DIY Local Learner Corpora: Bridging Gaps Between Theory and Practice

Neil Millar
Lancaster University, UK
n.millar@lancaster.ac.uk
Ben Lehtinen
Kanda University of International Studies, Japan
lehtinenb@yahoo.com

Corpus-based research in the past two decades has undeniably had a considerable impact on language teaching. Nevertheless, there still remains a gap between what applied corpus linguistics can offer and what teachers do (or don't do) with corpora in their teaching practice (Mukherjee, 2006). This article illustrates how compilation and analysis of corpora of student writing by teachers at a local level can help bridge this gap between theory and practice. Aimed at language teachers with no or little knowledge of corpus linguistics, we present a 'how-to-do' discussion of the basics of corpus creation and analysis, and suggest possible uses of data from such corpora in language teaching.

1. Introduction

While corpora come in all shapes and sizes, there are several types of particular relevance to language teaching. Large *general corpora* of native speaker production, such as the British National Corpus (BNC), can provide a global model of the language students seek to learn. Similarly, *specialized corpora* can provide information about more specific types of language usage, such as for academic or professional purposes. *Parallel corpora*, which contain aligned texts from the learners' native language and translations in the target language (or vice versa), can provide a good basis for studying how an idea in one language can be expressed in another language. A third type of corpus, which this article will focus on, is *learner corpora* – structured collections of language produced by language learners. Such corpora can provide information about how learner production differs from a target model and thus can inform the field of second language acquisition (SLA) (for a review of learner corpora, see Granger, 2004; Pravec, 2002). However, as Nesselhauf (2007) points out, although the insights from learner corpora would seem to be of value to language teaching, exactly *how* they might implemented in the classroom is often not made explicit.

Corpora have influenced language teaching in several tangible ways. Corpus-based

The JALT CALL Journal

descriptions have provided a model of the target language that is often different to 'traditional' intuition-based models presented to learners in textbooks (see for example, Fulcher, 1991). This can be seen the *Touchstone* series (McCarthy, McCarten, & Sandiford, 2005) – a textbook which draws extensively on the Cambridge International Corpus of North American English and descriptions of how English is actually used. Such descriptions include features of *spoken* grammar which have traditionally been absent from ELT textbooks (see McCarthy & Carter, 1995; Timmis, 2005). In addition to textbooks, findings from learner corpora have been used to tailor information in learner dictionaries to the specific kinds of errors that students frequently make (for example, the Longman Activator series). Corpus findings have also impacted on teaching methodology. The highly formulaic picture of language depicted by corpus studies has led to new approaches to language teaching – for example, *The Lexical Syllabus* (Willis, 1990) and *The Lexical Approach* (Lewis, 1993; 2000). These approaches have sought to challenge the dominance of the structural paradigm in language teaching and assign greater prominence to frequently occurring chunks of language.

Corpus linguists have often argued that the use of corpora *by teachers and students* "can significantly enrich the pedagogic environment" (Aston, 1995, p. 261), but the question remains *how* this might be done. One practical suggestion is that students can act as 'researchers' in the classroom by using authentic corpus data to identify language patterns – an approach termed Data Driven Learning (see Johns, 1997). The ever increasing availability of easy to use online corpora[1] certainly removes some of the practical difficulties of this. Text frequency profiling tools[2] also enable students and teachers to take a more empirically based approach to vocabulary learning/teaching. In the classroom parallel corpora have also been used, to a limited degree. Chujo et al. (2006), for example, report on the use of a Japanese-English parallel corpus by learners as a tool for vocabulary acquisition. Despite these advances, we agree with Mukherjee (2006, p. 20) when he suggests that there often appears to "a gap between what applied corpus linguistics has to offer and what teachers actually do (or don't do) with corpora in their teaching practice."

In seeking to address this gap, this article shows how the compilation of learner corpora by teachers on a *local* level might help realise some of the benefits that Aston suggests. Such corpora compiled in a specific context to address questions specific to a particular group of learners we refer to as *local learner corpora* – a term put forward by Seidlhofer (2002). After providing an overview of how teachers can go about creating a local learner corpus in a relatively 'quick and dirty' way, a description of the necessary tools for analysis is given. The following section offers an outline of the basics of how teachers can approach the analysis of their corpus with examples from a local learner corpus created at a university in Japan. Finally, suggestions are offered on how local learner corpus data can be used by both teachers and students in a meaningful way. Where possible, references to a wealth of freely available online resources are cited. Useful online resources are listed at the end of the article.

1 For example, those hosted at Brigham Young University – see online resources.

2 Such those available on The Compleat Lexical Tutor website – see online resources.

2. From student writing to student corpus

Planning what and how much to include in a corpus are indispensible first steps. As most teachers receive a constant flow of student texts for marking, these make the most practical departure point for creation of a local learner corpus. Because corpus data needs to be in electronic format, having students submit written work in electronic format is obviously desirable – although transcription is not completely without merit (for example, it allows the simultaneous standardization of spellings). Basic information about the text (in the language of corpus linguists, *mark-up*) is invaluable to structuring the texts in a meaningful way – for example, who wrote it, the class, the type of writing task or whether reference tools were used. The simplest way, of doing this is by naming files according to pre-determined system based on some alphanumeric code. Further information about analytic features of the text might also be added during this preparation stage – what is known as *annotation*. A simple but useful type of annotation, especially for essays, is the addition of tags indicating the start of paragraphs - usually enclosed in angled brackets. Thus, paragraph one could be preceded by <P1>, paragraph two by <P2>, and so on. Files should be saved in text format (i.e. with the file extension .txt) - when dealing with a large number of files saved in Word format (.doc), a file conversion program can prove useful. Each piece of student work should ideally be saved as an individual file. In combination with a meaningful naming system, a large number of files can thus be re-sorted to compile smaller sub-corpora for more specific comparisons. File merging software can then be used to combine sub-corpora into single files representing groups to be compared (for example, student writing at the start of the course and at the end). Readers should note that the overview of corpus compilation outlined above is brief and superficial. While this does suffice for many of the applications proposed in this article, for those keen to read in detail about issues surrounding corpus compilation, Wynne (2005 – available online) provides a comprehensive and accessible guide to good practice in developing linguistic corpora.

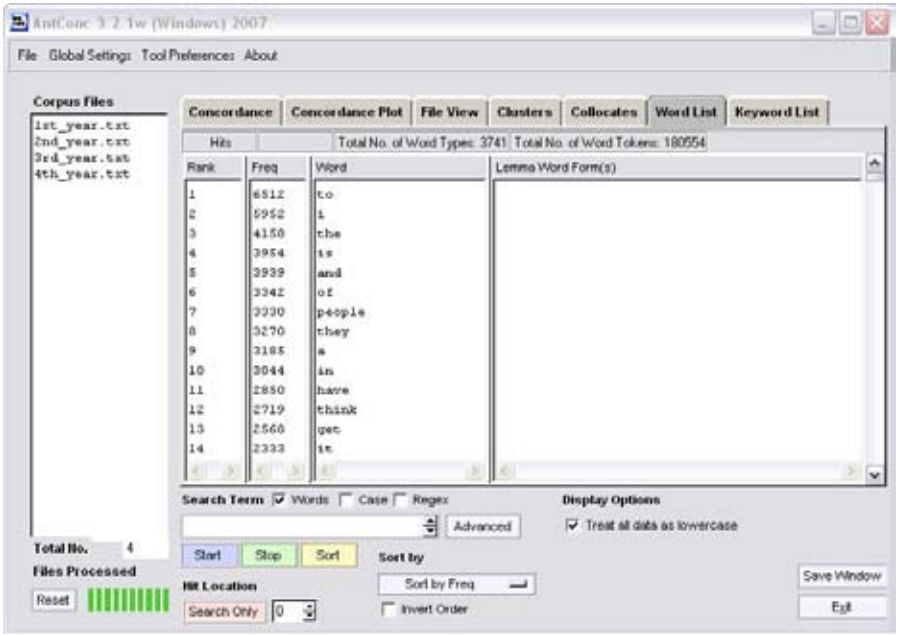
Needless to say, meaningful analysis of a corpus can only be carried out with specialised software. There is a variety of *concordancers* (corpus analyse software) available – some free and others not. *AntConc* is one of the best *free* concordancers available. Developed by Laurence Anthony and available for download from his website (see *Online resources*), AntConc has excellent functionality and is very easy to use. Popular commercial programs include *WordSmith Tools* and *MonoConc Pro* – see David Lee's meta-site (see Online resources) for an overview of available software. A concordancer, in conjunction with a simple spreadsheet, such as Excel (in which most teachers are already well versed), is all that is required to perform a wide range of analyses.

4. Corpus analysis – the basics

Most concordancers offer a range of features that may not necessarily be of immediate interest to teachers – therefore, only the most basic are discussed here: word lists and concordances. The *word list*, the most elemental of corpus analyses, is a useful departure point for exploration. This, quite simply, is a list of which words occur in the corpus and how often. Corpus linguists however make an important distinction with regard to what constitutes a word – *types* refers to distinct word forms and *tokens* to occurrences of *all* words irrespective of form. The local learner corpus considered here for the purposes of illustration comprises just over 960 essays.

These were written by students majoring in English at a Japanese university[3] under exam conditions and cover all four years of study. Figure 1 shows a screenshot of the word list in AntConc for this learner corpus. Individual *types* are listed in the central main column adjacent to their frequencies, while the *total number of types* (3,741) and *tokens* (180,554) are displayed above this.

Figure 1: Screenshot of a Word List in AntConc



Such frequency lists become of interest when compared across corpora. In the case of the example above, it may be of interest to compare it to a 'bench-mark' *reference corpus* to see how learner writing differs from a target 'norm'. Bear in mind however, that in order to compare two corpora, we need to *normalise* frequencies to ensure that they are comparable – this is done by converting *raw* frequencies, those listed in the central '*Freq*' column, to percentages of the total number of tokens. This is easily calculated by saving the frequency list as a text file (under *File*) and importing the data into an Excel spreadsheet to perform the simple calculation – i.e. raw frequency divided by total number of tokens and then converted to a percentage (words-per-million is also commonly used by corpus linguists). The question of what constitutes a suitable reference corpus is, needless to say, a thorny topic (Barlow, 2005, p. 345). In particular, it is open to question whether native speaker production (often performing a different task) is an appropriate model for learners. Despite these issues, frequency-lists of various native

3 The local learner corpus was created by the authors at Kanda University of International Studies in Japan – a four year university specialising in foreign languages.

speaker corpora, such as the BNC (see David Lee's meta-site), can serve as a useful *starting point* for investigation of learner corpora.

Eyeballing two frequency lists to look for differences is however limited. A more rigorous method for comparison of differences, which generates another type of word list, is a *keyword analysis*. This involves automatic comparison of frequencies in the corpus in question with frequencies in a reference corpus (loaded under *Tool Preferences* settings) to generate a statistical measure of the differences – log-likelihood (see Dunning, 1993). The software generates a list of *positive keywords* – these are words with an unusually high frequency in comparison to the reference corpus. *Negative keywords*, on the other hand, are those with an unusually low frequency. It is worth noting that for teachers involved in materials creation, keyword analysis is a powerful tool for identification of vocabulary characteristic of a particular set of texts. This, in combination with vocabulary profiling, can be useful for generation of meaningful vocabulary lists (see Millar & Budgell, 2008).

Figure 2: Screenshot of a Keyword List in AntConc

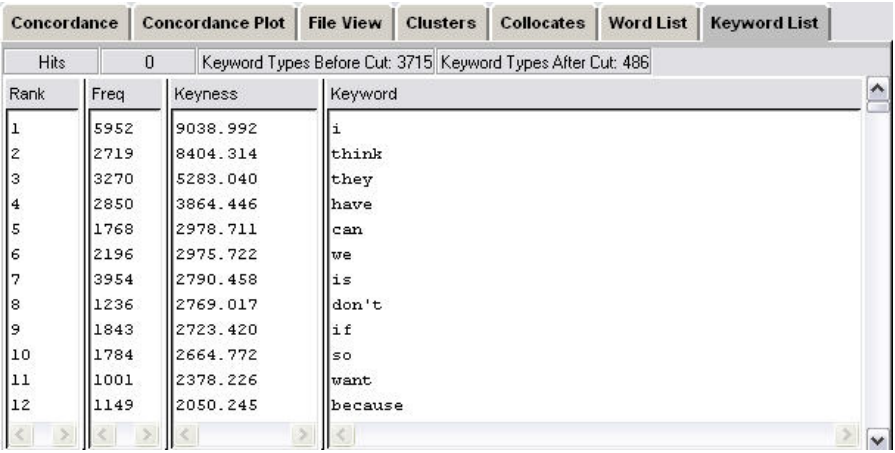


Figure 2 shows a screenshot of positive keywords in our example learner corpus (excluding content words occurring in the essay prompts[4]). Looking at the key words, one can see that the first person pronouns the verb *think* and conjunction are heavily overrepresented in student writing – no surprises there for language teachers. Word list data, such as this, can serve a useful departure point for more 'focused' analyses. These involve searching the corpus, much in the same that one carries out a Google search on the Internet. Bear in mind that corpus software will only return hits for *exactly* what you specify in the search term – therefore, as with the Internet searchers, wildcards, such as 'any character' or 'OR' can be useful (see *Global Settings* in AntConc). Results are returned in the form of a *concordance line* – a list of all the hits with the

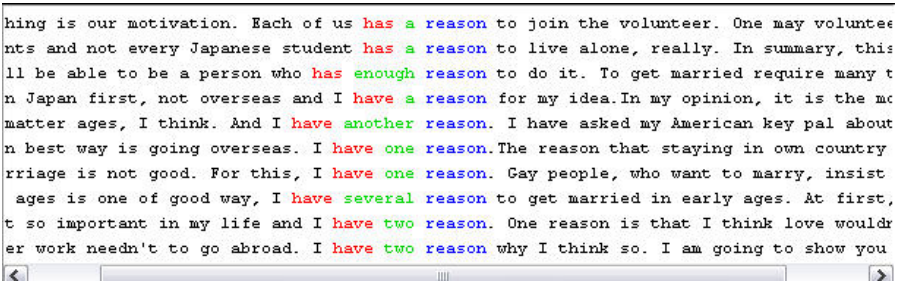
4 Content words occurring in the essay prompts were added to a *stop-list* – a list of words which the program will ignore during analysis. This can be accessed under *Tool Preferences* in AntConc.

immediate context in which the word occurs (a set number of characters to either side of the search term). This can be sorted to identify patterns, and, where necessary, quantified.

For example, we could investigate further student use of the first person pronoun. In our corpus, findings show that over a third of the occurrences are in combination with *think/ don't think* (making it by far the most frequently recurring phrase in the corpus). Making use of punctuation (periods) and paragraph tags (i.e. <P I >) as search terms, we can identify how learners start sentences – i.e. the words that occur to the right of these terms. In the example corpus, first person pronoun is used to start about six in every ten essays. It is also the first word of over twenty percent of all sentences. Expanding this analysis of how students start sentences, we can see that *but*, *so*, *and* and *because* account for over seventeen percent of all sentence initials, and are more often than not used grammatically incorrectly. Investigation of contractions shows that students have an overwhelming preference for contracted forms (*don't*, *can't*, *I'm* etc.) over the non-contracted equivalents - by over ten times in the case of *don't*. Similarly, negative keywords can serve as a starting point for explorations of learner writing. They show, for example, the underrepresentation of articles (again no surprises). What is however interesting is that past tense forms (*was*, *had* and *were*) are much underrepresented in student writing – indeed they don't even rank in the one hundred most frequent words in learner writing. Examination of concordance lines appears to confirm our teacher intuitions that tense inconsistency (students reverting to the present tense while writing about a past event) plays a considerable role in the lack of past tense markers.

Teacher intuitions and observations are likely to form the basis of many of the analyses of student writing. After noticing that students are, for example, consistently misusing a certain word, or translating from the first language (Japanese in the case of our corpus), concordance analysis can give the teacher a better understanding of how students are using that particular feature. For example, although we observe that in essays student use of the word *reason* is often 'odd', determining the exact nature of the problem is not easy. Analysis of concordance lines is however of assistance in showing the following: (1) there are frequent problems with word grammar (e.g. *reason of*); (2) it is often used in vague non-target-like combinations with only one clause, such as *I have some reasons* or *The reason is X* (see Figure 3); and (3) it is rarely (never?) used in any of the most common native-like collocational patterns, such as, *there is no reason why* or *for the simple reason*. Data such as this can be of value in addressing these very same issues, a point that we illustrate in following section.

Figure 3: Screenshot of a concordance line in AntConc



So far we have discussed comparison of a local learner corpus to external measures and searching within the corpus as a whole. It is also possible to compare different sections of the learner corpus against each other. The observant readers might have noticed that the corpus files displayed in Figure 1 (left hand column) represent years one to four. As our local learner corpus contains student writing from all four years of students, essays have been 'batched' according to year of study. This can enable a semi-longitudinal analysis of how students' writing develops over the course of four years – here the *Concordance Plot* in AntConc function can be helpful in returning number of hits per file. Analysis of, for example, the first person pronoun and conjunctions (*but*, *so*, *and* and *because*) shows that their use (normalised frequencies) falls considerably over the four years. The use of contractions also decreases. Length of learner essays (in number of tokens) increases steadily over four years, as does the average length of sentences. Such evidence, although very general, is encouraging as it appears to point to improvements students appear to be making in their writing[5].

4. Putting local learner corpus data to use – some suggestions

How might data from a local learner corpus be of value to language teaching? Language teachers, especially those with experience in Japan, might say that they *know* that students are likely to have difficulties of the type discussed in the examples above. The question is, therefore, *how* these can be addressed. We would like to suggest two ways in which local learner corpus data can be of practical value to teachers and students.

Even though corpus data might provide a picture of learner writing that confirms intuitions, teacher reflection on quantitative evidence of students' performance is, we believe, of value. Empirical evidence of student production can encourage teachers to reflect on their own teaching, investigate and, where possible, address the causes of phenomena observed in the corpus. For example, results from our local learner corpus point to general lack of awareness of academic writing style and essay structure across *all* years of study (for example, overuse of the first person pronoun, preference for contracted forms, and the restricted range of sentence initials). In our local context it has encouraged reflection on the writing curriculum, in particular reassessing competing priorities – namely, the need for balancing development of learners' fluency with the need for more activities focused on accuracy and complexity. It can, therefore, represent a potentially powerful tool for classroom-based action research – a point also made by Seidlhofer (2002) and Mukherjee and Rohrbach (2006).

Secondly, both frequency data and concordance lines from local learner corpora can be put to good use in the classroom – perhaps a more practical incentive for teachers to create corpora of their students' writing. The rationale for this is based on the notion of *noticing the gap* – the argument put forward by Schmidt (1990) that for input to become intake, learners must make a comparison between their own interlanguage and the input and consciously 'notice the gap' between the two. Some of the issues related to learner writing that we touched upon in the examples above (for example, over use of *I think* or lack of target-like collocational use of the word *reason*) are notoriously difficult to address in the classroom – in fact, often teachers often find it difficult to put their finger on what exactly the problem is. By using authentic

5 Of course in addition to descriptive statistics of the type outlined here, in order reach empirically sound conclusions, inferential statistics are needed. Discussion of inferential statistics is beyond the scope of this article.

learner production in the classroom it is possible to foster conditions for students to notice how their own production is different from that of a model. With regard to frequency data, just as corpus-based dictionaries have usage 'warnings' based on common errors in extensive learner corpus data (for example, the Longman Activator series), there is no reason why local learner corpus data should not be put to good use in the classroom. In materials focusing on essay structure and style, it would seem helpful to have a 'warning' box stating, for example, that students frequently overuse *I think* when starting an essay or frequently misuse contractions. The use of learner concordances in the classroom by students also has great potential for teaching. Mukherjee suggests that local learner corpora can be analysed by both teachers and students: "learners can use a concordance display of their own mistakes as a starting point for data-driven learning activities" (2006, p. 18). To clarify how this might actually be applied, we will provide an example of materials that are currently being piloted based on this idea. The example addresses the word the problems that learners have with use of a single word *reason* (as discussed above), although materials have been developed to address a range of issues in learner writing. The materials follow these four main stages:

1. First students are presented with pre-selected and sorted concordance lines of native speaker usage (from the BNC). From these students, identify the recurrent grammatical pattern (Figure 4.1). Students then produce these patterns in some simple exercises (not shown).
2. In the following stage students compare concordance lines of the same feature from the learner corpus (Figure 4.2), which have been pre-selected and sorted to illustrate problematic usage. Questions guide students to identify differences between learner usage and target-like usage.
3. The subsequent stage involves drawing students' attention to frequent native speaker collocation patterns using the word *reason* (Figure 4.3) and some production exercises.
4. Finally students are referred to their own writing portfolio to compare how they have used the word *reason* in their writing.

The materials, thus, follow a path which encourages learners to notice any gap between their own writing and a model, expand on that model and reflect on their own writing.

Figure 4.1: Identifying basic native speaker patterns

Figure 4.2: Comparison to learner patterns

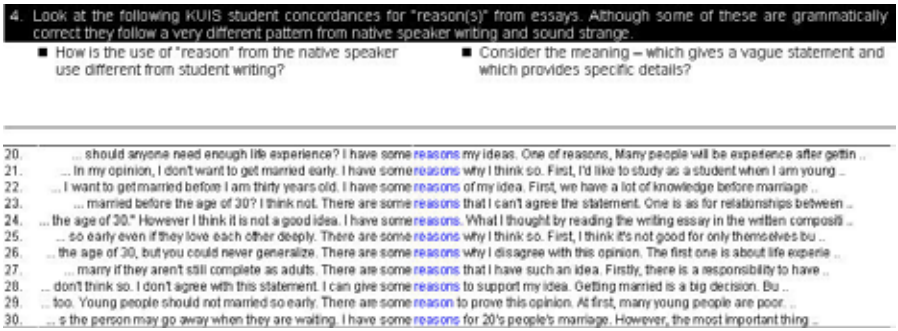
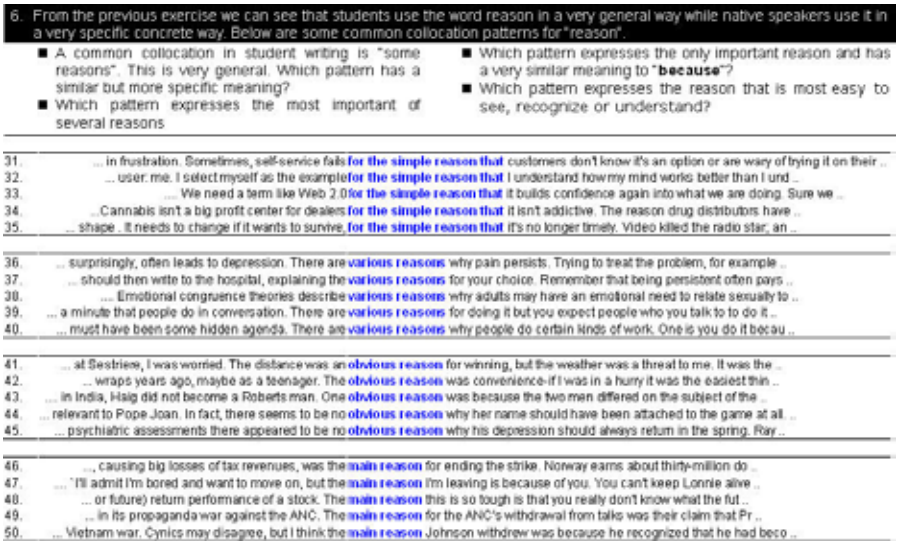


Figure 4.3: Expanding collocational knowledge



Of course, materials can also take a much less structured format. For example, a simple print out of concordance lines of a certain feature of student writing can constitute a departure point from which students explore online reference corpora (such as those listed in Online resources). In such tasks students can be encouraged to compare native speaker collocational patterns with those in their own writing. Alternatively, learner concordance lines can simply be transferred to an OHP transparency and used for class discussion in feedback session. It should be stressed again that DIY corpus creation *can*, depending on the intended use, be carried out in a 'quick and dirty' way (for example, creation of a small corpus based on a class assignment) – after that, preparation of concordance lines requires little time.

6. Conclusion

Although some teachers may be aware of advances in corpus linguistics, it seems that for most, it remains the remote pursuit of researchers far removed from practical issues of the language classroom. However, in this article we have illustrated how corpora can be complied and analysed on a local level to form the basis of classroom-based action research. Furthermore, we have also shown how direct use of learner corpus data in the classroom can be used to create materials to address the attested needs of specific groups of learners. This is a particular strength of local learner corpora and is in contrast to frequently used 'global' materials, which can often be based on the "fuzzy, intuitive, non-corpus-based needs of an archetypal learner" (Granger, 1998, p. 7). Therefore, creation of local learner corpora has a lot to offer: an accessible, hands-on resource for teachers and learners with the potential of greatly enriching the pedagogic environment and bridging gaps between what corpus linguistics has to offer and what teachers do (or don't do) with corpora in their teaching.

Online resources

1. David Lee's meta-site: http://devoted.to/corpora
David Lee's site provides an extensive collection of web-links related to all aspects of corpus-based research.
2. Corpora at Brigham Young University: http://corpus.byu.edu
This site provides free access to a range of large online corpora through an easy to use, fast and innovative interface developed by Mark Davies. In addition to the British National Corpus (100 million words), the recently released TIME corpus (100 million words) and BYU Corpus of American English (360 million words) are available.
3. Compleat Lexical Tutor: http://www.lextutor.ca
This site provides a suite of web-based corpus tools, created by Tom Cobb, geared to Data Driven Learning.
4. Laurence Anthony's Homepage: http://www.antlab.sci.waseda.ac.jp
This is the website of the developer of AntConc, Laurence Anthony, from which this, and other software, can be downloaded.

References

Aston, G. (1995). *Corpus evidence for norms of lexical collocation*. Oxford: Oxford U.P.
Barlow, M. (2005). Computer-based analyses. In R. Ellis & G. Barkhuizen (Eds.), *Analysing learner language* (pp. 335–369). Oxford: Oxford University Press.
Chujo, K., Utiyama, M., & Miura, S. (2006). Using a Japanese-English parallel corpus for teaching English vocabulary to beginning-level students. *English Corpus Studies, 13*, 153–172.
Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19* (1), 61–74.
Fulcher, G. (1991). Conditionals revisited. *ELT Journal: English Language Teachers Journal, 45*(2), 164–168.
Granger, S. (1998). *Learner English on computer*. London; New York: Longman.

Granger, S. (2004). Computer learner corpus research: Current status and future prospects. In U. Connor & U. T. A. (Eds.), *Language and computers: Studies in practical linguistics* (pp. 123–145). Amsterdam & Atlanta: Rodopi.

Johns, T. (1997). Contexts: The background, development and trialling of a concordance-based call program. In A. Wichmann, S. Fligelstone, A. McEnery, & D. Knowles (Eds.), *Teaching and language corpora* (pp. 100–115). London; New York: Longman.

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. Hove: Language Teaching Publications.

Lewis, M. (2000). *Teaching collocation: Further developments in the lexical approach*. Hove: Language Teaching Publications.

McCarthy, M., & Carter, R. (1995). Spoken grammar: What is it and how can we teach it? *ELT Journal: English Language Teachers Journal, 49* (3), 207–218.

McCarthy, M., McCarten, J., & Sandiford, H. (2005). *Touchstone*. Cambridge, UK; New York: Cambridge University Press.

Millar, N., & Budgell, B. (2008). The language of public health – a corpus-based analysis. *The Journal of Public Health, 16* (5), 369–374.

Mukherjee, J. (2006). Corpus linguistics and language pedagogy: The state of the art – and beyond. In S. Braun, J. Mukherjee & K. Kohn (Eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods* (pp. 5–24). Frankfurt: P. Lang.

Mukherjee, J., & Rohrbach, J.-M. (2006). Rethinking applied corpus linguistics from a language-pedagogical perspective: New departures in learner corpus research. In B. Kettemann & G. Marko (Eds.), *Planing, gluing and painting corpora: Inside the applied corpus linguist's workshop* (pp. 205–232). Frankfurt am Main: Peter Lang.

Nesselhauf, N. (2007). The path from learner corpus analysis to language pedagogy: Some neglected issues. *Language and Computers, 62*, 305–315.

Pravec, N. A. (2002). Survey of learner corpora. *ICAME Journal, 26*, 81–114.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11* (2), 129–158.

Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learning-driven data. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 213–234). Amsterdam: John Benjamins.

Timmis, I. (2005). Towards a framework for teaching spoken grammar. *ELT Journal: English Language Teachers Journal, 59* (2), 117–125.

Willis, D. (1990). *The lexical syllabus. A new approach to language teaching*. London and Glasgow: Collins.

Wynne, M. (2005). Developing linguistic corpora: A guide to good practice. Oxford: Oxbow Books. Accessed April 24, 2008, from http://ahds.ac.uk/linguistic-corpora/

Author Biodata

Neil Millar has taught English in Japan, New Zealand and the UK. He is currently studying for a PhD in applied linguistics at Lancaster University where he also is co-coordinator of the Corpus Research Group (Departments of Linguistics and Computing). Research interests focus on the practical applications of corpus linguistics.

Ben Lehtinen is a lecturer at Kanda University of International Studies in Chiba, Japan. His interests include second language writing, curriculum design, interactional analysis and the development of academic writing centres.