Extracting English Verb-Noun Miscollocations from NNS Academic Writing based on Corpus Comparison

Hao-Jan Howard Chen and Fiona Yu-chen Cheng (National Taiwan Normal University, Taiwan)

There have been more and more studies showing that collocation played a significant role in second language comprehension and production (Ellis, 1996; Lewis, 2000; Lien, 2003). However, collocations were often found to be problematic for many second language learners. Previous studies have consistently revealed that EFL learners had insufficient knowledge of English collocations (Bahns & Eldaw, 1993; Gitsaki, 1997; Liu, 1999; Chen, 2008). Laufer and Waldman (2011) found that L2 writers produced far fewer verb-noun collocations than native speakers. Moreover, the number of verb-noun collocations errors, particularly interlingual ones, continued to persist even at advanced levels of proficiency.

While verb-noun collocations "form the communicative core of utterances where the most important information is placed" (Altenberg, 1993, p. 227), researchers indicated that the most frequently occurred miscollocations were verb-noun miscolloations, and they were particularly difficult for second language learners (Liu, 2002; Nesselhauf, 2005).

Empirical studies have offered clear evidences that EFL learners lacked verb-noun collocational knowledge and made suggestions for explicit instructions on collocations. According to Woodlard (2000) and Lewis (2000), helping students to observe and notice their own miscollocations would enhance students' awareness of acceptable collocations. Therefore, investigating learners' miscollocations can help teachers and researchers to better understand learners' general patterns of collocational errors, thus shedding lights on what collocations to teach and how to help language learners.

In most existing studies on miscollocations, researchers in these studies collected L2 learners' compositions, assignments, or examination essays, and manually marked these miscollocations. There are several studies targeting on L2 learners' miscollocations through the use of learner corpora (Nesselhauf, 2003, 2005; Shih, 2000; Liu, 2002; Chang & Yang, 2009). One of the most comprehensive studies is the verb-noun collocation study carried out by Nesselhauf (2005). She investigated the use of verb-noun collocations produced by advanced German learners of English based on International Corpus of Learner English (ICLE). Nesselhauf manually extracted and analyzed the verb-noun combinations in the 318 essays selected from the sub-corpus (GeCLEE) which contained around 150,000 words. The study was able to retrieve 507 V-N miscollocations produced by German EFL learners. It is evident that Nesselhauf spent a great amount of time in searching through the learner corpus and provided useful information about learners' miscollocations. It is, however, very labor-intensive and time-consuming to examine second/foreign language learners' miscollocations through manual error-tagging. However, several much larger learner corpora are now available. These include corpora such as ICLE (International Corpus of Learner English), International Corpus Network of Asian Learners of English (ICNALE), The TECCL corpus (Ten-thousand English Compositions of Chinese Learners) and TOEFL11: A Corpus of Non-Native English. The manual error tagging and error analysis methods might not be able to process these larger learner corpora. Second language researchers need to adopt some new approach to analyze these learner corpora and to uncover various miscollocations more efficiently.

In this paper, we will introduce a useful corpus research tool called Sketch–Diff, a tool which was provided in Sketch Engine (SKE) developed by Adam Kilgarriff and his associates. This powerful tool can help language researchers to compare various

collocations used in a native corpus and a learner corpus. In this study, we first uploaded the following two corpora onto SKE: one native reference corpus and one NNS academic writing corpus. The native corpus includes the academic sub-corpora taken from BNC (British National Corpus) and COCA (Corpus of Contemporary American English) academic corpus (about 100-million words). The NNS academic writing corpus is a large collection of 494 Taiwanese masters' theses in TESOL from 10 different graduate programs in Taiwan (10-million words). After these two corpora were uploaded into SKE, the SKE can automatically tag and analyze these corpora. After the pre-processing stage, the Sketch-Diff tool can then be used to compare all the collocations used by native speakers and non-native speakers. When the researcher input any target word, he/she can then find all collocates of the target word used in the uploaded native corpus and non-native corpus. Various collocation patterns can be extracted and the significant differences between the usage of native speakers and non-native speakers can be revealed. The system also used different colors to show the users about the significant differences between native and nonnative usage.

With the help of SKE tools, 171 types (1171 tokens) of miscollocations were discovered in this study. All miscollocations were categorized into four main major types: misused verb, misused noun, misused or missing preposition, and nonexistent combination. 142 types (946 tokens) of miscollocations were categorized as misused verb; 16 types (124 token) were misused noun; 6 types (36 tokens) were misused or missing preposition and 7 types (75 tokens) of miscollocations were non-existent. Moreover, many of these collocation errors were attributed to L1 influences. 95 types (608 tokens) of miscollocations were possibly influenced by learners' native language. It seems clear that first language plays a crucial role in second language learners' production of collocations.

Based on our empirical tests, the Sketch-Diff can be an effective tool in comparing collocations used in English native and non-native corpora. With this new corpus comparison tool, researchers do not need to painstakingly examine each and every line of students' essays and manually mark each and every collocation error. This new corpus processing tool provides a more convenient and thorough way of uncovering the possible differences between native and non-native's collocational competence. At the current stage, the corpus tool is surely not fully automatic in uncovering all the collocation errors made by L2 learners. Researchers still need to further verify the differences recommended by SKE. However, the tool has already made some contributions to studies on L2 acquisition of collocations.

References

Altenberg, B. (1993). Recurrent verb-complement construction in the London-Lund Corpus. In J. Aarts, P. Haan, & N. Oostdijk (Eds.), *English language corpora: Design, analysis, and exploration: Paper from the thirteenth International Conference in English Language Research on Computerized Corpora, Nijmegen 1992.* (pp. 227-245). Amsterdam: Rodopi.

Bahns, J. & Eldaw, M. (1993). Should we teach EFL students collocations? *System* (21), 101-114.

Chen, M. H. (2008). *A study on the English collocation competence of college students in Taiwan*. Unpublished master's thesis. National Taiwan University of Science Technology, Taiwan, R.O.C.

Chen, P. C. (2002). *A corpus-based study of the collocational errors in the writings of the EFL learners in Taiwan*. Unpublished master's thesis. National Taiwan Normal University, Taiwan, R.O.C.

Laufer, B. & Waldman, T. (2011). Verb- noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning 61*(2), 647-672.

Lewis, M. (1997). *Implementing the lexical approach*. London: Language Teaching Publications.

Li, C. & Thompson, S. A. (2005). *Mandarin Chinese: A functional reference grammar*. Berkeley: University of California Press.

Lien, H. Y. (2003). *The effects if collocations instruction on the reading comprehension of Taiwanese college students*. Unpublished doctoral dissertation. Indiana University of Pennsylvania, United States.

Liu, C. P. (1999). A study of Chinese Culture University freshman's collocational competence: "Knowledge" as an example. *Hwa Kang Journal of English language & Literature 5*, 81-99.

Liu, C. P. (1999). An analysis of collocational errors in EFL writings. *Proceedings of the Eighth International Symposium on English Teaching* (pp. 483-494). Taipei: Crane.

Liu, L. E. (2002). *A corpus-based semantic investigation of verb-noun miscollocations in Taiwan learners' English*. Unpublished master's thesis. Tamkang University, Taiwan, R.O.C.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics 24*, 223-242.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Wang, C. J. (2001). *A study of the English collocational competence of English majors in Taiwan*. Unpublished master's thesis. Fu Jen Catholic University, Taiwan, R.O.C.

Woolard, G. (2000). Collocation: Encouraging learner independence. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 28-46). London: Language Teaching Publication.

Wu, W. S. (1996). Lexical collocations: One way to make passive vocabulary active. *Proceedings of the eleventh conference on English teaching and learning in the Republic of China* (pp. 461-480). Taipei: Crane.