Application Research of Cognitive Linguistics Based on Big Data Internet Corpus Construction

Jianbo Xu*
School of Foreign Languages, Xiangnan University, Chenzhou, China, 423000
*Corresponding author e-mail: tonyxu5132@126.com

Abstract. Corpus plays a very important role in language research. With the development and application of computer technology, database has been widely used in linguistic research, and it shows a strong advantage in the collection, processing and processing of corpus. Corpora, which collect and organize a large amount of language information in real use, have become an important tool for linguists to conduct academic research.

1. Introduction
Pragmatics is an old and young subject, and the study of pragmatics is not only an innovation of theory and analytical method. Modern science shows that the updating of research methods and operating methods has become an opportunity for the development of theory. The study of linguistics needs to collect, add and process corpus, and emphasize the authenticity and objectivity of corpus. Scholars who study language can feel the importance of the way and means of corpus processing.

In the face of large scale text corpus, the first task for linguistic researchers is to process and select the large scale and real text corpus in order to explore new language rules or information needed from these texts. Using manual processing method to carry out statistics, classification, analysis and other operations on Marine corpus data, the work efficiency will be very low, and the labor intensity is very high. This has also led to changes in the way language materials are handled and operated. Database technology is a comprehensive technology of data management automation and a warehouse for organizing, storing and managing data. With the extensive application of computer technology and the popularization of network, data base has been widely applied and popularized in the fields of production, life and science and technology. It also shows a strong advantage in the study of linguistics, and has become an important tool for linguists[1].

2. Database and linguistics
A database, in the name of thinking, is a warehouse where data is stored. Except that the warehouse is stored in a certain format on computer storage devices and has powerful data organization, data processing and data management capabilities. The corpus of language learning research is characterized by large scale, multiple types, heterogeneity and fragmentation, etc., so that the use of database management can improve the work efficiency of linguistic researchers. In order to store and manage all kinds of linguistic materials effectively, the primary task of designing a linguistic material database is to define the linguistic material data reasonably and establish a linguistic material database that can process and process the data. The definition of data in a database includes the format of data storage, the type of data storage and the association of data. The definition of data is the key to the use of data library. A good definition can facilitate users to count, retrieve and process data. The database
management system has powerful data definition function, which provides the possibility for linguistics researchers to use the database. The application of database in linguistics is mainly reflected in the addition and processing of language materials. The strong data definition function of the database provides the possibility for the storage of language materials. More importantly, language researchers can process and process the vast amount of corpus through the defined language materials. The database has powerful data processing capabilities, and language researchers can obtain the information of language materials through a variety of operational methods, such as corpus screening, corpus sorting, corpus classification, corpus statistics and corpus update, etc. He Leshi for example, the function words in zuo "research"[2], the articles are written in the 1960 s, when the main open copy card accumulation of data, analysis, statistics, these work need to spend a lot of time and energy, and obtained a few days or even more than ten days to get a result, investigators now are using the data library or corpus, between the instantaneous input search content was able to get accurate results. This is also the biggest benefit of the database, save time and effort.

At present, with the popularization of computer network and the establishment of various language material databases, it is possible for language researchers to maximize the possession of language materials, which fully embodies the resource sharing of database. Language research also has different research objectives and corpus requirements. In order to make the database full of different topic requirements, a more extensive data resource database needs to be established. Of course, the sources and types of linguistic materials are different, so these heterogeneous and heterogeneous resources should be properly defined and grouped. The study of linguistics needs the results of existing language studies. The database of language materials can ensure the reliability of language materials and meet the needs of the continuous development and update of linguistics, which is also the characteristic of the data base itself.

With the development of computer, database has been widely used in linguistics and has become an important tool for linguists. Using database technology to make it structured, relational, integrated, stable and other special points, the data warehouse of language materials - corpus is established. Corpus is a data warehouse for storing the original language materials, which is a database specially integrated for research through scientific collection and organization of a large number of language information used in the real situation. With the development of computing technology in recent years, researchers and organizations have established a wide range of corpora.

3. Corpus linguistics
The design of corpus requires the continuous collection and accumulation of a large amount of real corpus. The so-called corpus is the repository of a large amount of factual information. But a strictly defined corpus refers only to "a text library consisting of a large collection of written or spoken languages stored and processed by computers for linguistic research "Linguistic researchers conduct linguistic research on the basis of corpus[3].

Corpus linguistics reflects two levels of signification: one is a new means of linguistic research, that is to use corpus to study a certain aspect of linguistics; The second is the new viewpoint and theory of linguistics reflected in corpus, which is a branch of linguistics, such as sociolinguistics, pragmatics and so on. From the present literature, most of them regard it as a new means of scientific research on linguistic corpus.

3.1. The function and development of corpus
Corpus is a basic resource base for the collection, processing and processing of linguistic text information in accordance with certain linguistic principles. Corpus is a basic resource base for the collection, processing and processing of linguistic text information in accordance with certain linguistic principles[4]. The traditional language research is to acquire language materials by manual methods, and it is impossible for any language research researcher to memorize and process the whole language data. After the establishment of corpus,
accurate information can be quickly obtained from the vast language database. The most immediate use of corpus is to provide a large number of real and accurate examples for the compilation of lexicography. For example, there are seven kinds of dictionaries compiled according to the COBUILD corpus. In addition, Leech(1993) also listed more than a dozen uses of corpus for language theoretical research, grammar research, semantic research, discourse analysis research, etc, explaining the importance of corpus in linguistic research field and the promoting role of corpus in linguistic research field.

Since the 1990s, the establishment and research of corpus have developed rapidly, from millions of words to tens of millions of words, and the depth of corpus processing has developed to the level of discourse, realizing quantitative analysis of corpus, machine translation, lexicon compilation, etc. For example, the COBUILD corpus in the UK has published more than 20 kinds of grammar, dictionaries and English textbooks on the basis of this corpus[5]. The Chinese corpus started relatively late, such as "modern Chinese grammar research corpus" of Beijing language and culture university, CCL corpus of Peking University and Beijing language and culture university interlanguage corpus, etc.

3.2. Research on language learning based on corpus
Corpus-based language research corpus has become a powerful tool for linguists to study pragmatics. The corpus-based language research has the following characteristics.

The trueness of corpus: All the materials in the learner's corpus are written by students collected by random sampling, which is the real language operation. The conclusions drawn from the analysis of the learner's Middle language are all grounded and factual. The English teaching in Our country has made great achievements, but there are also many problems. Only through scientific analysis and in-depth study of the teaching situation, can we put forward a specific and practical improvement plan, receive practical results, and avoid aimless and meaningless disputes.

Quantitative analysis: One of the salient features of research based on linguistic databases is data-driven. The storage and language processing power of computer provide powerful means for the observation of language, including the learner, which is difficult to imagine before. Quantitative analysis makes the description objective, and statistical inference can also avoid the subjectivity of judgment. Data-driven quantitative analysis allows us to see problems that were previously invisible to our intuitions. These statistics make it possible to explore the characteristics of Chinese students' English learning from a quantitative perspective. Of course, quantitative analysis must be supplemented by qualitative analysis in order to provide practical solutions for learning English well[6-7].

Group analysis: through data-driven quantitative analysis, we can also find that some speech errors are problems in other students' language operation and individual actions. Some speech mistakes are common among English learners in China, which makes us have to study their causes carefully. Is it the transfer of mother tongue? Or is it overgeneralization? Or is it due to an omission in the textbook or teaching? Only by finding the original cause can we find ways to improve teaching and improve the quality of teaching.

Longitudinal analysis: Learning is the corpus collected from all levels of the students' language output, at the stage of learning is not with him, there is a beginner, have already to a higher level of learning stage, so that we could longitudinal analysis on the development of English learning process, look to see what is easy to occur in the early scholars say language errors, what words mistakes to high level stage is less now, but on the other kind of speech errors of the frequency is increased, Thus we can find the law of learner's intermediate language development.
Corpus provides real linguistic data, statistical data, verification of existing theories and construction of new research theories for linguistics researchers, and plays a very important role in supporting language research from high abstraction to practical application of language.

4. Conclusion
Linguistic research must be based on linguistic facts, and the emergence of detailed and abundant linguistic data databases makes the collection, consolidation and processing of linguistic materials get rid of the tedious, time-consuming and laborious manual labor. Corpus has become an important tool for linguistic researchers. In recent years, corpus has gradually evolved from an auxiliary tool for linguistic research into a subject for analyzing linguistic phenomena from multiple perspectives. This reflects the updating and scientific trend of the ideas in language research field. The use of linguistic databases has permeated every field of language research step by step and has become an indispensable research tool for natural language researchers, lexicographers and language researchers.

Acknowledgments
1. This research was supported by Scientific Research Project of Hunan Education Department in 2019: Research on Cognitive Linguistics in Foreign Language Teaching under the Background of Internet + and big Data Corpus, (Grant No. 19C1718).
2. This research was supported by Research Project on Ideological and Political Construction of Curriculum of Hunan Education Department in 2020: Teaching Reform of Advanced English under the Background of Ideological and Political Concept of Foreign Language Curriculum, (Grant No. HNKCSZ-2020-0572).

References
[1] Jiang Hu: Database is an important Tool for linguists[J], Journal of Chinese Academy of Social Sciences, 2004,6.
[2] Lu Xia Li: The application of database in academic research is discussed from the perspective of linguistics[J], Cloud dream journal, 2008, 3.
[3] Naqin Ma: The application of Corpus Linguistics in the study of linguistics[J], Journal of Gansu Union University (social science edition), 2007, 23.
[4] Zhang Xiang Li: Design and Implementation of corpus System based on content topic[J], Computer application research, 2004, 21.
[5] Xiao Li Mu, Zhuan Yun Yi: The application of IX corpus in language research[J], journal of foreign languages institute, 2001,9.
[6] Feng Jia Gao: The application of Corpus linguistics in the study of vocabulary and discourse[J]Journal of shanxi datong university (social science edition), 2010, 24.
[7] Wen Jia: A Review of corpus Linguistics in China[J], Journal of Fuyang Normal University, 2006, 5.