Application of Recommendation Algorithm and Big Data Technology in Computer English Corpus Database

Qingling Wang*
Weifang Engineering Vocational College, Qingzhou City, Shandong, 262500, China

*Corresponding author: wangqingling2020@wfec.cn

Abstract. Corpus linguistics is one of the branches of modern linguistics. It is also a cutting-edge research problem of modern linguistics spawned by the background of the information technology era. Corpus linguistics can reset the teaching role of English majors. College English vocabulary teaching makes use of the corpus's abundant real language materials and the computer's ability to analyze and process large-scale corpus, guide students to explore the usage of vocabulary in the context, and use the corpus to deepen students' autonomous learning. This article will start with the development of corpus linguistics, explain the necessity of combining corpus and English language teaching, and propose a corpus linguistics-based English language learning platform system in colleges and universities, and combine with semantic recommendation algorithms to verify that the system is in English learning.

Keywords: Computer corpus, expected material, English language learning, corpus system, recommendation algorithm modern corpus linguistics.

1. Introduction
Nowadays, corpus linguistics, as an independent subject, has its deep theoretical foundation and unique research paradigm, and it has become a system in research tasks, objects, and methods. The corpus, as a data source and method, is also widely used in other different types of research, including language teaching. In fact, the application of corpus in language teaching is not a new topic. As the founder of modern corpus linguistics, J. Sinclair has shown his feelings for English teaching very early in his academic thoughts, and this feeling has always been consistent. Enduring for a long time, and profoundly influenced his contemporaries and a group of scholars afterwards [1]. The basic requirement of effective English teaching for teachers is to have good English ability and understand some of the essential characteristics of English. Corpus linguistics research provides teachers with a new perspective on language. Language is not a universal system that can be abstracted from specific speech activities, nor is it a natural language ability that exists in the human brain. Language is true. The natural language used in the process of language communication. Therefore, learning language is not about learning abstract and arbitrary language symbols, nor is it the universal grammar advocated by the theory of linguistic innateness, but learning the real and naturally occurring language. However, natural language is huge, diverse, and varied, which makes people feel at a loss. Without a grasp, it is of course impossible to link up the complex language use, let alone systematic teaching. Corpus linguistics is not only a research method that uses probabilistic methods to analyze language facts, but
also has formed new research fields from early word frequency statistics, dictionary compilation, to speech recognition and generation, and data-driven linguistics [2]. Corpus Linguistics adopts quantitative or qualitative methods, and focuses on the application of language rather than language ability. It is mainly concerned with description rather than exploration of the universality of language. Corpus and corpus linguistics have become an indispensable basic resource in linguistic theoretical research and applied linguistics research.

2. Design of English corpus teaching platform

The curriculum resources for constructing the network English phonetic corpus mainly include: (1) The construction of basic knowledge of English phonetics and basic training content. The basic principles of the research and formulation of the corpus content are based on the goals and requirements of the English phonetics syllabus for English majors in colleges and universities. The training of English phonetics is closely integrated, emphasizing the training of basic skills, and paying attention to practical exercises. (2) The construction of cross-curricular training content in English phonetics teaching: English phonetics lessons are closely related to the learning of other professional courses, so basic English, listening, and general the teaching materials of reading and speaking courses are included in the corpus to help students further strengthen vocabulary memory, text comprehension, and spoken expression while performing voice training, and fully reflect the practical value of voice learning. (3) Voice training content based on network resources Construction: In order to enhance the communicative, interactive, and instantaneous nature of voice learning resources, collect and store online English voice corpus (including culture, economy, politics, history, education, religion, sports, etc.) according to the voice training content for students to speak Practice. (4) Establish a friendly and personalized speech corpus: In order to fully meet the individual learning needs of learners, the network speech corpus platform allows students to independently select and collect corpus that is compatible with their speech level and learning interests for learning and training.

2.1. Overall architecture

The system adopts a three-tier architecture system, which is divided into a presentation layer, a logic layer, and a data layer. Construct a flexible and extensible platform through a three-tier structure, as shown in Figure 1.

![Fig. 1 Three-tier system framework of the system](image-url)

The presentation layer module realizes the human-computer interaction between the management system and the user. Beginners can enter the virtual English environment through smart wearable devices, and realize various interactive operations between the system and the user through the module. The registration/login module provides user management functions. After registering an account in the system, learners get the assigned default configuration file and choose the module by themselves for
personalized customization. In the virtual environment, learners participate in courses according to their own learning process and hobbies. The course system is organized in a tree structure [3]. A large course is associated with multiple sub courses, and each sub course is connected to the parent course in a tree. Each course and series of courses are marked with tags, which is convenient for learners to search and choose directly. The learner confirms the selected course through the course center of the application interface, and the user center provides personal information, personalized learning path, and previous learning test evaluation results.

The middle logic layer includes various modules of middle logic. The Langue/AI module includes some English language frameworks and some AI functions, such as intelligent dialogue robots, speech recognition, and writing evaluation. In order to efficiently organize story paths and training scenarios, this article splits the system logic layer and actual functional applications, and introduces scripts to control the training ground and ensure system flexibility. The control module mainly manages the 3D environment and animated expressions. The virtual tools include a series of virtual tools, such as using whiteboards, virtual tools, and videos, to simulate graffiti through the whiteboard interface, and share the graffiti content with others in real time, supporting multiple people at the same time write. In addition, the system provides documents and videos for learners to display. The social module is used for learners’ social interactions. Friends lists and specific information are added to achieve mutual visits and exchanges between learners. Other functional modules include VR equipment management, multi-control modules for dictation and so on.

The data layer includes data servers, file servers and synchronization servers. The data server includes a small cache server, which is convenient for quickly obtaining data, creating different scene environments and dialogue interfaces. The file server includes file data such as course documents, PPT, writing documents, and comparison documents. Real-time data synchronization of each client is realized through the synchronization server, including role positions, actions, and simulated environment content.

2.2. Function analysis
The main tasks of computer corpus processing include corpus establishment, corpus processing and corpus information retrieval. In view of the characteristics of domestic English teaching, the CMS system should be able to manage the corpus of both English and Chinese. The main functions that should be provided are: (1) Corpus management users can create an empty corpus according to their needs; open the existing corpus; close the corpus; Delete the corpus; (2) Corpus management text corpus storage, corpus deletion, corpus modification, corpus printing, corpus is saved as a separate text file; (3) Document conversion can convert Word documents, HTML format documents, etc. and save them Enter the corpus; (4) The corpus classification can classify the corpus according to the specified category, and the category can be multi-level; (5) The corpus labeling can add various features to each corpus, including: title, author, and time of entry, Original publication time, source, warehousing, etc.; (6) corpus retrieval can be searched by category and label, or full-text search can be conducted by corpus content. It can be used to conduct research on word usage, phrase collocation, etc., and to find example sentences

And other work; (7) Word frequency statistics Count the usage frequency of words or phrases.

2.3. Immersive learning platform
Immersion learning uses a smart, interactive, and interesting virtual environment to improve learners’ enthusiasm and make the learning room closer to native English speakers. The main scene of the system simulates a virtual group in an all-round way. Many learners can enter a scene at the same time and communicate with each other. Multiple AI characters are designed in the scene. When the learner approaches the AI character, the AI character takes the initiative to greet the learner. In different scenarios, there will be different AI roles to provide various scenarios, as shown in Figure 2.
Learners interact with AI characters in the corresponding scene according to their own needs. At the same time, each learner in the scene can see the dynamics of other learners. Several everyday language scenarios are designed in the system, such as formal speeches, debates, broadcasts, and interviews. Take the interview scene as an example, as shown in Figure 3.

This scenario sets the interview scenario where learners are going to study in the United States and enter the embassy to apply for a visa. The entire simulation scene is carried out in the mock interview room, and the interviewer is a virtual character constructed through AI technology [4]. During the interview process, the interviewer will ask the learners about the details of visa and study abroad according to the actual interview questions given in the system, ask the learners to answer correctly, and make different responses according to the learners’ responses, indicating that they are responding...
to the interview Approval or doubts of the answer, the interview result is given at the end, and the problems and advantages existing in the interview are given for the learner to correct.

3. Interest recommendation algorithm for student corpus materials
There are four main types of corpus retrieval: English word or single Chinese character retrieval, phrase retrieval, fuzzy retrieval, and comprehensive retrieval. This kind of retrieval is easier to implement under the first index structure, which has been explained in the previous section. The other three are introduced below.

3.1. Phrase search
Taking the retrieval of two English words or two Chinese characters as an example, it can be divided into two situations. One is the "AB" type phrase, that is, word A and word B are joined together to form a phrase, such as "look after" and "university". The second is the "A... B" type phrase, where two words are in the sentence but not connected together, such as "take...in" and "as...as". This kind of phrase retrieval is not very used in general full-text retrieval, but it is very needed in English teaching and research.

1) For "AB" type phrase retrieval, the following algorithm can be used: Find the corpus CA containing word A and the corpus CB containing word B in the index structure;

\[\text{For all } C \in CA, \text{ do (3)-(6)}\]

\[\text{If } C \in CB \text{ then (4) else }\]

\[\text{We take out the position chain } \text{PA of word A in corpus C, and take the position chain } \text{PB of word B in corpus C.} \]

\[\text{For all } P \in PA \in \text{do (6)}\]

\[\text{If "AB" is an English phrase, and} \]

\[P + \text{len (A)} + 1 \in PB\]

Then P is a position of the phrase "AB" in corpus C; if "AB" is a Chinese phrase, and \(P + 2 \in PB\), then P is a position of the phrase "AB" in corpus C.

2) The retrieval of "A... B" type phrases is relatively difficult. In order to judge whether word A and word B appear before and after the same sentence, the sentence index of the corpus was established in the design to record the appearance position of each sentence in the corpus. The sentence index structure is shown in Figure 4. Among them, CorpusId is the corpus number, s-count is the number of sentences in the corpus, and s-link points to the end position list of each sentence in the corpus [5]. For example, the corpus Ci in Figure 4 has 100 sentences, and the corresponding ending position is

\[P_{i1}, P_{i2}, P_{i3},..., P_{100}\]

The idea of compressing the sentence ending position list is the same as the word index. For "A... B" type phrase retrieval, the following algorithms are used: 1. Find the corpus CA containing word A and the corpus CB containing word B in the index structure.

\[\text{for all } C \in CA, \text{ do (3)-(8)}\]
if \( C \in \text{CB} \), then (4) else (2) \hspace{1cm} (7) \\

Take out the position chain \( PA \) of word A in corpus \( C \), take out the position chain \( PB \) of word B in corpus \( C \); take out the sentence position list \( SC \) of corpus \( C \);

for all \( P \in PA \), do (7)~(8) \hspace{1cm} (8) \\

Find the ending position \( PE \) of \( P \) in \( SC \). If there is \( P^* \in PB \) and \( P < P^* < PE \), then \((P, P^*)\) is a position of the phrase "A... B" in the corpus.

![Fig. 4 Sentence-level index](image)

### 3.2. Fuzzy search

In English teaching, I often want to check the usage of the derivative words of the same root in the corpus, so the fuzzy retrieval function is designed, that is, the wildcard \(^*\) is allowed. For example, "object*" will automatically find out where the words "object", "objective", "objection" etc. appear. According to the index structure we designed, it is easier to realize this kind of retrieval.

### 3.3. Comprehensive search

The comprehensive search can not only specify the structured information such as the title, category, author, source, and year of the corpus to be retrieved, but also specify the words or phrases contained in the corpus [6]. The comprehensive search can be realized by combining structured search and full-text search.

### 4. Anticipation library application effect detection

In English, there are a lot of phrases composed of \( \text{at} \) + the + noun + of. Using corpus tools for reference can quickly and effectively help students master this type of phrases or the frequently occurring parts of such phrases. Search for at the [nn*] of in the website http://corpus [7]. byu. edu /bnc / to get a series of phrases, ranked as follows according to frequency

|   | AT THE END OF   | 24866 |
|---|----------------|-------|
| 2 | AT THE UNIVERSITY OF | 15353 |
| 3 | AT THE TOP OF    | 6607  |
| 4 | AT THE TIME OF   | 5532  |
| 5 | AT THE BEGINNING OF | 5244  |
| 6 | AT THE AGE OF    | 3818  |
| 7 | AT THE BOTTOM OF  | 3715  |
| 8 | AT THE CENTER OF  | 3590  |
| 9 | AT THE HEART OF  | 3234  |
| 10| AT THE EDGE OF   | 3145  |
It can be seen from Table 1 that the most frequently used phrases in the structure of at + the + noun + of are at the end of, followed by at the university/top /time /beginning of. By observing the above phrases, we can find that this type of structure is mostly used to indicate the beginning and ending points (at the beginning /end of), time or age (at the time /age of), or spatial orientation and location (at the top /centre /bottom/edge / University of) Teachers can consciously guide students to pay attention to the main manifestations and meanings of such phrases in teaching.

If you switch to the chart method, you will find the stylistic features of the at + the + noun + of structure: it appears the most in journals and magazines, followed by academic articles and novels, and the least in spoken language [8]. From the above results, it can be found that this type of segment is mainly used in formal and rigorous written language, while the frequency of use in spoken language is relatively low. As shown in table 2.

**Tab. 2** Stylistic features of at + the + noun + of structural phrases

| SECTION   | FREQ   | PER MIL | SEE ALL | SUB - SECTIONS | AT ONCE |
|-----------|--------|---------|---------|----------------|---------|
| SPOKEN    | 19355  | 214.90  |         |                |         |
| FICTION   | 29897  | 351.87  |         |                |         |
| MAGAZINE  | 31177  | 345.29  |         |                |         |
| NEWSPAPER | 26367  | 304.22  |         |                |         |
| ACADEMIC  | 30742  | 358.33  |         |                |         |
| 1990−1994 | 33325  | 320.44  |         |                |         |
| 1995−1999 | 32707  | 316.17  |         |                |         |
| 2000−2004 | 33364  | 324.11  |         |                |         |
| 2005−2009 | 32643  | 319.95  |         |                |         |
| 2010−2011 | 7001   | 275.94  |         |                |         |

The corpus can also be used to help teachers set up exercises. In the KWIC mode, the above search results can show the usage of this type of phrase in various contexts. The source of the examples comes from the real experience of native speakers, and they are by no means a product of imagination based on grammatical rules. They are very useful for helping students master the original English. Teachers can also design these sentences into practice forms so that students can master different phrases of similar types as soon as possible.

5. Conclusion
The article discusses the function and structure of the corpus management system CMS and the solution of some technical problems. The CMS system has been implemented with Delphi and VC++ on the WIN95/NT platform. The use of the School of Foreign Languages of our school shows that the system has basically reached a practical level and can play an active role in English teaching and research.

Acknowledgments
This work was financially supported by the fund project: Shandong Education Science "13th five-year plan" special project "Research on Information Literacy Cultivation Mechanism of higher vocational foreign language teachers", project number: 2020WGYB003.

References
[1] Futrell, R. Gibson, E. Tily, H. J. Blank, I. Vishnevetsky, A. Piantadosi, S. T. & Fedorenko, E. The Natural Stories corpus: a reading-time corpus of English texts containing rare syntactic constructions. Language Resources and Evaluation, vol.55, pp.63-77, January 2021.
[2] Kaity, M. & Balakrishnan, V. An automatic non-English sentiment lexicon builder using unannotated corpus. The Journal of Supercomputing, vol. 75, pp.2243-2268, April 2019.
[3] Yoo, I. W. & Shin, Y. K. Determiner use in English quantificational expressions: A corpus-based study. TESOL Quarterly, vol. 54, pp. 90-117, January 2020.

[4] Sadia, S. & Ghani, M. Modality in editorials of Pakistani English newspapers: A corpus based study. International Journal of English Linguistics, vol. 9, pp. 144-151, January 2019.

[5] Quesada, T. & Lozano, C. WHICH FACTORS DETERMINE THE CHOICE OF REFERENTIAL EXPRESSIONS IN L2 ENGLISH DISCOURSE?: NEW EVIDENCE FROM THE COREFL CORPUS. Studies in Second Language Acquisition, vol. 42, pp. 959-986, May 2020.

[6] Karimullah, K. Sketching women: a corpus-based approach to representations of women's agency in political Internet corpora in Arabic and English. Corpora, vol. 15, pp. 21-53, January 2020.

[7] Liu, D. & Myers, D. The most-common phrasal verbs with their key meanings for spoken and academic written English: A corpus analysis. Language Teaching Research, vol. 24, pp. 403-424, March 2020.

[8] Kotowski, S. The semantics of English out-prefixation: a corpus-based investigation. English Language & Linguistics, vol. 25, pp. 61-89, January 2021.