Advances in Language and Literary Studies
ISSN: 2203-4714
Vol. 5 No. 1; February 2014

Methodological Flaws in Corpus-Based Studies on Malaysian ESL Textbooks

Abdolvahed Zarifi (Corresponding author)
Department of Language and Humanities Education, Faculty of Educational Studies
Universiti Putra Malaysia, 43400 UPM SERDANG, Selangor, Malaysia
E-mail: vahed_zarifi@yahoo.com

Jayakaran Mukundan
Department of Language and Humanities Education, Faculty of Educational Studies
Universiti Putra Malaysia, 43400 UPM SERDANG, Selangor, Malaysia
E-mail: jaya@educ.upm.edu.my

Seyed Ali Rezvani Kalajahi
Department of Language and Humanities Education, Faculty of Educational Studies
Universiti Putra Malaysia, 43400 UPM SERDANG, Selangor, Malaysia
E-mail: ali.rezvani85@gmail.com

Doi:10.7575/aiac.alls.v.5n.1p.35
URL: http://dx.doi.org/10.7575/aiac.alls.v.5n.1p.35

Received: 04/01/2014
Accepted: 25/02/2014

Abstract

With the increasing interest among the pedagogy researchers in the use of corpus linguistics methodologies to study textbooks, there has emerged a similar enthusiasm among the materials developers to draw on empirical findings in the development of the state-of-the-art curricula and syllabi. In order for these research findings to have their impact felt in language pedagogy, the researchers should, however, follow the well-established principles of the methodology. In an attempt to investigate the extent to which the researchers abide by these standards, the current paper reviews a number of corpus-based studies carried out on the ELT textbooks in Malaysia. It was specifically intended to shed light on the possible methodological flaws committed by the researchers. The study, it is the hope of the researchers, could provide some guidelines for the researchers to carry out further corpus-based studies on EFL/ESL instructional materials as well.

Keywords: Corpus Linguistics, ESL Textbooks, Malaysia, Methodology

1. Introduction

The development of corpus-based materials, like dictionaries, serves as evidence of the relevance of corpus linguistics to language description and language instruction as well. In recognition of this notion, different researchers have managed to investigate the ELT materials within the framework of corpus linguistics and compare textbook materials with some reference corpora. It might perhaps come as a great shock that a noticeable body of such corpus-based research has indicated the inadequacy of ELT materials in describing evidence from real language use. For instance, corpus evidence enabled Kettermann (1995) to argue that the prescription of the backshift rule for tenses in reported speech constructions in pedagogical grammars fails to comply with actual language use. Likewise, Tognini-Bonelli (2001) laments the remarkable degree of lack of agreement between descriptions of the indefinite pronoun ‘any’ in pedagogical grammars and the way it is used in natural corpora, holding that about 50% of the occurrences of the item in a corpus cannot be explained by the descriptions given in pedagogical grammars.

2. Contributions of Corpus Linguistics to ESL Teaching

The impact of corpus linguistics on language description has been so profound that it is sometimes compared to the effect “of the telescope on astronomy” (Ranali, 2003, p. 2). In particular, it threw doubt on the conventional demarcation between vocabulary and grammar, the convention upon which much of the language teaching methodology and practice was built. Contention has even been made that it has revolutionized the field of lexicography. The impact of corpus linguistics on language teaching is, however, a different story with “changes arising from CL appear to be creeping into ELT slowly over time” (Ranali, 2003, p. 3). Aston (2001) looks at the contributions of corpus linguistics to the process of L2 teaching and learning with respect to three major areas, namely Language description, Corpus linguistics methods in L2 classes and Learner corpora.

The first important contribution of corpus linguistics to language teaching is the description of various language features. These descriptions are especially important for language teaching because ESL textbooks have commonly been shown to include incomplete or misleading explanations (Gabrielatos, 2003; Kettermann, 1995; Koprowski, 2005; Tognini-Bonelli, 2001).

Despite the illuminating empirical findings offered by employing the corpus linguistics principles, it is helpful to point out that the findings are sometimes questionable either because of a lack of attention to computational rules advocated by the field or inattention to the conceptual particulars of the discipline. For instance, it is often the case that the percentage may not convey meaningfully the frequency of use of a linguistic feature for the lack of balance between the sizes of the corpora. Likewise, failure to make a distinction between such concepts as token, type and family might lead the researcher to conclusions that are otherwise unjustified. In a similar way, interpreting the findings of the study within the purely corpus linguistics framework tend to result in remarks unwarranted in terms of the pedagogical issues. More specifically, while corpus linguistics has to do with the frequency of use of the language elements, pedagogy appeals to the usefulness, learnability and teachability of the language units whatsoever.

3. Development of ESL Corpora in Malaysia

A small number of ESL corpora have recently been developed or are being created in Malaysia. To begin with, the University of Putra Malaysia developed the EMAS corpus in 2002. The corpus includes half a million running words containing written and spoken language collected from students Form one and Form four of selected primary and secondary schools in three Malaysian states. Likewise, Mukundan and Anealka (2007) developed the forensic corpus of secondary school English language textbooks used between 1990 and 2002, containing a total of 280,000 running words. The same researchers also developed a corpus of the Malaysian ESL textbooks of current use with the secondary level students. Furthermore, Menon (2009) refers to three other corpora which are under construction, namely, the MACLE corpus based on students' essays, the COMEL corpus, a spoken corpus project, and the CALES corpus covering the argumentative essays being written by university students taking English proficiency courses at a number of Malaysian universities. Finally, Rezvani Kalajahi and Mukundan (2013) compiled a corpus of the argumentative essay writing composed by the Malaysian high school and college level students. This corpus consists of a total number of 500,000 running words.

With more and more attempts directed towards compiling and developing different corpora in Malaysia, more and more researchers have managed to study and explore different language forms in these compilations. Of particular interest in the present paper is a critical review of the corpus-based studies carried out on the corpus of English textbooks prescribed for use by Malaysian ESL students at the secondary level.

4. Corpus Studies on Malaysian ESL Textbooks

Different corpus studies with different foci have been carried out on the Malaysian ESL textbooks. In order to examine the adequacy of vocabulary load and distribution in two Malaysian ESL textbooks developed by two different writers for the same school level, Mukundan (2007) investigated the materials to document the degree of agreement between the two books in rendering the content of the syllabus. Having converted the textbooks and the syllabus word list into the Tagged Image File, he saved the files in the computer to be processed into computer text files. Then, he used the KeyWords, the WordList and the Concord tools of the WordSmith program 3.0 to analyze the files. Data analysis enabled the researcher to report a significant degree of disparity between the two books in covering the vocabulary suggested by the syllabus. While textbook One appeared to be an appropriate candidate, the other one failed to be representative of the syllabus in terms of coverage, distribution of words, density (type/token) and consistency(token/type) ratios, suggesting some degree of mismatch between the textbook and the syllabus requirements .

While the researcher has provided a list of words repeated in each unit of the books, no mention has been made as to the frequency counts of the items, however. Neither has any mention been made of the part of speech of the vocabulary items in the textbooks and those contributing to the intensity and consistency ratios. This is an important issue to consider as it is possible for the same orthographic form to serve different functions in the language. For instance, the word 'present' can function as a verb, a noun and an adjective based on the context in which it occurs.

In another study, Mukundan and Menon (2007) analyzed the language used in Science, Math and English language textbooks prescribed for Form One students in the southern zone of the nation. In order to identify the type of language Malaysian ESL students require in the process of learning Science and Math at schools, they looked for the most common word class among the key words in the corpora and the difference between the Science and Math textbooks. Results showed that nouns received a greater emphasis than verbs and adjectives in the three textbooks. Promising for the students, however, was that the nouns were mostly semi-technical and nontechnical words which were held to be easy and appropriate for the students' level. There seemed to be an equal number of lexicalized and delexicalized verbs in the Science and Math textbooks, with lexicalized items in need of special attention due to the special meanings they carried in these texts. Adjectives appeared to be a different story, however. The type of adjectives in Science, mostly being derived forms, was reported to be more complex than adjective type used in the Math text.

Despite the revealing findings of the study, care should be exercised in the application of the results. To begin with, the unit of measurement of the vocabulary items in the study was 'token' rather than 'lemma or type'. For instance, instead of 'ANGLE' and 'SHADE', 'angled' and 'shaded' were reported as the key word verbs in the Math text. However, if the unit of measurement had been 'lemma', that is if the other inflected forms like 'angle' and 'angles' or 'shade', 'shades' and 'shading' had been counted, the exclusion of some words from the list and the inclusion of some others would not have been unexpected. On the other hand, while the linking verb 'is' was reported as a key delexicalized verb in both Science and Math textbooks, it is not clear whether 'is' appeared as a single lexical verb or it was included in passive constructions. This is an important distinction to make as passive structures feature scientific texts, and the Science and Math textbooks are within the domain of scientific texts. As a result, differentiating 'is' as part of a passive structure from 'is' as a single lexical verb would give a clearer picture of the function of this verb in the textbooks.

In another small scale study on the same data, Mukundan and Menon (2008) investigated the use of nouns in the Science and English language textbooks. More specifically, they studied the distribution of nouns and noun collocations in the corpora. Results indicated that most of the nouns in the Science textbook consisted of semi-technical and nontechnical words which fitted in with the students' level of schooling. The researchers, however, reported some syntactic structures of the science noun collocates with no parallel in the English language text, specifically the 'Verb + Noun' combination. While phrases like 'living things' and 'exhaled air' were raised as clues to the idiosyncrasy of the paradigm, the reader is not provided with any concordance line evidence to see them in the context. What is more, the phrase 'living things' fits in more with the pattern 'Adj + Noun' than with 'Verb + Noun', and so might the phrase 'exhaled air'.

There are also some other corpus-based studies carried out on Malaysian ESL pedagogic corpus. For instance, Mukundan and Roslim (2009) studied the use of prepositions in the textbooks corpus against the reference corpus of the BNC. Likewise, Mukundan and Khojasteh (2011) compared the use of modals in this pedagogical corpus and the BNC. The researchers in both studies argued that the frequency of occurrence of modals and prepositions differed between the textbooks corpora as a whole and the reference corpus of the BNC. They also provided the reader with a set of figures indicating that the use of these two grammatical forms differed from textbook to textbook.

Although the two studies raised the difference between the two corpora as some sort of deficiency on the part of the textbooks, some reservation has been raised against using corpora as a basis upon which the instructional materials should be developed, however. For instance, Widdowson (2000) argued that corpus-based studies describe use, whereas pedagogy is concerned with usefulness. Additionally, the L2 learner community is not the same as that of the L1 native speaker.

As far as the difference between the textbooks are concerned, it should be pointed out that they were developed for learners with different levels of language proficiency and did not need necessarily to deal with the issues under study identically. Moreover, they were not of equal size. Thus, if the researchers, following the standard corpus linguistics analytical procedures, had normed the raw frequencies to occurrences per 100,000 words, for example, to allow for comparisons across textbook corpora of unequal size, it would be likely that the differences they found between the Forms might have been insignificant.

Moreover, although Mukundan and Roslim (2009) provided a good account of forms that almost exclusively function as prepositions of place like 'behind, in front of, between, etc.', they failed to discuss other elements like 'in, on, down, over, back, away, etc.', that might serve not only as prepositions of place, but also as prepositions of time, adverbs and verb particles. Such a distinction is necessary as the writers indicate "The nature and complexity of prepositions have consequently led to problems with prepositions for ESL teachers and learners" (p. 14). In addition, a major problem facing the ESL learners is distinguishing between the functions that these forms might serve as prepositions, adverbs and verb particles. The growing attempts in the related literature for developing various tests to differentiate these functions from one another (Bolinger, 1971; Fraser, 1976; Darwin & Gray, 1999; Zarifi, 2013) show the significance of the issue.

Finally, while the Mukundan and Khojasteh (2011) attributed the misuse/ungrammatical use of main verb forms by the learners to the lack of appropriate instruction of these forms by the teacher or insufficient repetition of these forms in the teaching materials, care should be taken that some other factors might be at work. First, some languages like Malay tend to show the notion of tense mainly through the inflection of the main verb or adverbs of time; therefore, the learners often prefer to indicate tense by inflecting the main verb, overlooking the well-established grammatical rule that simple form of the main verb follows the modal. Second, inaccurate use of these forms by the learners might be an indication of their developmental stages of language learning or inter-language phenomenon.

In a very recent corpus-based study on the Malaysian ESL Textbooks, Mukundan, Chiew Har and Nimehchisalem (2012) investigated the presentation and distribution patterns of the In/Definite Articles 'a', 'an' and 'the' in the Form One through Form Five textbooks prescribed for use by the secondary school level students. The study provided a good account of the frequency and distribution of the articles used in the corpus. It also revealed the frequency and distribution of the colligation patterns of the articles. Despite the useful information the study offered on the treatment of the articles in the textbooks, the framework employed for the identification of article colligation structures is not comprehensive at all. First, there are some common colligational patterns associated with the Particles overlooked in the framework. For instance, the framework does not account for the structure 'A + Quantitative expressions' like 'a few' and 'a little' which is a well-known structure in the language and had a large number of frequency counts in the textbook corpus. Likewise, the colligational pattern 'The + Comparative forms' exemplified by such strings as 'the more, the better, etc.' is not included in the study though the Concordance query gave a number of instances of this structure in the corpus. Second, while 'only' is an adjective in sequences like 'the only person', the framework, surprisingly enough, assigns a separate structure to the occurrences of the sequence 'The + Only', distinguishing it from the general pattern 'The + Adj'. Third, it seems not sensible why, if the string 'A + Adj' deserves a separate colligational pattern, the sequences 'An + Adj' and 'The + Adj' have not been given a separate pattern. Fourth, colligation of the definite article 'the' with descriptive adjectives like 'poor, rich, good, etc.' has not been accounted for by the framework. It should, however, be pointed out that the combination 'the + Adj' is a well-established pattern in the English language that is used to change the adjective into a plural noun when the speaker or writer refers to all the people that the adjective describes.

In addition to the inadequacy of the framework used, the study seems to be suffering in still another aspect as well. Despite the researchers' conclusion that the indefinite article 'an' was under-presented in the textbooks and they, therefore, called for the provision of supplementary teaching materials that could make up for this deficiency, comparison of the data against the BNC revealed that it was the article 'a' that was overused rather than the article 'an' being underused in the textbooks. Normalization of the data showed that the indefinite article 'an' in the pedagogic corpus enjoyed almost the same frequency counts as in the BNC.

The last but not the least, in another recent corpus-based study on the Malaysian ESL textbooks, Philip, Mukundan and Nimehchisalem (2012) explored the treatment of conjunctions in this pedagogic corpus. Although the accuracy of the data from the BNC that were used as a basis for comparison is questionable, they provided some insightful information on what types of conjunctions were used in the materials and how the use of these elements in the corpus compared with their presentation in the BNC. For instance, they found that the distribution of coordinating conjunctions was higher as compared to subordinating and correlative conjunctions. Moreover, correlative conjunctions appeared to be the least occurring conjunctions among the three types. Furthermore, the different types of conjunctions appeared to enjoy a similar rank order in both the Textbook Corpus and the BNC. This study, unfortunately enough, failed to paint a clear picture of the semantic functions of these forms in the corpus. Conjunctions can be used to link different segments of a text such as nouns, phrases, clauses and sentences. Despite the various grammatical functions of these forms, the paper also remained silent on what constituent types these elements combined in the corpus and with what proportion.

5. Conclusion

The present study has reviewed a number of the corpus-based studies investigating the treatment of different language aspects as used in the Malaysian ESL textbooks prescribed for use by the secondary level students. Despite the remarkable insight they provided into the textbooks language, they were shown to be suffering from some conceptual, methodological and analytical problems. First, some of the studies appeared not to take the differences between such concepts as 'token', 'type', and 'family' into consideration in their interpretation of the findings, hence coming up with conclusions that could be challenged. For example, Mukundan and Menon (2007) adopted 'token' as their unit of measurement of the vocabulary items and reported 'angled' and 'shaded' as the key word verbs in the corpus. Adopting 'type' as the unit of measurement might have led to a different conclusion. Such a stance would give a clearer picture of the way vocabulary categories were treated in the corpus for 'type' involves items of the same nature and meaning but with different inflectional forms. A few comparative studies dealing with the congruence between the textbook language and the BNC ended with counting only the frequency occurrence of the different aspects of the language without normalizing the data. Such treatment often leads to inappropriate conclusions simply because of the probable imbalance between the corpora in terms of their sizes. For instance, normalization of the data rejected the disagreement between the textbooks and the BNC in terms of the indefinite article 'an' as was indicated by Mukundan, Chiew Har and Nimehchisalem (2012). Likewise, the framework they employed for the identification of article colligation structures was incomprehensive. For illustration, some of the patterns commonly associated with articles like 'A + Quantitative expressions', 'The + Comparative forms', etc. were not incorporated in the framework. Finally, some studies managed to take the BNC as their reference corpora and made pedagogical suggestions simply because of the divide between the textbooks and the general corpora in dealing with a given phenomenon. These suggestions would, therefore, be thrown into question as pedagogy and corpora do not stick to the same principles.

Refernences

Aston, G. w., Paul (Ed.). (2001). *Learning with corpora.* Houston: Athelstan.

Bolinger, D. (1971). *The phrasal verb in English*. Cambridge, MA: Harvard University Press.

Darwin, C., M., & Gray, L. S. (1999). Going after the Phrasal Verb: An Alternative Approach to Classification. *TESOL Quarterly, 33*(1), 65-83.

Fraser, B. (1976). *The verb-particle combination in English*. New York: Academic Press.

Gabrielatos, C. (2003). *Conditionals: ELT typology and corpus evidence.* Paper presented at the 36th Annual Meeting of the British Association for Applied Linguistics (BAAL), University of Leeds.

Kettermann, B. (1995). Concordancing in English language teaching. *TELL and CALL, 4*(95), 4-15.

Koprowski, M. (2005). Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal 59*(4), 322-332.

Menon, S. (2009). *Corpus-based analysis of lexical patterns in Malaysian secondary school science and English for science and technology textbooks*. Universiti Putra Malaysia, Serdang.

Mukundan, J. (2007). Irregularities in vocabulary load and distribution in same level textbooks written by different writers. *Indonesian JELT, 3*(1), 99-118.

Mukundan, J., & Anealka, A. H. (2007). A forensic study of vocabulary load and distribution in five Malaysian Secondary School Textbooks(Forms 1-5). *Pertanika Journal of Social Science and Humanities, 15*(2), 59-74.

Mukundan, J., Chiew Har, A. L., & Nimehchisalem, V. (2012). Distribution of articles in Malaysian secondary school English language textbooks. *English Language and Literature Studies, 2*(2), 62-70.

Mukundan, J., & Khojasteh, L. (2011). Modal auxiliary verbs in prescribed Malaysian English textbooks. *English Language Teaching, 4*(1), 79-89.

Mukundan, J., & Menon, S. (2007). Lexical similarities and differences in the mathematics, science and English language textbooks. *Kata, 9*(2), 91-111.

Mukundan, J., & Menon, S. (2008). Nouns and their extended units of meaning: A corpus analysis of nouns in the Science and English language textbooks. *Journal Sastra Inggris, 8*(2), 90-111.

Mukundan, J., & Roslim, N. (2009). Textbook representation of prepositions. *English Language Teaching, 2*(4), 13-24.

Philip, A., Mukundan, J., & Nimehchisalem, V. (2012). Conjunctions in Malaysian secondary school English language textbooks. *International Journal of Applied Linguistics & English Literature, 1*(1), 1-11.

Ranali, J. M. (2003). *ELT coursebooks in the age of corpus linguistics: constraints and possibilities*. Birmingham: University of Birmingham.

Rezvani Kalajahi, S.A. & Mukundan, J. (2013). *Malaysian Corpus of Students' Argumentative Writing (MCSAW).* Australia, AIAC.PTY.LTD.

Tognini-Bonelli, E. (2001). Corpus linguistics at work. Amsterdam: John Benjamins Publishing Co.

Widdowson, H. (2000). On the limitations of linguistics applied. *Applied Linguistics, 21*(1), 3-25.

Zarifi, A. (2013). *Establishing and evaluating phrasal verb use in a Malaysian secondary school textbook corpus.* Universiti Putra Malaysia, Serdang.