Corpus Linguistics and the Classroom: Avenues for Innovation

Iain McGee

Abstract In this paper, I describe three ways in which corpus linguistics research and findings have influenced my own classroom practice over recent years in the Arabian Peninsula, including Oman, whether this be in general English language classrooms or in linguistics classes. The three general skill areas considered along with the specific corpus-based focus are vocabulary (specifically, synonym differentiation), grammar (a comparison of a function of *going to* and *will*) and writing (editing with shell nouns). I suggest that exposing students to corpus-based insights can make language and linguistics study more engaging and that the iconoclastic nature of some corpus-based findings can be a catalyst for significant learning moments in our classrooms.

Keywords Corpus linguistics · Data-driven learning · Semantics · Shell nouns · Grammar rules

1 Introduction

The field of corpus linguistics has been defined in a number of ways. A fairly representative definition is the following provided by Bennett (2010, p. 2): 'Corpus Linguistics approaches the study of language in use through corpora (singular: corpus). A corpus is a large, principled collection of naturally occurring examples of language stored electronically'. Like all definitions, there are some points of contention in how Bennett defines this field. Firstly, there is the issue of size. Stating that a corpus is, by default, *large* is not necessarily true. This requirement is largely related to the issue of function. As Francis (1982, p. 11) noted, a very small corpus can help determine the relative frequencies of letters, their typical combinations and the use of punctuation marks. Further, Sinclair (1991, p. 100) suggested that one million words of data would suffice to document a language's grammar. This would certainly not be considered to be 'large' in corpus linguistics studies today. The

I. McGee (✉)
Majan University College, Muscat, Oman
e-mail: iain.mcgee@majancollege.edu.om

R. Al-Mahrooqi, C. Denman (eds.), *English Education in Oman*, English Language Education 15, https://doi.org/10.1007/978-981-13-0265-7_18

issue of size has become particularly important in relation to phraseology studies. De Beaugrande (1999), for example, suggested that a 200 million word corpus was too small to study phrases containing 'couldn't help…'. In sum, the 'large' requirement in Bennett's definition must be understood and interpreted with regard to one's own research interests.

The second contentious issue is that the data collected in the corpus be *principled*. Specialised corpora are principled: a 'Works of Shakespeare' corpus will contain the works of Shakespeare and only his works; a learner corpus will *only* contain language produced by learners. The issue becomes a little more complex when we consider general corpora, in which the data are mixed (e.g. spoken and written or from a variety of written sources). The designers of the British National Corpus (BNC), a large general corpus, have made much of its claim to representativeness and the principles behind the collection of data (see, e.g. Aston & Burnard, 1998; Leech, Rayson, & Wilson, 2001), but there are clearly issues: spoken language constitutes just 10% of this corpus. Is this 'principled'? It is, more accurately, 'pragmatic': spoken data is more expensive to transcribe and tag than electronic text. Like the word *large*, *principled* is open to interpretation. One last comment is in relation to the Internet. Is it a corpus? Some linguists have indeed called the web a corpus (e.g. Kilgarriff & Grefenstette, 2003), while others prefer to call it a text collection (e.g. Stubbs, 2000). The reason for this difference may well be over the issue of the 'principled collection' of data definitional requirement noted by Bennett (2010). Beyond these two contentious points, however, Bennett's definition is as good a starting point as any other.

Data from corpus linguistic studies are already informing what goes on in the second language classroom. A well-known series (Touchstone, see McCarthy, 2004, on the corpus-based nature of this series) is just one of many corpus-based text series, and though organised in a fairly traditional way, it contains elements which have been directly informed by corpus data in areas such as frequency of usage information, more 'natural' conversational exchanges and collocation or lexical word combination pattern information.

The utilisation of corpus data to inform the creation of text materials is typically termed an *indirect*, as opposed to *direct*, use of corpus data. The term *direct* is typically reserved for learner interaction with corpus data, and it is this specific use of corpus data which I wish to focus on in this paper. The three different skill areas I wish to consider are vocabulary, grammar and writing.

2 Vocabulary: Corpus Data and Synonym Differentiation

The first issue where I believe corpus data can help our students is in the study of semantics and vocabulary. The first response of teachers to student questions about differences between synonymous words tends to be paradigmatic in nature: two words are contrasted in terms of their denotational (i.e. difference in meaning), connotational (attitude) or stylistic (formality) differences (see Inkpen & Hirst, 2006).

So, a teacher may respond to his or her student that *tiny* is smaller than *small*, *slim* is more positive than *thin* and *purchase* more formal than *buy*. So far, so good. The problem is that there are many word pairs, or groups of synonymous words, for which such attempts at differentiation are not quite so straightforward. Students in my own classes in Oman have, at times, suggested to me that *large* is bigger than *big*. From where have they got these ideas – their teachers? It is, of course, possible that students have never really considered the syntagmatic environment of words, and this may explain such attempts at differentiation. It may also be that synonym exercises present in school books do not really help students think in the right way about words and their relationships. Engagement with corpus data can enable students to explore what the differences really are. Responding to student questions with corpus data is a good strategy, partly because student interest will be higher and partly because it begins to help students to try and answer their own questions. The specific advantage of corpus data when it comes to resolving issues of word meaning is that it enables us to expand our interest away from purely paradigmatic or narrowly semantic considerations of a word to the syntagmatic environment of the word. Rather than attempting to 'understand' a word in isolation, one can observe its environment and see how it is used, rather than merely consider what it 'means'. Indeed, whether words actually mean anything out of context is an interesting subject of debate (see Kilgarriff, 1997). Below, I contrast dictionary information from the Longman Dictionary of Contemporary English (2003) with corpus data information for the words *study*, *report* and *research*.

Study: A piece of work that is done to find out more about a particular subject or problem and usually includes a written report.
Report: An official piece of writing that carefully considers a particular subject and is often written by a group of people.
Research: A serious study of a subject that is intended to discover new facts or test new ideas.

The semantic differences presented here between *study*, *research* and *report* are rather contrived and artificial. A simplistic understanding of the semantic information provided for these words *suggests* the following:

- *Research* is characterised by its seriousness which, in turn, suggests that research is of high quality – which we know not to be true of all research.
- The focus of a *study* may be on a problem, whereas research is not so focused (and yet much research is problem-focused).
- It suggests that a *study* (alone) is written, whereas *research* is not – which, again, is rather confusing.
- The key characteristics of a *report* are that it is a group effort (but we know that most hard science research is multi-authored) and that it is official (suggesting that governments and government bodies do not research, per se, but, alternatively, 'carefully consider').

Clearly any student attempting to understand the differences between these words is likely to come out of the dictionary page either in a state of confusion, or,

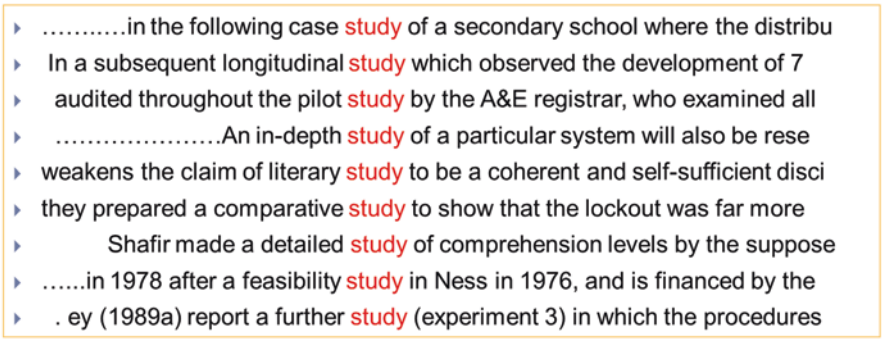

Fig. 1 Concordance lines for *study* from the BNC

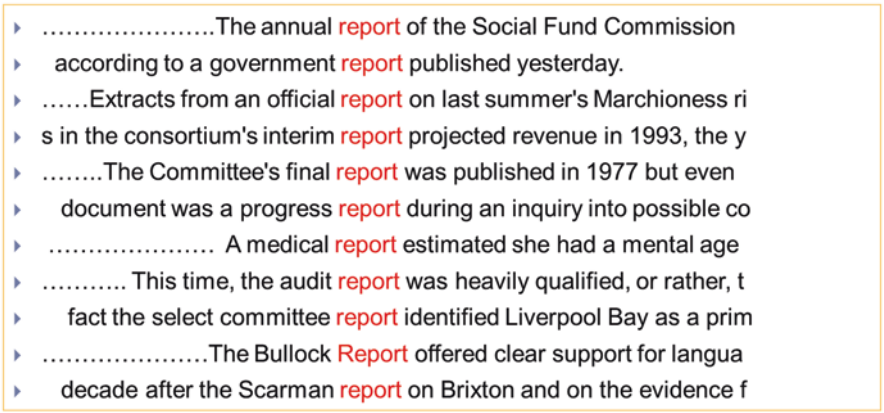

Fig. 2 Concordance lines for *report* from the BNC

possibly worse, with a highly questionable, wooden and ultimately inadequate understanding of these words. Below, I provide a number of concordance lines from the BNC for each word (Figs. 1, 2 and 3).

When students are given such data, they will notice different things about how these words are used. Indeed, I am constantly pleasantly surprised at the various insights provided by different students in my classes in Oman. I have found students quite open to sharing with myself and their peers what they have found and indeed being excited to do so. A helpful categorisation scheme (though just one suggestion) is to think about the words' collocation and colligation patterns and finally the words' meanings.

Collocation The adjectives collocating with *study* and *report* indicate that reports are often connected to who authored or sponsored them – words such as *government*, *audit* and *committee* indicate this. *Study*, on the other hand, seems to be more associated with the type of research which has been conducted, rather than who

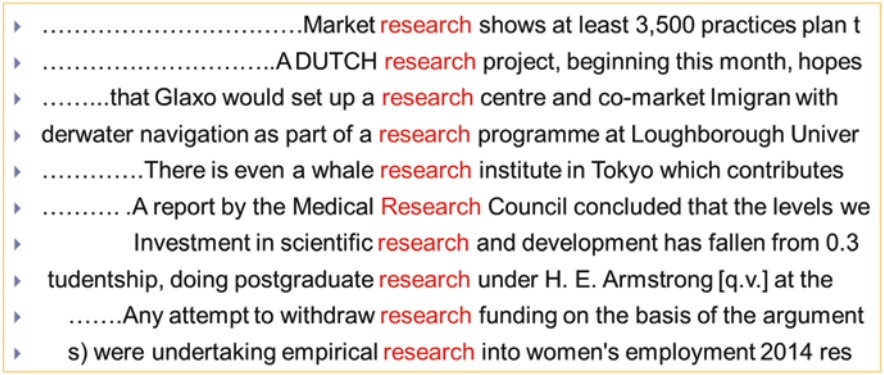

Fig. 3 Concordance lines for *research* from the BNC

Table 1 A colligation matrix for the words *study*, *report* and *research*

	Study	Report	Research
Part of a larger noun phrase			✓(Very strong tendency)
…into			✓
…on		✓	
…of	✓	✓	
In past tense context	(Some)	✓	(Not clear)

conducted it, per se, e.g. *literary*, *detailed* and *feasibility*. *Research* has fewer adjective or attributive nouns connected with its usage, though some may strike us as very strong pairings (e.g. *market research*, *empirical research*).

Colligation What can we notice about the grammatical patternings? This can be more easily investigated through the use of a matrix (see Table 1).

It should be noted that the conclusions drawn from analysing such data are tentative, being that the data are limited. However, with the data we have, we can note that *research* has quite a different grammatical patterning when compared to *study* and *report* and that for preposition patterns all the words are rather different from each other. Concerning meaning, one might, on consideration of the data in the table and figures, consider *research* to be what makes up part of a *study*, which might, in turn, be published as a *report*.

So, how exactly can students' analyses of the collocation and colligation patterning help them understand the differences between synonyms? Firstly, the data point to tendencies (e.g. in the semantic field of the collocates with which the node words occur or the colligational patternings) rather than absolutes: dictionary definitions tend to be too precise. In addition, the provision of concordance lines is a good way for students to pick up some very prototypical combinations: the focus on the word is balanced by the focus on the word's environment. Setting up tasks where students investigate the differences between *big* and *large*, and *happy* and *glad*, for example,

put students in the driving seat and empower them to become the authorities in the classroom, rather than their teachers. This healthy reversal of roles is often called for in the literature.

3 Grammar: Corpus Data and Referring to the Future

Insights from corpus linguistics research are affecting our views of grammar, though it would be more accurate, in today's climate, to speak of the grammars of language (see, e.g. Biber, Johansson, Leech, Conrad, & Finegan, 1999). Tensions between sentence-based and discourse-based grammar have been documented in the literature (e.g. Cook, 1989; Hughes & McCarthy, 1998), and I do not wish to focus on these here, beyond stating that a key strength of discourse-based grammar is that it allows us to examine, with more care, functions and usage patterns.

As an example case, we can consider how *will* and *going to* are traditionally taught in a decontextualised, sentence-based approach, and then contrast this with an inductive (corpus-driven) discourse-based approach. A fairly typical way of differentiating 'will' and 'going to' is given below from New Headway Plus (Pre-Intermediate): "Going to is used to express a future decision, intention, or plan made before the moment of speaking… Will is used to express a future decision or intention made at the moment of speaking" (Sorars & Soars, 2006, p. 136). As EFL teachers, we have probably all taught the above and, perhaps, religiously corrected student 'mistakes' accordingly. There are, of course, instances where the rule 'holds'. However, corpus data indicates that there are many cases where it does not. The following extracts, found on the Internet after just a few minutes' search, do not fit in well with the above-noted differentiation.

Extract 1 (Political News)

> *"On Thursday, Assistant Secretary of State Victoria Nuland is going to visit Ukraine, along with several foreign ministers of European countries," the agency quotes. Head of EU diplomacy Catherine Ashton and European Commissioner for Enlargement and European Neighbourhood Policy Stefan Fule will also visit Ukraine this week.* (Voice of Russia, 2014, para. 1)

Extract 2 (Rolls Royce Board Meeting)

> *So I'm going to give you an overview of our performance in 2013 and provide some longer-term context, and then I'm going to cover guidance for 2014..... and I'm going to spend a few minutes explaining them. Mark will then talk you through the numbers, and then we'll have a Q&A.* (Thomson Reuters Streetevents, 2014, p. 2)

Extract 3 (Sports News)

> *And how does Ehlers plan to celebrate his birthday?*
> *"I'm going to school and then to practice and then I'll open some gifts when I get home," he added with a smile.* (Metronews, 2014, para. 9)

Extract 4 (Entertainment News)

> *I'm working on some new songs for a new record that I'm going to start recording, hopefully in late April. In July, I've got a tour out to the Pacific Northwest. Then I'll come back here,….* (Opoien, 2014, para. 11)

In all of these instances, the speaker or writer begins talking about the future with *going to* and then switches to *will*. The key question that must be asked of the data is whether, functionally, *will* is being used differently from the initial *going to*. On balance, I would suggest not: it seems to be used to achieve the same function. The only difference appears to be that *will* does not open the series of plans, whereas *going to* does. On the basis of this admittedly small set of data, we could, therefore, hypothesise the following:

When people have, or report on, a number of sequentially related plans, they start with *going to* and then might switch to *will*.

This hypothesis would, of course, need to be investigated on a larger data set. What the above kind of mini-study suggests is that we are probably too 'tight' and overly prescriptive, in some aspects of our grammar teaching: the rules we present to our students may actually be at variance with the data. From these data, we can see that *will* is indeed used for pre-arranged plans, contrary to the grammar point noted earlier.

An inductive corpus-based approach to teaching and learning grammar may well help students develop more 'reasonable' and, ultimately, more accurate ideas about grammar usage. The alternative for our students is an awful lot of 'unlearning' and then relearning. Indeed, my own students in Oman have expressed their deep concern when faced with such data. Having placed great hope in the rules they were taught in school, they are brought face to face with data which challenge the rules and, at the same time, challenge parts of their previous learning. This is not a helpful state of affairs. With data and guidance from the teacher, students can begin to form their own hypotheses about functional differences in English tense and aspect usage.

4 Writing: Corpus Data and the Encapsulation/ Interpretation of Previous Discourse Elements

The final area of teaching I would like to touch on is writing, more specifically editing, and how corpus data insights can help in this area. When we consider the kind of feedback typically given to students on drafts of reports or term papers, we normally consider the following:

- Grammar issues (e.g. run-on sentences)
- Signposting (e.g. conjunctions)
- Mechanics (e.g. punctuation, spelling)
- Lexical issues (e.g. collocational and colligational issues)
- Organisation

These are all legitimate and useful areas to consider. However, in addition, an area I have explored with my own students in Oman and across the Arab Gulf when at the editing stage of their writing is the use of shell nouns. Some of the most common nouns in the English language are abstract nouns (e.g. *idea*, *problem*, *situation*), and corpus studies have not only highlighted their frequency but also their use. A key function of these words is often underutilised by our students, i.e. the shell noun function. Schmid (2000, p. 4) defines shell nouns as 'an open-ended functionally-defined class of abstract nouns that have […] the potential for being used as conceptual shells for complex, proposition-like pieces of information'. What these nouns do is encapsulate, and possibly interpret, previous discourse of various sizes ranging from units larger than the noun phrase to the paragraph. The common anaphoric use of these shell nouns should be noted. However, in addition to having a backward-focused orientation, authors also utilise these nouns to move the argument forward as well. Exercises such as the one noted below, using corpus data, can help students realise the importance and function of these nouns.

Example Exercise
Which of the following nouns goes into the following gaps?
plan/situation/achievement

1. The skills required to build such systems were rare and also the required combinations of software and computer hardware were expensive. This _____ has more recently been reversed.
2. The growth in revenue enabled a total of £25,000 to be spent on research and development. This _____ was unheard of in the company's history.
3. The IBA hoped to raise half the capital for a new company from Midlands money, with ATV providing the rest. This _____ failed on two counts.

In drafts of reports, I have required students to use shell nouns and highlight their usage to me through underlining. Additionally, students can be asked to develop the interpersonal element of their writing, often in a later draft, through the use of attributive adjectives before these nouns. The exercise below was developed to address this specific point and to give students the opportunity to invest personal opinion in the text they are writing.

Example Exercise
Match the adjectives on the left with the nouns on the right with which they typically co-occur.

1. Great, real, welcome, remarkable	A. Plan
2. Unsatisfactory, encouraging, existing, unusual	B. Situation
3. Ambitious, original, controversial, strategic	C. Achievement

While students may face challenges in using these shell nouns accurately (see McGee, 2009), appreciating the facilitating function of these special nouns, and using them intelligently, is an essential part of making writing more natural and facilitates the flow of written discourse. Students can also be encouraged to examine previously published written work for its usage of shell nouns, see how they are

used by their authors, and begin to use them in their own writing. I have done such with my own students in Oman: getting students to actually notice such usage and appreciate how this can improve the quality of their own texts is critical.

5 Conclusion

In this paper, I have touched on three possible uses of corpus data in the classroom – uses to which I have put corpus data in my own teaching in the Sultanate, and which, I believe, have been reasonably well-received by students. In terms of considering the pedagogical implications from the experiences described above, I believe that the following are amongst the most significant:

1. Synonymy is a minefield: Moving away from single word synonym pairs and raising awareness of collocation and phraseology can only be a good thing for our students.
2. Teachers do not know it all: Our attempts to explain differences between words can, quite frankly, be embarrassing. Corpus data show teachers and students what is typical and frequent/infrequent. As such, they can only be helpful for students (and teachers, too).
3. Real texts and corpus data must live in harmony with our 'rules': If students can see the disparities, we should too. The implications of disharmony should impact our teaching approach and assessment.
4. Inductive learning will not disappoint: Learners of (overly simplistic) rules will always be troubled with language data. This is not because there are no rules, but because the rules are more subtle than we typically state. This is a fact to which the data eloquently point.
5. Cohesive linking in texts needs more focus: The power of shell nouns to manage discourse flow, allow interpersonal engagement, and add to the quality of written text needs to be appreciated by students.

In closing, it is important to stress that corpus data are not the panacea of English language teaching. Corpus linguistic information is simply another tool which can be employed by teachers for the benefit of their learners. I am not suggesting that corpus-based insights be adopted wholesale into our teaching. Corpus data is not a methodology, and the pedagogical implications of these data are not entirely clear: simplification, teachability, motivation, assessment, etc. are all additional areas which must be considered and valued in our classrooms and institutions.

Various attempts to consider how corpus data can inform teaching are being made. For example, a special issue of the journal *ReCALL* (14 February 2014), entitled *Researching uses of corpora for language teaching and learning*, is devoted to the same focus as this paper with a number of papers specifically considering the links between corpus data and how they can help in teaching and learning writing skills. These are important developments which instructors and students in Oman and, indeed, across the entire world, would benefit from becoming more familiar with.

Finally, it is my own conviction that corpus data are not just of use for students – teachers too can experience 'significant learning moments' in their classrooms, together with their students, as they consider corpus data. One way to keep fresh as a teacher is to challenge cherished beliefs on a daily basis, whether they be about vocabulary, grammar and writing, and corpus data constantly do this.

References

Aston, G., & Burnard, L. (1998). *The BNC handbook*. Edinburgh, UK: Edinburgh University Press.

Bennett, G. R. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. Ann Arbor, MI: University of Michigan Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Grammar of spoken and written English*. Harlow, UK: Pearson Education.

Cook, G. (1989). *Discourse*. Oxford, UK: Oxford University Press.

de Beaugrande, R. (1999). Reconnecting real language with real texts: Text linguistics and corpus linguistics. *International Journal of Corpus Linguistics, 4*(2), 243–259.

Francis, W. N. (1982). Problems of assembling and computerizing large corpora. In S. Johansson (Ed.), *Computer corpora in English language research* (pp. 7–24). Bergen, Norway: Norwegian Computing Centre for the Humanities.

Hughes, R., & McCarthy, M. (1998). From sentence to discourse: Discourse grammar and English language teaching. *TESOL Quarterly, 32*(2), 263–287.

Inkpen, D., & Hirst, G. (2006). Building and using a lexical knowledge base of near-synonym differences. *Computational Linguistics, 32*(2), 223–262.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities, 31*(2), 91–113.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics, 29*(3), 333–347.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. London: Longman.

Longman Dictionary of Contemporary English (LDOCE). (2003). London: Pearson Longman.

McCarthy, M. (2004). *Touchstone: From corpus to course book*. Cambridge, UK: Cambridge University Press.

McGee, I. (2009). Traversing the lexical cohesion minefield. *ELT Journal, 63*(3), 212–220.

Metronews. (2014, February 14). *Birthday boy: Ehlers ignites shorthanded Mooseheads to 5-1 victory over Gatineau*. Retrieved from http://www.metronews.ca/sports/2014/02/14/back-on-track-ehlers-ignites-mooseheads-to-5-1-victory-over-gatineau.html

Opoien, J. (2014, February 13). A homecoming, of sorts, for Josh Harty. *The Capital Times*. Retrieved from http://host.madison.com/ct/

Schmid, H. J. (2000). *English abstract nouns as conceptual shells: From corpus to cognition*. Berlin, Germany: Walter de Gruyter.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.

Soars, J., & Soars, L. (2006). *New headway plus: Pre-intermediate student's book*. Oxford, UK: Oxford University Press.

Stubbs, M. (2000). Using very large text collections to study semantic schemas: A research note. In C. Heffer & H. Sauntson (Eds.), *Words in context: A tribute to John Sinclair on his retirement* (pp. 1–9). Birmingham, UK: University of Birmingham.

Thomson Reuters Streetevents. (2014). *Preliminary 2013 Rolls-Royce Holdings PLC earnings presentation*. Retrieved from http://www.rolls-royce.com/~/media/Files/R/Rolls-Royce/documents/investors/results/archive/rr-2013-fy-results-transcript-tcm92-55370.pdf

Voice of Russia. (2014, February 3). *Victoria Nuland to-visit Ukraine's Kiev Thursday*. Retrieved from http://voiceofrussia.com