Abs-96

Yukio Tono (Tokyo University of Foreign Studies, JAPAN)

Identifying new verb co-occurrence patterns as criterial features: Using ICCI and JEFLL

This paper aims to present an interim report on the analysis of younger learners' interlanguage development in English using the International Corpus of Crosslinguistic Interlanguage (ICCI). ICCI is a government-funded 5-year project of compiling corpora of English writings by primary and secondary school students in eight different regions (Japan, Spain, Israel, Austria, Poland, Hong Kong, Taiwan, and China), initiated by the author at Tokyo University of Foreign Studies.

The ICCI project focuses on data collection from younger learners of English. Most learner corpora available so far mainly cover intermediate to advanced learners at upper-secondary to university levels. In order to investigate the acquisition processes in its entirety, it is necessary to gather data at learning stages much earlier than that. To this end, we collected beginning-stage learners' data, by administering 20-minute in-class essay tasks without the use of a dictionary to primary and secondary school students. Another unique feature of ICCI is its comparability with the JEFLL Corpus (Tono 2007), a corpus of Japanese EFL learners' writings, covering 10,000 students from Year 7 to 12 (English is introduced in Year 7 in Japan). Together with JEFLL and ICCI, the data consists of more than 17,000 students ranging over 6-8 years of English learning at the beginning stage (the total size of ICCI and JEFLL is well over 1.5 million words).

In this paper, the overall design of the ICCI project will be described and an interim report on the research into the new verb co-occurrence patterns at early stages of learning will be presented as an example of the research using ICCI and JEFLL. The new verb co-occurrence patterns are said to serve as "criterial features," linguistic features distinguishing proficiency levels of learners in terms of the Common European Framework of Reference (CEFR). Preliminary findings using Cambridge Learner Corpus (CLC) show that early stages of features, especially A1 and A2 are hard to identify, due to the lack of data in CLC. In this study, we first assigned CEFR levels to sampled compositions (200 samples from each country for each CEFR level, totaling approximately 1,000 samples for A1, A2 and B1 levels respectively). Then all the data were syntactically parsed, and verb co-occurrence patterns were extracted using a tgrep-like search engine. The results were statistically analysed in order to find significant differences in frequencies between different CEFR levels. The results showed how ICCI and JEFLL could fill the gap in identifying criterial features for beginning-level learners.

References

Tono, Y. (ed.) (2007) JEFLL Corpus: A Corpus of 10,000 Japanese EFL Learners. Tokyo: Shogakukan.