Abs-25

Lan-fen Huang (University of Birmingham)

The selection of corpora for the investigation of discourse markers in learner English

The availability of learner corpora has attracted research attention to the area of comparative studies between learners and native speakers (NSs) and between learners with different first language backgrounds (e.g. studies in Granger (ed.) (1998), Granger, Hung and Petch-Tyson (eds.) (2002) and Meunier and Granger (eds.) (2008)). Despite these demonstrations of the increasing interest and new approaches to learner language, there is little discussion about the comparability of corpora and difficulties in comparing corpora. This paper will seek to address some questions raised in comparative studies of learner corpora as well as the selection of corpora for the investigation of discourse markers.

Since learners in the environment of English as a foreign language do not talk in English in everyday life, it is very difficult to collect naturally-occurring speech produced by learners. Therefore, most learner corpora of spoken English consist of contrived data, which are elicited in a rather restricted context. When a learner corpus is used for comparative studies of learners' and NSs' speech, it is difficult to obtain a truly comparable NS corpus. Although it seemed appropriate to recruit NSs to do the same tasks as learners had done for the compilation of a corpus, whether or not NSs are trained to take an oral exam in their first language and whether or not the context is properly duplicated are open to doubt. If the so-called comparable NS corpus can be carefully designed and compiled, this raises another question about the 'un-naturalness' of the elicited NS speech.

The Spoken English Corpus of Chinese Learners (SECCL) is used to investigate discourse markers. I set out some arguments in this paper against compiling a comparable NS corpus. It is difficult to ensure comparable conditions with respect to such factors as exam-oriented and it is also challenging to compile a corpus of NSs' speech with similar size and number of participants. The compromise that is chosen is comparing the uses of discourse markers in the learners' spoken English in SECCL with those in the NS speech in MICASE and ICE-GB. Also, I argue against using the terms 'underuse' and 'overuse', which seem to assume NSs' use of discourse markers is the target norm for learners as well as suggesting learners' lack of competence. Instead, the neutral terms 'under-represent' and 'over-represent' are used.

This paper will discuss if I can legitimately compare the use of discourse markers in SECCL, MICASE and ICE-GB and how I overcome the problem of the issue of comparability of corpora as well as the difficulties in investigating discourse markers across three different corpora.

References

Granger, S. (ed.) 1998. *Learner English on Computer*. London: Longman.

Granger, S. 2002. 'A bird's-eye view of learner corpus research' in Granger, S., J. Hung and S. Petch-Tyson (eds.). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Meunier, F. and S. Granger. (eds.) 2008. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins.