Corpus Linguistics Based on Integrated Learning and Its Role in Second Language Acquisition

Zhuofu Sun

Shenyang Aerospace University, No.37 Daoyi South Avenue, Shenbei New Area, Shenyang, 110136, China

Keywords: Integrated Learning; Corpus; Linguistics; Second Language Acquisition

Abstract: Corpus is mainly used in linguistic research and linguistic text base. It consists of a large number of collected written, spoken or other forms of corpus, and is stored and processed by computer. By introducing some basic situations of corpus research, people attach importance to the role and role of corpus research in SLA research. This paper reviews the important role of corpus linguistics in the process of second language acquisition and proposes an entity relationship extraction method based on ensemble learning. This method combines and transforms the feature features into feature vectors to illustrate the importance of establishing a learner corpus, and points out the advantages of this emerging subject area and the issues to be explored. The learner corpus is a corpus resource for teaching and can be used directly or indirectly for linguistics. Contrastive corpus linguistics can predict, diagnose and interpret learners' interlanguage errors, thus providing a powerful tool for second language acquisition research.

1. Introduction

Corpus linguistics is a distinctive linguistic research discipline. With the rapid development of computer technology, empirical-based research methods in linguistic research have been affirmed [1]. The ensemble learning method is chosen as the classification method. This method overcomes the over-fitting problem of training set, and it has strong generalization ability, and can improve the accuracy of classification. Monitoring corpus can also allow lexicographers to track the changes of word meanings in the diachronic corpus, so as to present more accurately the latest meanings of words and their origins and backgrounds [2]. In addition, the empirical evidence of the corpus can also verify the personal intuition of the lexicographer, because this intuition is not completely reliable, which can make the dictionary entries more accurate. From a methodological point of view, it can be used not only to study all aspects of the language system, but also in other fields beyond linguistics [3]. Due to the continuous upgrading of software, more and more functions such as word matching, alignment, automatic term extraction, bilingual/multilingual automatic balancing, annotation processing, random sampling and statistical analysis are being continuously developed [4]. In the study of corpus linguistics, the establishment of theoretical models for the analysis of grammar, semantics, vocabulary and statistics using the machine corpus, the development of practical tools, and the exploration of analytical methods are the basic disciplines of corpus linguistics research. It plays a leading role in promoting the deep development of corpus linguistics [5].

As long as learners pay attention to the content of language input, the acquisition of language forms will naturally occur in the subconscious; the learning system is conscious [6]. However, the function of this system is limited, and it can only monitor the acquisition system. Kernel function is needed to calculate the distance between relationships. The training and prediction speed of this method is too slow. It is not suiTable for processing large amounts of data. So most researchers who extract entity relations adopt machine learning method based on eigenvector. This method needs to construct training data in the form of eigenvector [7]. Corpus technology is especially useful. For example, you can query the various situations in which the collective nouns are consistently used. On the other hand, the survey also shows that many methods that are not suiTable for this inductive method, and that this method of learning may not be suiTable for all types [8]. A corpus-based lexicographical change has changed this model, using real corpus as an example. Therefore, the use

  DOI: 10.25236/aisallc.2019.043

of real corpus as an example sentence plays an important role in corpus-based learner lexicography [9]. It is exerting more and more influence on many fields of language research. It can be seen that corpus linguistics plays an extremely important role in the process of language research. Especially the study of second language acquisition is more important [10].

2. The Application of Corpus Linguistics in Second Language Acquisition

The corpus approach provides a huge source of raw language materials for empirical linguistic research, which makes inductive research more strongly supported by linguistic facts. In terms of collocation of words. Collocation retrieval and statistics are the main basis for studying the change of word meaning, the pattern of phrase structure and the meaning of phrases. The combination of collocation statistics and index analysis provides a large number of contexts and facts for language observation. After a series of suspicions and criticisms, corpus linguistics is emerging. It will form a complementary rather than opposite relationship with the mainstream explanatory linguistic methods, emphasizing learners to learn independently, challenging textbook interpretation and present. The intuition of the language user, and self-discovering the law of language use in learning.

Corpus undoubtedly provides a huge and objective database for second language research. Most of the previous studies on interlanguage focus on error analysis. Now we can get a more comprehensive understanding of interlanguage through corpus. As a research tool and method, corpus can be used not only to test research hypotheses, but also to generate new hypotheses. The corpus is drawn not only from the daily conversations of native speakers, but also from a variety of academic literature. The principle of writing is also based entirely on the linguistic data of this large corpus, rather than the original concept and understanding of the second language grammar by the author himself. Only a statistical analysis of a large number of real language data can be used to find a significant co-occurrence of words and words. It can be found that these co-occurrences are not arbitrary combinations of words but units of meaning that were previously unknown. Comparing interlanguage with target language can help researchers determine the use or overuse of specific language features. A learner's interlanguage in a native language context is compared with the learner's interlanguage in another native language context, so that common features in the second language acquisition process can be found.

As far as the relationship between corpus and SLA research is concerned, corpus can make up for the lack of data sources in traditional SLA research. As second language acquisition is a psychological process, researchers can not directly observe and study this process. By using corpus and new achievements in linguistic research, the corpus reflects the true face of contemporary foreign languages. Dictionary classifies the most commonly used words and frequency of use, and arranges meanings, idioms, collocations and example sentences according to frequency of use. It gives a new look to the teaching dictionary. In other words, the learner's choice of subject will result in a style of persuasion or emphasis, which is very different from the dissertations of native speakers. Reasons why learners use too many types of topics, including the migration of native language structures and cultural conventions. The comparative corpus linguistics mainly considers the cross-language comparison between the learner's mother tongue and the target language. This cross-language comparative analysis can help us to analyze the interlanguage. One of the great advantages of corpus linguistics research is that it can present all the contexts of a word, enabling learners to identify the different meanings associated with the word.

The process of language acquisition is more important to the acquisition of language form than to the understanding of language content, because it is the former, not the latter, that really helps to improve the accuracy of language. The main idea of entity relation extraction based on ensemble learning feature vectors is to find all possible entity pairs in the text, and construct these entity pairs as candidate relation instances through entity and entity context. In order to take full account of the co-occurrence of features and categories, and to adjust the phenomenon that the amount of mutual information can not correctly reflect the expressive ability of feature items to text categories due to the uneven distribution of text among different categories in text set. The corpus is labeled for the types of exercises and can be used to provide vocabulary for a specific level of practice. The corpus

can also be used to investigate meta-language types in textbooks to study whether the use of terms is consistent. To create a corpus, the corpus design needs to meet the learner's requirements. Only in this way can corpus-based learning activities become an integral part of linguistics. Compare, analyze, and study the vocabulary, grammar, textual behavior patterns, linguistic error characteristics and causes, and learning strategies of learners in the process of second language acquisition. This is the result of the combination of corpus linguistics and second language acquisition.

3. The importance of building Learner Corpus

The method of ensemble learning is a popular algorithm in the field of machine learning in recent years to improve learning accuracy. ensemble learning trains multiple classifiers to solve the same problem. In the chain of information from sensory memory to short-term memory, and then from short-term memory to long-term memory, short-term memory is the key, and attention is the key to short-term memory, which can be said to be the key. To describe the relevance of the text. The focus of the study is to establish collocation contours and collocations of keywords, and to compare specific patterns of a text with reference corpus. In the corpus that has been grammatically annotated, the combination of various part-of-speech tokens and the use of various types of sentence patterns are quantitatively analyzed. This is where corpus linguistics is superior to other linguistic theories. It is also unmatched by other linguistic theories in second language acquisition. The existing corpus is still too concentrated on written corpus, and the degree of automation of existing data analysis software can not meet the needs of researchers. Nonetheless, the corpus provides an unprecedented accurate description of the second language corpus, which can help researchers discover more facts.

There are still some limitations in the study of tagged corpus, because although there are better tagging software now, there is not a handy analysis software tool yet. However, the advantages of using marked corpus for grammatical and textual research are obvious compared with using unmarked corpus. There are enough language learners through contacting examples of a word used in multiple contexts. They can explore the implied meaning, semantic prosody and collocation of a word through self-induction and summary, so as to improve their language awareness and self-learning ability. The corpus will be larger, better and better, and more and more popular. Corpus linguistics has not only created a new linguistic research method, but also opened up a new field of language subject. The church uses its corpus to conduct its own learning and research. Once you have mastered the necessary corpus knowledge and skills, language learning activities will become learner-centric. “Using a corpus to teach“ means using a corpus-based approach to teaching topics related to language or linguistics. It means that the learner is not only the producer of the corpus data, but also the user. By collecting the data as part of normal activities, you can analyze the words you produce and help develop the learner's ability to learn independently.

The essence of linguistic process is to adopt effective methods to make learners'sensory memory enter into short-term memory, and then from short-term memory into long-term memory. To improve the performance of text categorization by ensemble learning, it is necessary to establish an effective feature evaluation function to select feature items reasonably. Here, the feature selection method of ensemble learning mutual information method is used to select the features of the class string. Second language acquisition based on the achievements of Phrasology can improve the authenticity of language output. Phrases play an important role in daily communication. It has a great influence on the core elements of communication such as the fluency of language output, and separates vocabulary and grammar. It is considered that grammar is the subject of language, vocabulary is the material that fills the grammatical structure, and the meaning of utterance is expressed through grammar. Corpus linguistics emphasizes the importance of lexicology. The learner corpus of different backgrounds is analyzed to reveal the law of second language acquisition in different backgrounds, to distinguish the affected language features in interlanguage and the stage characteristics in the process of foreign language acquisition and its significance to linguistics.

From the perspective of strict second language analysis. In fact, even the literature on second

language acquisition which is not based on corpus shows that this so-called "contrastive misunderstanding" exists widely in a hidden form. Corpus research can also be used to improve foreign language grammar teaching, compile dictionaries for assisted learning, compile writing guides and develop computer-aided software. The primary task of ensemble learning is to identify entities first, which requires labeling named entities. At present, we only consider the relationship between two entities in a sentence. Regardless of the relationship between the entities crossing the sentence, the different development stages of the interlanguage are distinguished in the linguistic process; the learner's spoken corpus is compared with the written corpus. To help learners improve their linguistic awareness by discovering the differences between learners' spoken and written language. The language data it provides is not only rich, but also real and reliable. In addition, the use of computer technology to compare and analyze the corpus, which increases the accuracy and credibility of the research. The processing of linguistic content precedes the processing of linguistic forms. The learner's brain cognition mechanism first processes the components that express the meaning of the input linguistic data, and secondly the components that express the grammatical function.

4. Conclusion

This paper analyses corpus linguistics based on ensemble learning and its role in second language acquisition. The second language knowledge satisfies the communicative needs and tests the newly acquired language forms. The process of language output is also a process of language experimentation, which enables them to know which second language knowledge they have learned is available. Corpus provides a large number of authentic corpus for language learning. Data-driven learning mode helps to cultivate the ability of autonomous learning. By observing corpus, learners analyze the syntactic features and typical collocations of words. This discovery learning mode helps to improve the enthusiasm of learning. The development of the corpus will be greatly accelerated. More and better features of the corpus will also be developed in practice. It will also fundamentally change all aspects of linguistics, including "teaching" and "how to teach." We should not only think of the corpus as a resource that helps to decide what to teach, but also as a resource that can be learned directly from it. In short, corpus linguistics has studied the study of two-language acquisition.

References

[1] Egbert, Jesse. Corpus linguistics and language testing: Navigating uncharted waters[J]. Language Testing, 2017, 34(4):555-564.

[2] Crossley S,Salsbury T,Titak A,et al. Frequency effects and second language lexical acquisition: Word types, word tokens, and word production[J]. International Journal of Corpus Linguistics, 2014, 19(3):301-332.

[3] Xie, Qin. Recent developments in corpus linguistics and corpus-based research / Department of Linguistics and Modern Language Studies at the Hong Kong Institute of Education[J]. Language Teaching, 2015, 48(01):156-160.

[4] Tang C,Rundblad G.When safe means "dangerous": a corpus investigation of risk communication in the media[J]. Applied Linguistics, 2017, 38:págs. 666-687.

[5] Durrant P.Corpus frequency and second language learners' knowledge of collocations[J]. International Journal of Corpus Linguistics, 2014, 19(19):págs. 443-477.

[6] Crossley S,Kyle K,Salsbury T.A Usage-Based Investigation of L2 Lexical Acquisition: The Role of Input and Output[J]. The Modern Language Journal, 2016, 100(3):702-715.

[7] Chen M H,Huang S T,Chang J S,et al. Developing a corpus-based paraphrase tool to improve EFL learners\" writing skills[J]. Computer Assisted Language Learning, 2015, 28(1):22-40.

[8] Kreyer R.”Funky fresh dressed to impress“: A corpus-linguistic view on gender roles in pop songs [J]. International Journal of Corpus Linguistics, 2015, 20(2):págs. 174-204.

[9] Tseng J J,Lien Y J,Chen H J.Using a teacher support group to develop teacher knowledge of Mandarin teaching via web conferencing technology[J]. Computer Assisted Language Learning, 2016, 29(1):127-147.

[10] Saito K,Shintani N.Do Native Speakers of North American and Singapore English Differentially Perceive Comprehensibility in Second Language Speech?[J]. TESOL Quarterly, 2015, 50(2):421-446.