Analysis on Linguistics Research Directions in the Age of Big Data

Tingyu Lu
Henan University, Kaifeng, 475000, China
haochideni@163.com

Abstract. With the development of science and technology in recent years, concepts such as big data and databases frequently appear in people's fields of vision. Among them, big data and the construction of foreign language corpora have become common practice in the field of academic research. It has a certain understanding by consulting related journals, papers and books. This article starts with what big data is, what it means to build a corpus for linguistic research, and the situation of linguistic research in the context of big data, and briefly explains the current status of data and linguistic research in the context of big data discussion and at the same time hopes to have a more comprehensive understanding and understanding of the database.

1. Introduction
With the rapid development of computer science and technology, people have ushered in the information age. An important feature of the information age is the emergence of large amounts of data. This has caused changes in people's lifestyles and ways of thinking, and also has a major impact on linguistics research. Linguistic research is inseparable from the collection, processing, and integration of language materials. Traditional language research is very inefficient in collecting corpus, which requires a lot of manpower to go through the classics and make records. Now through the combination of big data and language research, these problems are solved. According to different language research purposes, scholars will build a corpus of Tong content to complete the processing and processing of corpus. It can be said that big data opens up a new world of language research [1].

2. Big data
The most popular word in the field of computer science in recent years is "big data". For "big data", people were initially a hazy state of understanding, but with the development of technology and the use of big data, people began to realize the huge role and significance of big data. Big data not only changes people's daily life, but also has a subtle influence on people's way of thinking. At the same time, it also shortens the distance between different fields, which is of great significance to the development of interdisciplinary subjects [2]. It drives the development of various fields and makes people more clearly understand the world we live in. Big data is not only reflected in the hugeness of its data, but more importantly, through the hidden value of information behind these huge data-the pattern of human cognition, behavior, and the law of human, social, and natural interaction. In the past scientific research, the resources we can use are limited, or it is difficult to collect more comprehensive information content. In many cases, people summarize the general laws through limited examples. This often makes people ignore some important laws and even go in the wrong
direction. Big data can provide us with corresponding data information quickly and efficiently, and also provide relevant information, which is an operation that has not been in traditional research methods.

In this era of big data, we are inevitably affected by it. Through the network, we can quickly and easily find the information we need, and can also generate various data through editing and searching. We always receive, retrieve new information and process existing information. Through the retrieval and processing of information within a certain period of time in different regions, different fields, and different countries, we can have a preliminary understanding and pre-judgment of its status during that period of time. For example, In the Spring Festival Transport in 2018, we use big data to obtain the passenger traffic between and around the country and the transportation method they choose. The pressure relieves the pressure of the Spring Festival.

Language, as a human communication tool, is widely used in various fields. Throughout the ages, people have constantly opened up new things and new fields, and the corresponding language system has gradually formed and expanded accordingly. Therefore, big data provides a large amount of factual corpus for linguistic research. At the same time, it is also affecting the method of language research. Facing a large amount of factual corpus, we need to analyze, process and sort it scientifically, which requires a complete system tool to complete this task [3].

3. Corpus linguistics research tools

In the 1960s, corpus linguistics was gradually established. Corpus linguistics is based on empiricism and scientific rationalism as its philosophical foundation, with bottom-up induction as its methodology. Since its establishment, it has been based on real corpus research, based on a large number of daily terms, based on probability, and established models through statistical quantification. The study of linguistics should be based on linguistic facts and search linguistic materials extensively and meticulously. With the development of computers, it has become a trend to conduct language research by constructing corpora. A corpus is a warehouse that stores many factual corpora. Linguistics should pay attention to the study of the meaning of natural real corpus, so the data of corpus should not only pursue "quantity", but also pay attention to its "quality".

As a tool for language research, corpus needs to contain a large amount of real corpus. When constructing a corpus, text collection must have certain standards, and the amount of text collected must also reach the scale that the computer can handle. Among them, it is necessary to process the collected text by marking, coding and other processes. Corpus plays an important role in all fields of language research. In the field of language education, analysis and statistics of vocabulary word frequency, pronunciation, sentence pattern, collocation, etc., can have a very clear grasp of corpus. It is also widely used in the field of applied linguistics. For example, using related system software, the corpus can be studied for syntax, vocabulary, spoken language, semantics, language variation, and discourse analysis. In addition to this, speech synthesis and recognition tasks can also be performed by using corpus technology. In the field of sociolinguistics, analysis of sociolinguistic phenomena and language variation based on large-scale real corpus can obtain more objective and accurate data and conclusions. In the field of literature, develop a new research model that combines literary research with a corpus. By constructing a corpus of literary classics and work reviews, and processing and processing such as annotation and retrieval. The analysis of the characters, work styles, themes, images, etc. in the separate studies can provide more objective data support for the views [4]. In the field of translation, in order to improve the efficiency and accuracy of translation, a bilingual parallel corpus will be used. At the same time, the bilingual parallel corpus can also provide empirical material support for translation research and practice. In the field of quantitative language analysis, by using large-scale real corpus to design the knowledge points to be quantitatively analyzed and the various question types used, the efficiency and recognition of quantitative analysis results can be improved.

In the context of big data, we can share resources more conveniently and quickly, as is the corpus. For example: the free library "Blue Sky Library" constructed in Japan, which contains a large number of literary works of famous Japanese writers. This provides many researchers of Japanese literature
with an efficient way to collect first-hand information. Played a great role in language research. We can retrieve the required language information through the corpus, and then analyze and research through the corresponding software. This shows that the corpus is an important tool for language research. When retrieving information through the Japanese corpus, I accidentally found that they not only established their own corpus of their own language, but also built a relatively complete corpus for other languages. Through the use and observation of these corpora, we can recognize the degree of Japanese understanding of Chinese characters, and at the same time, it is of great help in the translation and acquisition of the two Chinese languages. The corpus is not only in the position of both languages, it should be multi-party. For example, both China and Japan will establish their own corpora of English, French, etc. By examining these corpora, we can see that countries that have acquired a second language have a degree of understanding and development of this second language. In fact, there are too many unknowns in the corpus that we want to retrieve.

4. Linguistics research situation
In the context of big data, language research has renewed its vitality. As a tool for information exchange, language is closely related to various fields. And big data has shortened the distance between various fields, which makes language research no longer an isolated subject. We must pay attention to the research and development of its related disciplines and inject new vitality into linguistic research. Brisnan's experience of language research in the era of big data is from "garden" to "bush." She believes that "linguistics in the garden" is mainly to analyze linguists' linguistic phenomena based on introspection or careful selection, and to qualitatively summarize and summarize through symbols such as syntactic trees. "Linguistics in the bush" is a study of the real language people use in real life, often using conditional probability, information, etc. for quantitative analysis. During this process, the situation of linguistic research will change significantly. The research situation of linguistics shows the following characteristics in the context of big data: First, the search for the causal relationship within the language facts turned to the correlation between language facts. Second, no longer excessive pursuit of precision at the micro level will allow us to have better insight at the macro level of the language. Third, language research has turned from finding causal relationships and constructing various language interpretation models to studying the relationship of language facts. All in all, every change of the times will leave traces in the language. We must trace and explore these traces and continuously promote the understanding of the language.[5] We must pay attention to the research of linguistics as well as the research of its related disciplines. In the context of big data, language research has new opportunities and challenges.

5. Interdisciplinary research on linguistics in the era of big data
In recent years, interdisciplinary research has become one of the hot words in academia. We know that at the earliest, the division of disciplines did not exist, and history. The number of people with the same knowledge and knowledge is innumerable. Later, due to the development of technology,
exploration methods have emerged in an endless stream, tending to be complex and diverse, and it is impossible for one person to master so much knowledge and skills at the same time, so the academic division of labor is more refined and a division of disciplines has been formed. After decades of refined research, we found that the refined method is similar to the method of using a blind person to touch the elephant, and the understanding of the elephant as a whole needs to be combined. Therefore, when discussing the same research object, they tend to adopt different methods and tools. For example, when studying languages, using the methods of biology, physics, or mathematics, so-called interdisciplinary or interdisciplinary appears.

The author's team has also achieved some results in the interdisciplinary research of linguistics in recent years. For example, it borrowed the method of complex networks to conduct typological research on Slavic languages. The mainstream of linguistic typology today is word order typology. When analyzing the Slavic language family with relatively rich morphological changes and relatively free word order, the past word order typology methods are not very suitable. We borrowed the method of complex networks from statistical physics. Based on the real texts of twelve Slavic languages, we used complex network indicators to classify these languages. You can refer to our two articles published in "Biophysics Review" that use complex networks to study the laws of human language and how to use dependent distances to discover the linearization pattern of human language.

In the two cases mentioned earlier, "cross" does not cross into physics. Of course, from the point of view of physics, it also expands the application field of complex network methods and provides real universal examples. In terms of language, the use of complex networks has helped us solve linguistic problems that were not easy to solve in the past. Of course, as the two disciplines borrow more frequently from each other and become more and more closely related, it is very likely to form a cross-discipline and may even form a new research paradigm. The deepening of the blending makes this new discipline different from the original Ren-discipline, for example, one day it may be unclear whether it is physical linguistics or language physics.

So does big data also contribute to interdisciplinary development? From a practical perspective, interdisciplinary research in linguistics requires a certain understanding of the "trans" field. If we define linguistics as the subject of "research on language structure patterns and evolution laws", this is of course a very narrow and traditional definition, because linguistics also covers a lot of content, but ultimately it still has to deal with linguistic data. When dealing with linguistic data, knowledge of statistics, mathematics, and computer science must be used, such as the software used by biology to study networks to study networks constructed from linguistic data, which also belongs to interdisciplinary research in linguistics. In addition, from the perspective of language as a complex adaptive system, the laws derived from real text materials may guide the computational linguistics and natural language processing that have great potential and development prospects today, so we are actually dealing with language data. Therefore, the data-based method obviously can also promote the interdisciplinary research and development of linguistics.

6. Conclusion
Linguistics is from the garden to the bushes, which is what the big data of the information age gives us. In the future, whether we will go from bush to forest depends on the development of technology and linguistics. As far as linguists who are new to bushes are concerned, they are faced with infinite novelties and challenges. How to make better use of it for linguistic research has become an important topic.

References
[1] Liu Haitao, Lin Yanni. Methods and trends of language research in the era of big data [J]. Xinjiang Normal University News, 2018 (1): 72-83.
[2] Lin Hailun, Wang Yuanzhuo, Jia Yantao, etc. Overview of knowledge fusion methods for network big data [J]. Chinese Journal of Computers, 2017 (1): 1-27.
[3] Yuan Changwan, Liu Chengyu, Chang Shuli. The combination of systemic functional linguistics and corpus [J]. Foreign Languages, 2016 (1): 104-109.

[4] Li Huayong. A new perspective on language research from the perspective of big data [J]. Journal of Chongqing Jiaotong University, 2015 (4): 134-137.

[5] zhan weidong. Research on Chinese language in the era of big data [J]. Journal of shanxi university (philosophy and social sciences edition). 2013 (5) : 70-77.