Abs-61

Hao-Jan Howard Chen

Constructing a Chinese as Second Language Learner Corpus and a Web-based Concordancer

Many researchers and language teachers believe that language corpora have great potentials for language learning and teaching. The learner corpora in particular have received much attention recently. Currently, several English learner corpora are available; however, learner corpora for many other languages are not easy to obtain. Recently, because of the rapid economic growth in China, an increasing number of students are learning Chinese as a second language. Although the number of CSL learners is increasing, very few CSL learner corpora are available for teaching and research. For CSL research, the learner corpus can play an important role. Researchers can conduct research on learners¡¦ interlanguage development, language assessment, and language pedagogy. In addition, the research findings from the learner corpus can also be used in developing Chinese teaching materials.

This paper will introduce a new Chinese as second language learner corpus and related corpus search tools developed by MTC (Mandarin Teaching Center) and SC-TOP (Steering Committee of Test of Proficiency) in Taiwan. MTC is located at National Taiwan Normal University and it is the largest Chinese learning centers in Taiwan. There are more than 1700 students enrolled in each quarter, and there are more than 200 teachers in this center. Students from more than 70 countries are studying in this center. SC-TOP is a research center sponsored by Ministry of Education for developing various Chinese as a second language tests. Based on the data provided by these two centers, a 3-million-word Chinese as a second language learner corpus has been developed. The leaner corpus includes the following three different types of learner data-

1.CSL learners¡¦ short essays written in various TOP tests.

2.CSL students¡¦ writing assignments at MTC

3.CSL students¡¦ writing in the MTC achievement tests at each proficiency level

The learner corpus was further automatically tagged with a Chinese tagger called CKIP (Chinese Knowledge Information Processing) tagger developed by Academia Sinica, Taiwan. The POS-tagged CSL corpus is very useful for research and teaching. In addition to the learner corpus, a web concordancer which has several different search options was also developed. This web concordancer allows users retrieve specific words and phrases from CSL learner corpus. Thus, various CSL learners¡¦ errors can be retrieved and studied more easily and systematically. Furthermore, the POS-tagged learner corpus can be used to search for collocates used by learners. When more data are collected, users of the web-based system can also find errors and patterns produced by CSL learners from different native language backgrounds. The availability of this CSL learner corpus and the web concordancer should be able to help more researchers uncover CSL interlanguage patterns. Moreover, many teachers and students can use the learner corpus to enhance their teaching and learning.