrieval, collocation calculation, word block extraction, and lexical assignment, and develop more functions for phraseology research. In addition, the algorithm will be optimized to improve the computational speed and efficiency of the corpus linguistics research. At the same time, this study will integrate existing semantic analysis research tools for secondary development, which will enable the reintegration of existing research tools.

2. Demand Analysis

2.1. System Demand Analysis. In software engineering, nonfunctional or functional system requirements analysis has an essential role. To be specific, both of them are the starting point of system development. In other words, the requirements analysis must be very precise in order to consider later implementation and testing issues. In addition, a thorough analysis of the requirement analysis can greatly shorten the development cycle by multiplying the effort in the design process. The advent of the 21st century has brought with it a number of very advanced modern technologies. In this context, many outstanding technological achievements have been presented to the world. At the same time, these technological products have brought more convenient production methods and a more enjoyable lifestyle to people's daily life. However, these new changes have also challenged the ability of human language to represent them. As a result of technology, a large number of scientific and technical-English language papers have emerged. At the same time, more and more countries are communicating internationally through English. The vocabulary of technical English has been enriched by a large number of scientific and technical terms that have been created as a result of scientific and technical updates. As a result, researchers who want to study the latest technologies or publish scientific articles on the frontiers of certain aspects of science and technology must be aware of the latest and most cutting-edge technologies at this stage. In addition, scientific and technical English is not only a crucial tool for the international exchange of scientific and technical materials and technology but also a leader in the development of science and technology and the advancement of science and technology in society.

In fact, as shown in Figure 2, the development of the analytical model of technical English has gone through two stages. The first stage was to use the general approach of stylistics to count and analyze the frequency of common tenses in English, and thus provide well-founded guidelines for the teaching of technical English. The second stage was to summarize the common patterns and individual characteristics of technical English issues through discourse analysis. Therefore, the exploration of research methods for scientific and technical English will be directly related to and affect the level of scientific and technical development in the

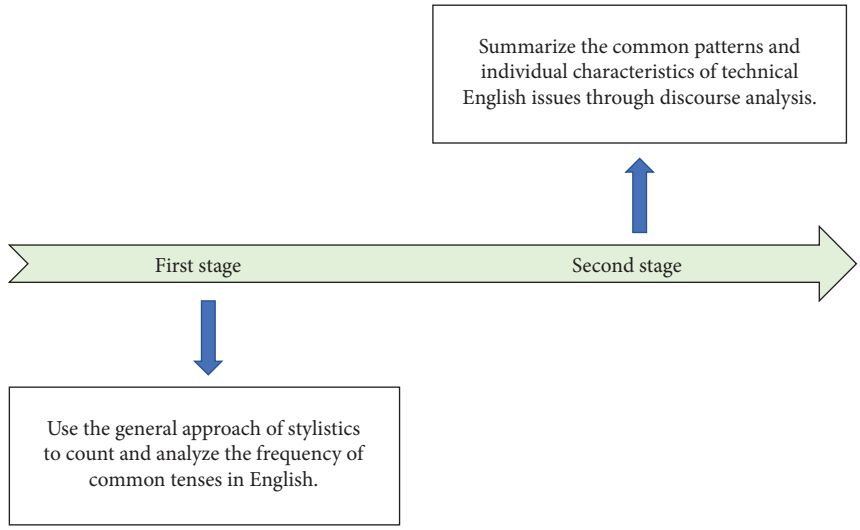

Figure 2: Development of the analytical model of technical English.

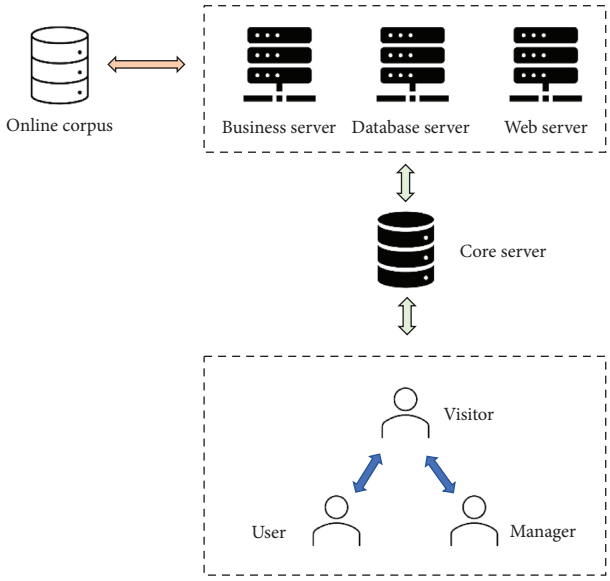

Figure 3: Basic structure of the online corpus.

country. This system is developed with such a goal in mind. As a result, a multifaceted and multifaceted analysis of the system's architecture, functional development, and integration is required to facilitate the completion and application of the system.

The basic structure of the online corpus proposed in this study is illustrated in Figure 3. The basic structure of the corpus is based on a browser and is divided into three levels: database, server, and user. All the studied articles are stored in the corpus as text files, which are used as the research corpus. At the time of use, they are retrieved by the server and made available to the user. As a result, users can easily access the corpus by simply visiting the web pages through their browsers. All business logic and functions are implemented on the server side. The physical server is the analysis system server, while the logical servers are the business server, the database server, and the web server. Among them, the corpus is connected to the database server. Therefore, users can use the functions by authenticating their logins. The roles include visitor, user, and administrator. To be specific, the visitor can authenticate and log in as a user, who can then use the business functions on the server. From the administrator login portal, visitors can authenticate and log in as administrators. After that, they can add, delete, change, and check the user data on the database server, as well as maintain the privilege of applying the online corpus.

2.2. Functional Module Demand Analysis. According to the frequency of use and research order, the system functions can be divided into general functions, special functions, and special functions. Among the general functions are data analysis and vocabulary search. The linguistic feature analysis cannot be performed without the selection of words or blocks of words, and there are simple exact searches in the requirements. In addition, there are also complex regular expression searches. Searching is the basis of research, so it is used frequently. Regardless of the choice of research tools or research factors, data analysis is necessary. After all, it is only the data that gives a visual picture of the value of the results. As a result, data analysis is a result of research and is used frequently. As shown in Figure 4, the feature function consists of three research methods, namely consistency search, word list creation, and subject terms. The feature function is the choice of research tools after the research object and research factors are established. In other words, it is in the middle position of the research. As a result, these three research tools are categorized as special features in this study. At the same time, these three instruments are only used in linguistic research and have the characteristics of linguistic research. The last category is the special functions, which can be divided into the cut, code assignment, and article upload functions.

In fact, the corpus is not a free and open-source project. As a result, it is necessary to set visitor status in the corpus. In other words, the visitor can only view the manual on the login

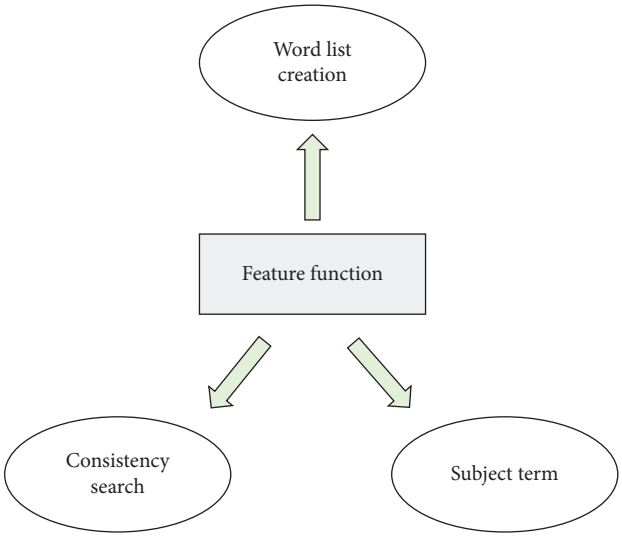

Figure 4: Components of feature function.

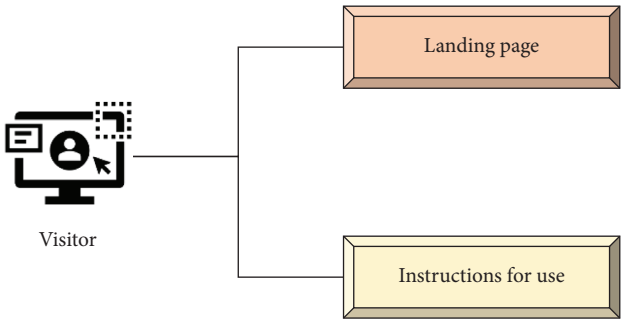

Figure 5: Use case for the visitor role to log in to the online corpus.

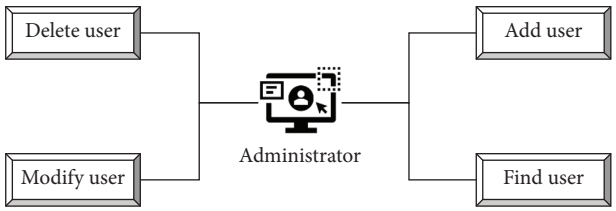

Figure 6: Use case diagram for the administrator role.

page to check how to use the corpus and the main features of the corpus. If the user is not logged in, he/she cannot use the features of the corpus. In this case, they can only get a general overview. If they need to use the corpus, they need to contact the administrator and obtain a username and password before they can use it. Figure 5 illustrates a use case for the visitor role to log in to the online corpus.

The administrator is only responsible for adding, deleting, changing, and checking users. Therefore, the administrator is responsible for all permissions in the corpus. If visitors request to use the system, they need to contact the administrator. The administrator can create user accounts and passwords according to the user's request. If a user wants to change or query their username or password, they need to contact the administrator to do so. However, administrators do not have access to the corpus. In other words, they exist only as user administrators. The use case diagram for the administrator role is given in Figure 6.

Table 1: Structure of the article information.

Field name	Type	Size	Description
txtitle	varchar	60	Title
txtauthor	varchar	80	Author
txtclassify	varchar	45	Classify
txtsource	varchar	45	Source
txtcontent	text	40	Content

Table 2: Clause information table.

Field name	Type	Size	Description
senid	bigint	8	Clause number
txtid	bigint	8	Article number
senindex	bigint	8	Clause index number
sencontent	varchar	1024	Clause content

2.3. English Text Collection. Although the preprocessing process can result in an electronic version of the corpus material, the material must be entered into a server-side database to achieve the established functions of the online corpus. As a result, database design is one of the key steps in the corpus design and development. In other words, a great database design can directly determine the efficiency of application development and database access. When the original text corpus is entered into the database, in addition to its own content, some relevant auxiliary information is added. This makes it easier to retrieve and analyze. In addition, it is necessary to set an article number field, which is used as the primary key and is automatically incremented. The structure of the article information is shown in Table 1.

In addition, in order to facilitate the analysis and statistics of linguistic phenomena in English sentences, the text of the corpus needs to be processed with automatic clause separation. Each sentence after processing is stored in the clause information, as shown in Table 2. The sentence index number field in the data table is applied to store the position of the first character of the sentence in the article, so as to facilitate the positioning of the sentence. The foreign key is used to correlate with the article.

In summary, according to the characteristics of the online corpus and its management system, in order to ensure the integrity and security of the data, four tables need to be created in the database, and these tables can store the original article information, clause information, word separation information, and user information.

3. Relevant Technologies Used to Construct Corpus

3.1. Data Analysis. Density analysis is the process of counting the number of individual words or word chunks compared to the vocabulary of the entire study factor and the

proportion of the study factor or corpus. As a result, density analysis provides a visual indication of the importance of a word in an article. If the weight is high, then the word has a high number of recurrences in the article, indicating that the word is important. This function can be combined with the vocabulary search function to retrieve not only single words but also blocks of words or regular expressions for density display.

Mutual information (MI) values are a common method used to calculate the strength of word collocations. The use of MI in linguistics is different from its common use in finance, especially in the range of values. The MI values in information science are in the range of 0 to 1, while the MI values used in linguistics are only in the 0-digit interval. The higher the value, the greater the mutual encounter and attraction between words. To be specific, MI values calculate the frequency of occurrence of a word in a corpus and provide information about the probability of occurrence of another word. The MI values can be calculated as follows:

$$M(x, y) = \log_2 \cdot \frac{Z(x, y) \cdot S}{Z(x) \cdot Z(y) \cdot 2D}, \quad (1)$$

where x and y refer to two words randomly distributed in the corpus, S indicates the total corpus capacity, D is the span, $Z(x)$ and $Z(y)$ denote the actual observed frequencies of their occurrence in the corpus, and the frequency of occurrence is $Z(x, y)$.

3.2. Data Cleaning. There is a class of methods in parameter estimation called maximum likelihood estimation. Since the estimated functions involved are often exponential families, taking the logarithm does not affect its monotonicity. However, this would make the computational process easier. Therefore, the logarithm of the likelihood function is used in this study. Depending on the model involved, the logarithmic functions will be different, but the principles are the same. Specifically, they are all determined by the density function of the dependent variable and involve the assumption of a distribution of random disturbance terms. In effect, it is a function value. However, it is possible to construct test statistics such as the likelihood function ratio based on the likelihood function value, and these test statistics can be used to test the significant properties of the model. In the online corpus, the log-likelihood function values are mainly used to determine the collocation relationship values of two words within the left and right span of the node word. Therefore, it can show the collocation rate of collocations within the left-right span and thus help to quickly locate the search for effective collocations near the nodal words. The calculation of the log-likelihood value can be determined as follows:

$$L = 2m \cdot \log_e \frac{m}{K \cdot m + n/K + J} + 2m \cdot \log_e \frac{m}{J \cdot m + n/K + J}, \quad (2)$$

where L refers to the log-likelihood value, K is the amount of vocabulary in the corpus K, and J is the amount of vocabulary in the corpus J.

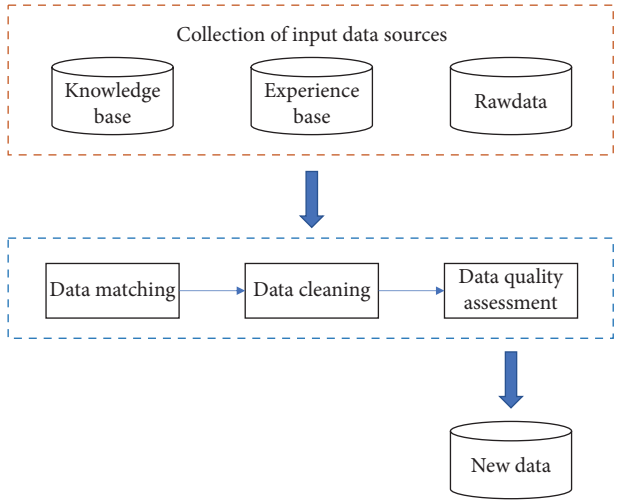

Figure 7: Structure of the data cleaning platform.

In the process of the corpus construction, data cleaning plays a very important role. The data cleaning platform mainly adopts the filter model and a hierarchical building block technology to realize data integration, conversion, cleaning, and validation in the process of data cleaning by using filter processing elements. At the same time, the corpus is equipped with data integrity analysis components, data conversion components, statistical analysis components, visualization components, and data mining algorithm components. As a result, it can effectively build a good, scalable, and easy-to-integrate soft bus architecture system. In this study, the structure of the data cleaning platform is shown in Figure 7.

4. Conclusion

This research mainly discusses the background of building an online corpus, based on the English text collection and data cleaning technologies, the technologies used to develop the corpus, the overall design of the corpus, and the detailed development process of the corpus management system. The purpose of building an online corpus is to integrate existing English resources and facilitate teachers' research and lesson planning. At the same time, it is conducive to improving classroom teaching methods, fully mobilizing students' learning motivation, and improving teaching effectiveness. This corpus allows users to easily retrieve relevant corpus information from university English textbooks and to analyze and count the corresponding linguistic phenomena. In addition, this corpus is based on the B/S model, and all operations are performed in the client's browser. As a result, this corpus can be divided according to the source of the corpus and can be maintained and managed by users. Compared with the existing analysis system, which only has the means to compare corpus with corpus, this system adds the function of comparing corpus uploads. The system extends the scope of application by adding the comparison of corpus to uploaded corpus and uploaded corpus to uploaded corpus.

However, the online corpus designed in this study still suffers from the following deficiencies. Although the current analysis system can handle the data volume of the current corpus comfortably, it is inevitable that the size of the corpus will increase gradually over time and as the discipline progresses. Since the current solution has limited load capacity, it may not be able to accommodate the rapid increase in data volume. As a result, building a quality concurrent processing solution will be the focus of this system in the future.

Data Availability

The labeled datasets used to support the findings of this study are available from the author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Acknowledgments

This work was supported by the Tangshan Normal University.

References

[1] D. Poedjiastutie, Z. A. Amrin, and Y. Setiawan, "English communication competence: expectations and challenges (a case in Indonesia)," *International Journal of Applied Linguistics & English Literature*, vol. 7, no. 6, pp. 184–191, 2018.

[2] M. H. Shin, "Study of English teaching method by convergence of project-based learning and problem-based learning for English communication," *Journal of the Korea Convergence Society*, vol. 10, no. 2, pp. 83–88, 2019.

[3] R. Fujita, M. Terui, T. Araki, and H. Naito, "An analysis of the English communication needs of people involved in tourism at Japanese rural destinations," *Journal of Global Tourism Research*, vol. 2, no. 1, pp. 53–58, 2017.

[4] S. Sutiyatno, "The effect of teacher's verbal communication and non-verbal communication on students' English achievement," *Journal of Language Teaching and Research*, vol. 9, no. 2, pp. 430–437, 2018.

[5] S. Degaetano-Ortlieb and E. Teich, "Toward an optimal code for communication: the case of scientific English," *Corpus Linguistics and Linguistic Theory*, vol. 18, no. 1, pp. 175–207, 2022.

[6] S. H. Ting, E. Marzuki, K. M. Chuah, J. Misieng, and C. Jerome, "Employers' views on the importance of English proficiency and communication skill for employability in Malaysia," *Indonesian Journal of Applied Linguistics*, vol. 7, no. 2, pp. 315–327, 2017.

[7] R. Qian, S. Sengan, and S. Juneja, "English language teaching based on big data analytics in augmentative and alternative communication system," *International Journal of Speech Technology*, vol. 25, no. 2, pp. 409–420, 2022.

[8] H. R'boul, "Intercultural communication dialectics in English language teaching," *International Journal of Society, Culture & Language*, vol. 9, no. 1, pp. 30–42, 2021.

[9] M. A. S. Khasawneh, "The degree of practicing organizational justice by teachers of learning disabilities in English language from their point of view," *Journal of Asian Multicultural Research for Educational Study*, vol. 2, no. 2, pp. 1–8, 2021.

[10] J. H. Han, Y. Wang, and M. Naim, "Reconceptualization of information technology flexibility for supply chain management: an empirical study," *International Journal of Production Economics*, vol. 187, pp. 196–215, 2017.

[11] S. A. Asongu and N. M. Odhiambo, "Foreign direct investment, information technology and economic growth dynamics in Sub-Saharan Africa," *Telecommunications Policy*, vol. 44, no. 1, Article ID 101838, 2020.

[12] P. Saeidi, S. P. Saeidi, S. Sofian, S. P. Saeidi, M. Nilashi, and A. Mardani, "The impact of enterprise risk management on competitive advantage by moderating role of information technology," *Computer Standards & Interfaces*, vol. 63, pp. 67–82, 2019.

[13] J. Ghosh, "The blockchain: opportunities for research in information systems and information technology," *Journal of Global Information Technology Management*, vol. 22, no. 4, pp. 235–242, 2019.

[14] T. Ravichandran, S. Han, and S. Mithas, "Mitigating diminishing returns to R&D: the role of information technology in innovation," *Information Systems Research*, vol. 28, no. 4, pp. 812–827, 2017.

[15] K. Ackerley, "Effects of corpus-based instruction on phraseology in learner English," *Language, Learning and Technology*, vol. 21, no. 3, pp. 195–216, 2017.

[16] L. Lowphansirikul, C. Polpanumas, A. T. Rutherford, and S. Nutanong, "A large English–Thai parallel corpus from the web and machine-generated text," *Language Resources and Evaluation*, vol. 56, no. 2, pp. 477–499, 2021.

[17] R. Esfandiari and F. Barbary, "A contrastive corpus-driven study of lexical bundles between English writers and Persian writers in psychology research articles," *Journal of English for Academic Purposes*, vol. 29, pp. 21–42, 2017.

[18] Y. Liu, L. J. Zhang, and S. May, "Dominance of Anglo-American cultural representations in university English textbooks in China: a corpus linguistics analysis," *Language Culture and Curriculum*, vol. 35, no. 1, pp. 83–101, 2022.

[19] M. Kytö and T. Walker, "A standardization process in its final stages: mine and thine in A corpus of English dialogues 1560–1760," *International Journal of English Studies*, vol. 20, no. 2, pp. 95–116, 2020.

[20] R. Love, C. Dembry, A. Hardie, V. Brezina, and T. McEnery, "The Spoken BNC2014: designing and building a spoken corpus of everyday conversations," *International Journal of Corpus Linguistics*, vol. 22, no. 3, pp. 319–344, 2022.

[21] W. Gong, "An innovative English teaching system based on computer aided technology and corpus management," *International Journal of Emerging Technologies in Learning*, vol. 14, no. 14, p. 69, 2019.

[22] J. Mackenzie, "Sentiment and confidence in financial English: a corpus study," *Russian Journal of Linguistics*, vol. 22, no. 1, pp. 80–93, 2018.

[23] N. Põldvere, V. Johansson, and C. Paradis, "On the London–Lund Corpus 2: design, challenges and innovations," *English Language and Linguistics*, vol. 25, no. 3, pp. 459–483, 2021.

[24] G. Wang, "A corpus-assisted critical discourse analysis of news reporting on China's air pollution in the official Chinese English-language press," *Discourse & Communication*, vol. 12, no. 6, pp. 645–662, 2018.

[25] A. S. Haider and R. F. Hussein, "Analysing headlines as a way of downsizing news corpora: evidence from an Arabic– English comparable corpus of newspaper articles," *Digital Scholarship in the Humanities*, vol. 35, no. 4, pp. 826–844, 2020.

[26] M. M. Ismail and F. N. Harun, "Modern standard Arabic online news discourse of men and women: corpus-based analysis," *Asian Journal of Behavioural Sciences*, vol. 3, no. 1, pp. 24–39, 2021.

[27] Z. Hua, M. Handford, and T. J. Young, "Framing interculturality: a corpus-based analysis of online promotional discourse of higher education intercultural communication courses," *Journal of Multilingual and Multicultural Development*, vol. 38, no. 3, pp. 283–300, 2017.

[28] S. M. Chua, "Compiling and analysing a large corpus of online discussions to explore users' interactions," *Applied Corpus Linguistics*, vol. 2, no. 2, Article ID 100017, 2022.

[29] P. Baker, "Language, sexuality and corpus linguistics: concerns and future directions," *Journal of Language and Sexuality*, vol. 7, no. 2, pp. 263–279, 2018.

[30] H. Abdul Al Salam Jassim and E. A. Jaafar, "the language of suicide notes: a corpus-based stylistic analysis," *International Journal of Language and Literary Studies*, vol. 4, no. 1, pp. 19–32, 2022.

[31] S. A. Joharry and S. Turiman, "Corpus stylistic analysis of Malaysian online columnists," *Journal of Modern Languages*, vol. 30, no. 2, pp. 53–80, 2020.