Chinese Learner Corpora and Second Language Research

WEN Qiufang
The National Research Center for Foreign Language Education
Beijing Foreign Studies University

Abstract

The paper starts with a general description of the three published Chinese learner corpora, followed by a survey of the corpus-based studies of Chinese learner English with a focus on two of my own studies on vocabulary. Finally, the advantages and disadvantages of the corpus-based studies are discussed.

Up to now there are three published corpora of Chinese Learner English available in the market, that is, Chinese Learner English Corpus (CLEC), Spoken and Written English Corpus of Chinese Learners (SWECCL) and College Learners’ Spoken English Corpus (CLOSEC), which have given fresh impetus to corpus-based studies on Chinese Learner English. The corpus-based studies on learner English in China have analyzed various aspects of the learner’s interlanguage, such as phonology, morphology, word forms, word combinations, syntax and discourse features, etc. which provide insights to a better understanding of the interlanguage of the Chinese learners.

The recent corpus-based studies carried out by myself are on the English vocabulary of Chinese learners. The first one is on the use of top twenty frequency adverbs (TTFAs) in English that attempted to investigate how Chinese tertiary-level English majors use frequency adverbs in their written and spoken English in order to provide empirical evidence for answering two questions: (1) whether English learners tend to overuse or/and underuse high-frequency words; (2) whether they tend to use oral discourse features in their written English or/and use written discourse features in their spoken English. The data used in this study included a written English corpus by English majors from the CLEC, and a spoken English corpus from the SWECCL. With these two corpora combined, the general learner English corpus contained 955,043 words. The reference corpus was the British National Corpus (BNC), which contains 90 million words of written English and 10 million words of spoken English, all by native speakers. Data analysis included a series of comparisons regarding the use of top twenty frequency adverbs (TTFAs), such as the comparison between the general learner corpus and the BNC, the comparison between the learners’ spoken corpus and the spoken portion of the BNC, the comparison between the learners’ written corpus and the written portion of the BNC. The results showed that the overuse tendency and the underuse tendency co-existed in the use of frequency adverbs by Chinese EFL learners but the former was much stronger than the latter both in their speech and writing. Furthermore, the use of frequency adverbs in speech by Chinese EFL learners deviates more than that in writing from native speakers’ use. It also showed that some written-register sensitive frequency adverbs were overused in the learner’s spoken corpus while some spoken-register sensitive frequency adverbs were overused in the learner’s written corpus.

The second study is on the developmental features of speaking vocabulary by English
majors in China, which was then compared with the native speakers’ performance. The English learners involved in the study were 56 students enrolled in a university in 2001. They were asked to complete an oral task by producing a three-minute monologue after three minutes’ preparation in a language lab in 2001, 2002, 2003 and 2004. The native speakers were 15 American college students requested to perform the same task as our learners. The developmental changes were measured in terms of three indexes: fluency, word variation and lexical frequency breadth. The performance on the three indexes of the learners across four years and the native speakers form a continuum. However, Year Four students were significantly lower in fluency and word variations than native speakers, but similar to the native speakers in lexical frequency breadth. The interval between Year Two and Year Three saw the most noticeable progress. The learners with different starting levels of the three indexes varied in terms of their changes. The low-level group made greater progress than the middle-level group that produced slightly faster progress than the high-level group.

The corpus-based studies examine the learner’s performance that show what learners can do rather than what they cannot do. The corpus is authentic, representative and retrievable. However, such studies cannot reveal what the learners tend to avoid and what learners are not able to do. Therefore, the analysis of the learner’s performance must be complemented by the analysis of the learner’s competence if we intend to get a comprehensive picture of the features of an interlanguage.

Corpora References

