Collocation Deficiency in a Learner Corpus of English: from an overuse perspective

Rebecca Hsue-Hueh Shih
Department of Foreign Languages and Literature
National Sun Yat-sen University, Kaohsiung,Taiwan, R.O.C.
E-mail: hsuehueh@mail.nsysu.edu.tw

Abstract

Collocational deficiency is a pervasive phenomenon in learner English. Language learners often fail to choose the correct combination of two or more words due to their unawareness of collocational properties in vocabulary. They are apt to adopt lexical simplification strategies such as using a synonymous or L1-influenced expression. This paper presents a corpus-based study on the collocational deficiency of Taiwanese learners of English. The work utilizes two pre-tagged corpora, Taiwanese Learner Corpus of English and British National Corpus, to examine the learner's use of collocations over a set of synonymous words: *big*, *large*, *great*.

The experimental findings indicate that among the three words the collocations with *big* are significantly overused by the learners when it is used to refer to abstract concepts. This overuse phenomenon is further investigated and it is found that the collocations of high frequency in the learner English tend to be used to express vague ideas when more specific meanings should be conveyed. It is also found that the learners are apt to apply those collocations to the cases where more concise expressions are preferred. Another finding shows that problematic collocations, pertaining to *big*, *large* and *great*, are produced as the result of learner's application of the L1-transfer and synonym strategies, which the Taiwanese learners commonly adopt for lexical simplification.

1. Introduction

Corpus Linguistics studies language features based on large databases of authentic language samples stored on computer. Because its automated quantitative analysis provides novel and refreshing insights into real language use [1], the corpus-based approach has rapidly spread into many language-related research. SLA (Second Language Acquisition) and EFL (English as Foreign Language) specialists, with no exception, consider large databases of learner English a useful resource for them to gain concrete evidence and a wider perspective on learners' inter-language acquired during the process of language learning. Therefore,

computer learner corpora (CLCs) with various mother tongue backgrounds have been subsequently constructed [2].

With gradual availability of CLC and the awareness of its potential, a wide range of CLC-research starts to boom. The research often involves comparisons between inter-language that learners possess and native language on various linguistic features. For instance, the frequency distributions of most commonly-used words in a native and seven eastern European learner corpora are compared on different parts-of-speech categories [3]; the use of complement clauses in four learner corpora as contrasted with their native counterparts [4] is studied; the use of adverbial connectors by Swedish learners in comparison with the natives' is examined [5]. This kind of cross-language approach helps SLA and EFL specialists find out what linguistic features the language learners are apt to overuse/underuse, what are the particular areas of language behavior shared by learners from different backgrounds, and to what extent these phenomena appear in learner English. The quantitative information as such guides the researchers to carry out insightful qualitative analysis.

While the CLC-research focuses on the core aspects of learners' lexis, grammar and discourse, collocation deficiency, a pervasive phenomenon in every learner corpus, remains intact. Collocation is the habitual co-occurrence of two or more words in a text, and is an important feature for vocabulary learning. However, due to the traditional grammar-based EFL pedagogy, the collocational property in relation to each item of vocabulary has been neglected in EFL class [6] [7]. When learners encounter a collocation problem, they tend to resort to one of the strategies of lexical simplification: synonym, avoidance, transfer and paraphrasing [8]. Table 1 lists the examples of the four strategies used by Taiwanese learners of English (* is used to indicate collocational errors):

Problematic Collocations	**Correct Collocations**	**Strategies Applied**
* rules are loose	Rules are lenient	Synonym
* great drinker	Heavy drinker	Avoidance
* age layers	Age groups	Transfer
Inconvenience in moving	Transport difficulty	Paraphrasing

Table 1: Examples of lexical simplification by Taiwanese Learners of English

Apart from paraphrasing which is considered a good strategy in L2 (second language) communication, the other three uses result in unacceptable collocational mistakes in language learning. The most commonly used strategy, synonym taking up 38% of total colloational errors [6], can be viewed as a direct consequence of the unawareness of collocational restrictions between lexical items. Avoidance strategy is adopted when learners avoid a correct lexical item in favor of another, and thus alters the meaning of collocation. The use of transfer creates L1-influenced collocational errors and is the result of learners' working

hypothesis that there is one to one correspondence between L1 and L2 [9].

Since learners tend to choose whatever word is easily and readily retrievable in their minds when putting the strategies in practice, certain commonly-used synonymous words are apt to be overused in learner English. The work in this paper is designed to examine the collocational deficiency of Taiwanese learners of English from an overuse perspective, and find out in what context the problematic collocations occur. The work investigates this issue over a set of synonymous words, *big*, *large* and *great*. It uses two pre-tagged corpora, Taiwanese Learner corpus of English (TLCE) and British National Corpus (BNC), which will be stated subsequently in Section 2.1. The computer-assisted tools for tagging and lemmatizing of the corpora and for quantitative analysis will be described in Section 2.2. A series of experiments and the results are shown and discussed in Section 3. Concluding remarks are made in Section 4.

2. Methodology

2.1 Corpora: TLCE and BNC

As stated in the introduction, CLC-research often compares non-native data with native data in order to reveal the overuse and/or underuse phenomena in a learner corpus. In this work, the Taiwanese Learner Corpus of English (TLCE) of 286,600 words is under investigation and the British National Corpus (BNC) of 100 millions words is used for comparison. TLCE is a growing corpus of English compositions and weekly journals written mainly by college English majors in Taiwan. They are freshmen, sophomores and juniors of age ranging from 19 to 22. The current data are from Sun Yat-sen and Chi-nan universities, and more data from other universities will be collected in the next couple of years to make the corpus more representative. The BNC contains modern British English and is a unique collaboration between three major U.K. dictionary publishers, two universities, and the British Library [10]. The work here utilizes mainly its subset of 1 million words (from BNC Sampler written text), but its complete set was consulted in the situation where more data is needed as in Section 3.4 and 3.4 .

2.2 Analysis Tools: TOSCA and CCS

The corpora are lemmatized and part-of-speech tagged with the TOSCA tagger [11]. TOSCA is a stochastic tagger, supplemented with a rule-based component which tries to correct observed systematic errors of the statistical components. TOSCA also gives each word form its lemma (basic form). For instance, word forms such as *takes*, *took*, *taken*, and *taking* have the same lemma *take*. This function facilitates the collocation analysis under the same lemma. TOSCA operates with a lexicon, which currently contains about 160,000 lemma-tag

pairs, covering about 90,000 lemmas. The TOSCA-ICLE tagset contains 270 different tags within 16 major word classes. For simplicity, only the major word classes are considered in the current study.

Corpus analysis tools such as WordSmith [12] and Qwick [13] are very popular and useful software for concordance and simple collocation search. However, they only take raw text as input, and thus fail to perform more sophisticated functions, such as the search of collocations in terms of their lemmas and parts-of-speech. To facilitate analysis required in this work, software for sophisticated collocation searching, Corpus Collocation Searcher (CCS), is specially developed. CCS takes TOSCA tagged data as input and enables users to enter either a word form, lemma or even part-of-speech as a search keyword. It provides the same mechanism for collocate specification. For instance, users are able to search the noun collocates immediately following the keywords *great*, *greater* and *greatest* by specifying the lemma form of the keyword, *great*, the part-of-speech of the collocates, NOUN, and the location of the collocates.

3. Experimental Results

3.1 Frequency Distribution of the Synonymous Words

The frequencies of *big*, *large* and *great* are calculated from both of the corpora. Figure 1 indicates the frequencies (per million words) for each of these synonymous words in TLCE and BNC. As shown in the figure, *big* and *large* both show a considerable discrepancy in the number of their occurrences between the two corpora, while *great* doesn't. The frequency of *big* in the learner corpus is almost double the number of occurrences in its native counterpart, whereas *large* in TLCE appears only one fifth of the number of occurrences in BNC.

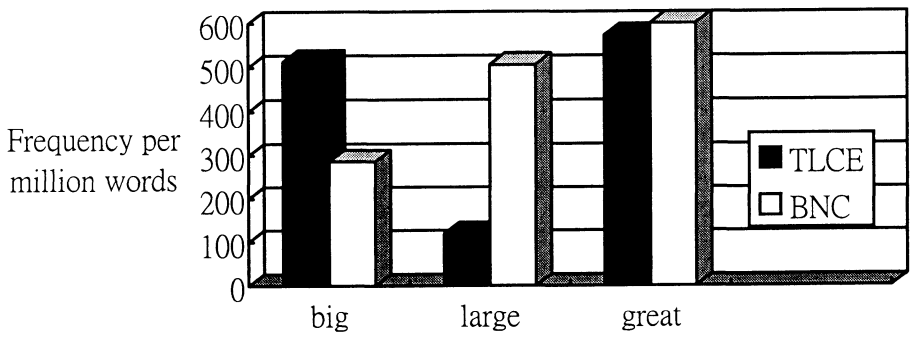

Figure 1: Frequency Distribution of the words concerned

As stated in the introduction, certain commonly-used simple words are apt to be overused by language learners in the course of lexical simplification (synonym, transfer and avoidance) due to their easy retrievability. To further investigate the phenomenon of overuse in relation to learner collocation deficiency, it is necessary to examine in great detail the collocational property of *big* in the learner corpus. As to the underuse of *large*, it is not in the focus of the present study.

Like many other adjectives, *big* has both attributive (*big* + noun) and predicative (noun + be + *big*) functions. Since the attributive *big* (here assumed to be immediately followed by a noun) takes up majority of the occurrences (nearly 70%) in the learner corpus, and the CCS tool so far doesn't provide a function to locate the subject noun of the predicative *big*, the following experiments are based on its attributive function only.

3.2 Abstract vs. Concrete Collocated Nouns

The study in [1] showed that the vast majority of occurrences of *big* in native corpora are used to refer to physical size of objects. Thus, this experiment was carried out to examine the properties of nouns that *big* collocates in TLCE. Figure 2 shows the ratio of abstract to concrete collocates in both corpora. As shown, nearly 70% of the occurrences of *big* in BNC are used to refer to concrete objects, but only 55% in TLCE. In other words, the Taiwanese learners use the word *big* much more often than native speakers do when describing abstract concepts. In contrast to 30% in BNC, the collocations of *big* and the referred abstract concepts take up 45% of total number of the occurrences in TLCE. As the use of *big* to refer to physical size is less problematic, next experiment are carried out to examine the use of its abstract noun collocates.

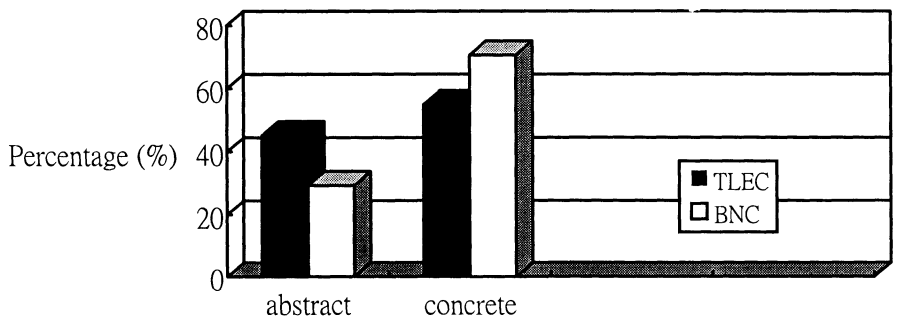

Figure 2: Distribution of Abstract and Concrete Nouns which collocate with 'big'

3.3 Distribution of Abstract Noun Collocates

To further study the abstract collocates of *big*, it is desirable to find out what these

abstract nouns are and how the Taiwanese learners use them differently from native speakers. Three kinds of data are concerned here: (1) abstract nouns (N) which collocate with *big* more than once in TLCE, (2) freq(*big*, N)–the frequency of co-occurrences of *big* and N, and (3) the ratio of freq(*big*, N) to freq(N). Table 2 shows the comparisons of these figures in both TLCE and BNC. As some collocations of *big* in TLCE do not occur in the subset of BNC, a complete BNC of 100 million words is consulted for comparison.

Abstract Noun (N)	**TLCE 0.287 million**		**BNC One million**		**BNC 100 millions**	
	freq(*big*, N)	**ratio(%)**	**freq(*big*, N)**	**ratio(%)**	**freq(*big*, N)**	**ratio(%)**
problem	10	3.8	3	0.6	128	0.4
trouble	4	7.5	1	1.4	47	0.5
surprise	4	12	1	1.9	41	0.8
deal	3	16.7	0	0	158	0.9
burden	3	16.7	0	0	2	0.0003
pressure	2	2.0	0	0	2	0.0002
joke	2	6.9	0	0	11	0.005
turn	2	5.0	0	0	0	0

Table 2: Distribution of highly collocated nouns

As shown in the table, the collocation of *big problem* gives the highest co-occurrence frequency with *big* in TLCE, and the frequency ratio for the use of *big problem* over *problem is* 3.8%. However, the same collocation appearing in BNC of 1 million and BNC of 100 million words only shows the ratios of 0.6% and 0.4% respectively. This large discrepancy in the ratio of using certain collocation between learner and native English spreads across other high frequency collocates, except the collocation *big turn*, which never appears in the native corpus and is treated as a problematic collocation in the next experiment.

It is observed from learner's writing in TLCE that the learners overuse certain collocations to deliver vague ideas in the situations where more specific meanings are acquired; they also apply some collocations which are easily retrievable in their mind to the cases where more concise terms are usually expressed. Two examples given below explain these phenomena:

(a) *"...but I have a big, big, big problem, that is, that I don't have a camera... ... However, a camera is really very expensive. ..."*

(b) *"... it will be a big trouble to move all my things to another place. ..."*

As can be seen in (a), *"big problem"* conveys a vague meaning which is realized later in the writing as *"financial problem"*, and the three-word phrase, *"a big trouble"*, in (b) can be

replaced by a concise expression, " *very troublesome*".

3.4 **Problematic Collocations**

Some collocations in TLCE, pertaining to *big*, *large* or *great*, do not appear in BNC and are considered to be problematic. Table 3 lists those collocations in question, with misused adjective collocates in TLCE and common collocates in BNC.

Abstract Noun	**Misused Collocates in TLCE**	**Common Collocates in BNC**
turn	big	sharp
wind	big	strong
slam	big	loud
nature	big	-
regret	large	great, big
trouble	large	great, big
jealousy	great	intense, acute

Table 3. Misuse of Collocates in TLCE

The collocation errors listed in Table 3 are due to the strategies of lexical simplification that the Taiwanese learners put to use in their writing. When describing the first four abstract nouns, Taiwanese learner directly translate the concept of "big", which goes naturally with these nouns in their mother tongue language, into English. This results in the problematic collocations: **big turn*, **big wind*, **big slam* and **big nature*, rather than the correct use of *sharp turn*, *strong wind*, *loud slam* and *Nature* in BNC. The misuse of **large regret*, **large trouble* and **great jealousy* can be viewed as the application of the synonym strategy.

4. Conclusions and Further Work

This paper presents a corpus-based study on examining the collocational deficiency of Taiwanese learners of English from an overuse perspective. The experimental results give the followings findings. Firstly, among the synonymous word, *big*, *large* and *great*, the Taiwanese learners overuse *big* significantly when it is used to refer to abstract concepts. Secondly, the phenomenon that certain collocations with *big* in TCLE appear far more frequently than in BNC can be explained by the observations that the learners use *big* to convey vague ideas when more specific meanings should be expressed and that they are apt to apply easily retrievable collocations to the cases where more concise expressions are preferred. Finally, transfer and synonym are the main simplification strategies that the Taiwanese learners adopted when they encounter a collocational problem.

In this work, the predicate function of *big* was not examined due to the difficulty of locating the precedent subject using the current CCS tool. In the future, CCS will be augmented to facilitate the investigation on this part for a thorough study.

Acknowledgements

The author would like to thank her colleagues for collecting student's compositions and Dr. Ching-yuan Tsai for his valuable discussion. This work is supported by the National Science Council, Taiwan, R.O.C.

References

[1] Biber, D., S. Cornard, and R. Reppen, *Corpus Linguistics Investigating Language Structure and Use*. 1998: Cambridge University Press.

[2] Granger, S., ed. *Learner English on Computer*. . 1998, Addison Wesley Longman Limited.

[3] Lorenz, G., *Overstatement in advanced learners' writing: stylistic aspects of adjective intensification*, in *Learner English on Computer*, S. Granger, Editor. 1998, Addison Wesley Longman Limited. p. 53-66.

[4] Biber, D. and R. Reppen, *Comparing native and learner perspectives on English grammar: a study of complement clauses*, in *Learner English on Computer*, S. Granger, Editor. 1998, Addison Wesley Longman Limited. p. 145-158.

[5] Tapper, M., *The use of adverbial connectors in advanced Swedish learners' written English*, in *Learner English on Computer*, S. Granger, Editor. 1998, Addison Wesley Longman Limited. p. 80-93.

[6] Farghal, M. and H. Obiedat, *Collocations: a neglected variable in EFL.* International Review of Applied Linguistics in language Teaching (IRAL), 1995. **33**(4): p. 315-331.

[7] Bahns, J., *Should We Teach EFL Students Collocations.* System, 1993. **21**(1): p. 101-114.

[8] Blum, S. and E. Levenston, *Lexical simplification in second language acquisition.* Studies in second language acquisition, 1978. **2**(2): p. 43-64.

[9] Bahns, J., *Lexical collocations: A Contrastive View.* English Language Teaching Journal, 1993. **47**(1): p. 56-63.

[10] Aston, G. and L. Burnard, *The BNC Handbook*. 1998: Edinburgh University Press.

[11] Aarts, J., H. Barkema, and N. Oostdijk, *The TOSCA-ICLE Tagset Software and Tagging Manual*, . 1997, The Department of Language and Speech, University of Nijmegen, The Netherlands.

[12] Scott, M., *WordSmith Tools*, . 1997, Oxford University Press.

[13] Sinclair, J., *et al.*, *Language Independent Statistical Software for Corpus Exploration.* Computers and the Humanities, 1998. **31**(3): p. 229-255.