Applying corpus linguistics to pedagogy
A critical evaluation*

Lynne Flowerdew
Hong Kong University of Science and Technology

This article reviews and discusses four somewhat contentious issues in the application of corpus linguistics to pedagogy, ESP in particular. Corpus linguistic techniques have been criticized on the grounds that they encourage a more bottom-up rather than top-down processing of text in which concordance lines are examined atomistically. One criticism levelled against corpus data is that a corpus presents language out of its original context. For this reason, some corpus linguists have underscored the importance of ‘pedagogic mediation’ to contextualize the data for the students’ own writing environment. Concerns relating to the inductive approach associated with corpus-based pedagogy have also been raised as this approach may not always be the most appropriate one. A final consideration relates to the issue of whether a corpus is always the most appropriate resource to use among the wealth of other resources available.

Keywords: top-down, decontextualisation, pedagogic processing, inductive, data-driven learning

1. Introduction

Corpus linguistics is usually associated with a phraseological approach to analysis, which takes a syntagmatic, as opposed to a purely paradigmatic, view of language. Corpus analysis, in fact, gives both a paradigmatic and syntagmatic view of language as concordance output can either be “read” vertically, i.e. paradigmatically, in line with the slot and filler notion espoused by substitution tables, or horizontally, i.e. syntagmatically, from a phraseological perspective. Following Sinclair (1999, 2004a) the lexical item has primacy, with its core meaning and semantic prosody as obligatory categories, and collocation, colligation and semantic preference considered as optional categories. An interweaving of some or all of these categories gives what Sinclair refers to as an ‘extended unit of meaning’, although in later work Sinclair (2004b: 280) extends this concept to the ‘maximal approach’ which:
“… would be to extend the dimensions of a unit of meaning until all the relevant patterning was included — all the patterning that was instigated by the presence of the central word. …[W]e should extend the unit until the ambiguity disappears”. In this phraseological approach recurring patterns in concordance output have revealed how language can follow certain tendencies according to Sinclair’s notion of an extended unit of meaning’ rather than being bound by hard-and-fast rules.

By way of example, Danielsson (2007: 18) presents a methodology using raw frequency data as a step towards the identification of meaningful units, for which as she points out “there are no accepted answers to simple questions such as ‘What exactly constitutes a multi-word unit?’ or ‘Where does a multi-word unit begin and end?’” Taking the word jam as the node, Danielsson identified its most frequent collocates in the BNC, with traffic the top collocate. Concordance lines for the node and its collocate were then generated, with a span of 4+4 words either side of the node. This investigation showed a to be the most frequent collocate in the lines including both traffic and jam, e.g., stuck in a traffic jam with your pulse; man in a traffic jam who curses (ibid.: 19). A further calculation was made to find out the most frequent collocate in the lines for traffic and jam, with the second collocate a generated. This showed in to be the next most frequent collocate, e.g., stuck in a traffic jam, you might reflect (ibid.: 20). This procedure was repeated until no other collocates were found to occur above the cut-off point of 5 (an arbitrary point that, Danielsson admits may have to be revised). The search ended with the unit stuck in a traffic jam: “From here, no other collocates occur with sufficient frequency to reach the cut-off point, and we may claim to have achieved the maximal unit [my italics] based on the distribution in this corpus” (p. 20). Danielsson then considers the paradigmatic axis testing each word to see if the unit allows for any alternatives, finding that the corpus offers sitting, waiting and caught as alternatives to stuck. It is to be noted that these alternatives reflect the syntagmatic nature of semantic prosody (the frequent co-occurrence of a lexical item with items expressing a positive or negative evaluation) and semantic preference (the frequent co-occurrence of an item with those from a particular semantic set): “In this particular context, the words seem to be related to offer a set of verbs that create a feeling associated with the annoying event of being held up in traffic” (p. 20).

Some researchers (cf. Tognini-Bonelli 2001) see this phraseological approach associated with corpus linguistics as a catalyst in redefining aspects of linguistic theory. This is regarded as the corpus-driven approach (as opposed to the corpus-based one), in which the data is approached without any pre-conceived notions in relation to how it should be analysed. Other corpus linguists take a less extreme approach; for example, McEnery et al. (2006: 6), while considering corpus linguistics as “a new philosophical approach to linguistic enquiry” with its own theoretical status, do not view it as a discipline in its own right with its own theory,
but rather as a methodology. The same position can be applied to data-driven learning (DDL); it is generally agreed there is no underlying theory as such, but rather it rests on a methodology which can uncover facts about language hitherto unexplored.

Notwithstanding the advantages of this approach for DDL, during the last few years some accounts in the literature have adopted a more critical stance, drawing attention to potential drawbacks of using corpora in DDL. This paper reviews the following key issues in the debate on applying corpus linguistics to pedagogy.

– Corpus linguistic techniques encourage a more bottom-up rather than top-down processing of text in which truncated concordance lines are examined atomistically.
– Corpus data are decontextualised and, for this reason, may not be directly transferable to students’ own context of writing.
– Corpus-based learning is usually associated with an inductive approach to learning, in which rules, or indeed patterns, are derived from multiple examples, rather than a rule-based deductive approach.1 This approach might not be the most appropriate choice for some students.
– There are different types of corpora (general, specialized, learner) and different types of online resources (dictionaries, grammars). Students may have difficulty in selecting the most appropriate corpus and resource for a particular query.

The issues outlined above are not in fact discrete issues but inter-related, as the following discussion shows. They are examined specifically with reference to corpora of written text.

2. Corpus linguistic techniques encourage a bottom-up processing of text

Corpus linguistic techniques have been criticized for encouraging a more bottom-up rather than top-down processing of text in which truncated concordance lines are examined in a somewhat atomistic fashion without recourse to the overall discourse (Swales 2002, 2004). Like Swales, Kaltenböck & Mehlmauer-Larcher (2005:71) have expressed similar sentiments: “There are, however, certain parts of a text that even a concordancer cannot reach. These are aspects of the macrostructure of a text, such as textual moves, i.e. a unit of text that expresses a specific communicative function”. However, in the last couple of years corpus linguistics research has paid much more attention to these two different modes of text processing (Flowerdew 2003, 2005, forthcoming a), with Biber et al. (2007a) explaining the concept behind these two different, yet complementary, approaches thus:
In the ‘top-down’ approach, the functional components of a genre are determined first and then all texts in a corpus are analysed in terms of these components. In contrast, textual components emerge from the corpus analysis in the ‘bottom-up’ approach, and the discourse organization of individual texts is then analysed in terms of linguistically-defined textual categories.

(Biber et al. 2007a: 11)

These two different starting points can be illustrated by reference to the studies of Kanoksilapatham (2007) and Jones (2007). Kanoksilapatham (2007) in her corpus-based examination of rhetorical moves in biochemistry research articles commences with a top-down analysis by first developing an analytical discourse-based framework through identifying the move types that can occur in each section of biochemistry research articles, before embarking on the corpus analysis. A corpus analysis, using Biber’s multi-dimensional analysis, was subsequently carried out to determine the linguistic characteristics of different rhetorical moves. Jones’ (2007) research, on the other hand, begins with a corpus analysis to identify the linguistic characteristics of vocabulary-based discourse units (VBDUs), using software which specifically highlights new words not found in the preceding adjacent stretch of discourse. For example, in the methods sections of biochemistry articles, new linguistic items such as was stored, was extracted, denoting “Procedural description of past actions”, were found.

As Biber et al. (2007b) point out, the different starting points of a top-down functional analysis vs. a bottom-up linguistic analysis will yield differences: “Fifteen different move types were identified in the analysis of biochemistry research articles…. In contrast, only 6 different discourse types were identified in the VBDU study” (p. 249). However, at the same time “the inherent structure of a genre would be reflected in analyses undertaken from both perspectives” (ibid.: 249), as there are areas of overlap in that the VBDU type of “Procedural description of past actions” can be mapped onto one of the moves in the Methods section “Describing experimental procedures”.

It is generally acknowledged that exploitation of corpus linguistics findings takes time to percolate through to pedagogic applications (Braun 2005; Frankenberger-Garcia 2006). Interestingly, it is in the area of English for Specific Purposes where corpora are now taking on an increasingly mainstream role (Belcher 2006; Flowerdew, forthcoming b) with the compilation of small “localised” corpora often compiled by the class tutor, or sometimes the students (Flowerdew 2004). Indeed, this is evidenced by the studies outlined below which report a more discourse-based pedagogic application of corpora, combining top-down and bottom-up approaches to text analysis.
2.1 Moving from top-down to bottom-up processing

Charles (2007), mediates between top-down and bottom-up processing of academic thesis writing in order to introduce her students to the rhetorical function of “defending your work against criticism”, a two-part rhetorical pattern, “in which the writer first concedes the possibility of criticism and then moves to neutralize its potentially negative effect” (p. 289). Charles achieves this by first having her students engage in initial macro discourse-based tasks as consciousness-raising activities of the discourse pattern. In this initial stage, students are introduced to the rhetorical function, its purpose in the text and the different ways in which it can be realised. Students discuss the insights that they have gained in their group discussion in a whole-class feedback session. Charles then moves from this more top-down to bottom-up analysis by having students perform corpus searches which focus on specific lexico-grammatical structures within a discourse-based framework. For instance, students concordance on salient (in the sense of noteworthy) items with the aim of formulating a generalization about the use and positioning of while in constructing a concession. Students are asked to examine the lexico-grammar for constructing a concession, noting that while co-occurs with acknowledge, as in Table 1 below, or that it may also appear in the context of appear/seem/may, e.g. While this may seem contradictory to the conclusions drawn above..

Table 1. Anticipated criticisms and writers’ defence (from Charles 2007: 294)

<table>
<thead>
<tr>
<th>Extract</th>
<th>Anticipated criticism</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.</td>
<td>Although the results of the experiments are not conclusive,</td>
</tr>
<tr>
<td>B.</td>
<td>There is <strong>nothing essential in these categories</strong>, and they <strong>may not appear tenable to other scholars</strong>.</td>
</tr>
<tr>
<td>C.</td>
<td><strong>While I acknowledge</strong> that in some cases the distinction between institutions and groups may seem rather arbitrary,</td>
</tr>
<tr>
<td>D.</td>
<td>Unfortunately, specimen preparation is especially laborious for the completed device structure, which meant that</td>
</tr>
</tbody>
</table>

A more top-down approach, similar to that adopted by Charles, is also found in Weber’s (2001) materials aimed at law students. First, Weber’s students were inducted into the genre of legal essays by reading through whole essays taken from the University of London LLB Examinations written by native speakers, and identifying some of the prototypical rhetorical features, e.g. identifying and/or delimiting the legal principle involved in the case. They were then asked to identify any lexical expressions which seemed to correlate with the genre features. This was followed up by consulting the corpus of the legal essays to verify and pinpoint regularities in lexico-grammatical expressions.
Another corpus linguistics practitioner who commences from a top-down perspective is Noguchi (2004). In fact, Noguchi had her science and engineering majors build their own mini-corpora of research journal articles first before classifying the genre features and then moving on to examining prototypical lexico-grammatical features.

It has been noted that Swales (2002: 163) has contrasted the “fragmented” world of corpus linguistics with its tendency to adopt a somewhat bottom-up, atomistic approach to text with the more “integrated” world of ESP material design with its focus on top-down analysis of macro-level features. However, Weber’s tasks, and those by Charles and Noguchi, seem to be achieving a “symbiosis” between these two approaches, as called for by Partington (1998: 145).

2.2 Moving from bottom-up to top-down processing

Hyland’s (2007, 2008) work on genre-specific phraseological routines commences from a bottom-up perspective. Hyland (2007) tabulates the most frequent 50 4-word bundles across four disciplines (biology, electrical engineering, applied linguistics, business studies), noting the great extent to which these are specific to particular disciplines. These bundles are then classified into three broad foci of research, text and participants, as outlined below (Hyland 2007: 13–14):

**Research-oriented** – help writers to structure their activities and experiences of the real world, e.g.:
- **Procedure** (*the use of the*, *the operation of the*)
- **Quantification** (*the magnitude of the*, *the surface of the*)

**Text-oriented** – concerned with the organization of the text and its meaning as a message or argument, e.g.:
- **Structuring signals** — text-reflexive markers which organize stretches of discourse (*in the present study, in the next section*)
- **Framing signals** – situate arguments by specifying limiting conditions (*in the case of*, *with respect to the*)

**Participant-oriented** – these are focused on the writer or reader of the text, e.g.:
- **Engagement features** – address readers directly (*it should be noted that*, *it can be seen*)

Hyland (2004: 220) notes the high productivity of the bundle *the* *of*, which he argues justifies its inclusion in courses assisting students to write effective academic papers in the sciences. For example, two frequent bundles in biology, *the presence of* and *the splicing of* “would seem to offer students valuable forms for expressing meanings relating to existence and to research processes in their writing” (ibid.).
Pedagogic applications with a bottom-up starting point moving to more top-down processing have also been noted by Flowerdew (2006). Besides lexical bundles, another type of phraseological routine is collocations. Usually, collocations are considered as word combinations, one of the most common being adjective + noun, which is a pairing of particular difficulty for advanced learners of English (Nesselhauf 2003, 2004). However, such collocations can also be involved in more top-down processing of text. For example, Flowerdew (2006) notes that in a module on business letter writing, students were not sure which adjective from a set of seemingly semantically synonymous adjectives was the “right” one to choose in the following sentence.

Thank you for your kind / sincere / cordial invitation to the alumni dinner.

A search on these different combinations in a Business Letters Corpus revealed the following patternings, shown in Figures 1 and 2 below.

hoping, in fact, that you will accept our cordial invitation to be our guest for the length of
May we extend to you a cordial invitation to call in at White’s and make the
Please accept our cordial invitation to visit and become acquainted with
have the pleasure in extending to you our cordial invitation to visit our organization at a data
and I shall be pleased to extend to you my cordial invitation to visit our Tokyo office at your

Figure 1. Selected concordance lines for cordial + invitation

I very much appreciate your kind invitation to join the University Club and I know
Thank you very much again for your kind invitation and I hope your conference will be a
I am therefore very happy to accept your kind invitation and look forward to attending a great
as if I shall be unable to accept your kind invitation this time because of a most important
Thanking you once more for your kind invitation to address the audience, I remain

Figure 2. Selected concordance lines for kind + invitation

In order to determine the most appropriate collocation for this context, students were required to look beyond the immediate collocation to an ‘extended unit of meaning’, which takes in the subject + verb and direct and/or indirect object of the sentence. In so doing, students were able to work out that cordial + invitation was used for offering an invitation (May we extend to you a cordial invitation) whereas kind + invitation was used for accepting an invitation or thanking the host (thank you very much for your kind invitation), or as one of my students expressed it: cordial is used from you to me, and kind from me to you. As Aston (personal communication) has pointed out, it is important that students do not just consult the corpus in a phrasebook type fashion but have something more substantial. Having students “read” the corpus paradigmatically to find alternatives to the verbs extend,
accept, thank, and their respective phraseologies would be a way of counteracting this narrow “reading” of the corpus. Stubbs’ (1996: 36) oft-quoted principle “There is no boundary between lexis and grammar: lexis and grammar are interdependent” is also of relevance here. Lexis and grammar have been shown to manifest interdependency; for example, the subject of the intransitive verb set in very often refers to “unpleasant states of affairs” such as bad weather (Sinclair 1991: 74). In a similar fashion, collocations and functions can also be viewed as interdependent, albeit at a more discourse-based level, as evidenced by the above analysis of the collocations of “invitation” with kind and cordial.

Another example of more top-down processing leading on from a bottom-up search is as follows. Flowerdew (2008b) notes that in a course on report writing one student query related to whether the active or passive voice was used in the following sentence:

This project focuses / is focused on the incidence of mosquitoes on campus.

A search on focus was conducted in an institutionally-compiled 7 million-word corpus of reports, which gave the results shown in Figure 3 below.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Left sort</th>
<th>Right Sort</th>
<th>Frequency Sort</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOUN + VERB + PREP</td>
<td>e.g. “study focuses on”</td>
<td>Show results</td>
<td>292</td>
</tr>
<tr>
<td>VERB + VERB + PREP</td>
<td>e.g. “has focused on”</td>
<td>Show results</td>
<td>231</td>
</tr>
<tr>
<td>TO + VERB + PREP</td>
<td>e.g. “to focus on”</td>
<td>Show results</td>
<td>95</td>
</tr>
<tr>
<td>ADV + VERB + PREP</td>
<td>e.g. “not focused on”</td>
<td>Show results</td>
<td>94</td>
</tr>
<tr>
<td>PRON + VERB + PREP</td>
<td>e.g. “we focus on”</td>
<td>Show results</td>
<td>57</td>
</tr>
<tr>
<td>CONJ + VERB + PREP</td>
<td>e.g. “that focus on”</td>
<td>Show results</td>
<td>56</td>
</tr>
<tr>
<td>DET + VERB + PREP</td>
<td>e.g. “which focus on”</td>
<td>Show results</td>
<td>38</td>
</tr>
<tr>
<td>VERB + VERB + ADV</td>
<td>e.g. “has focused almost”</td>
<td>Show results</td>
<td>34</td>
</tr>
<tr>
<td>NOUN + VERB + ADV</td>
<td>e.g. “efforts focused primarily”</td>
<td>Show results</td>
<td>33</td>
</tr>
<tr>
<td>TO + VERB + ADV</td>
<td>e.g. “to focus more”</td>
<td>Show results</td>
<td>31</td>
</tr>
</tbody>
</table>

Figure 3. Search for focus (all word forms) (Flowerdew 2008)

Besides the fact that the students were able to glean the different meanings between the active and passive forms of focus by examining the verb in a wider context, accessed via “Show results” (column three of the Table), I also found that this search encouraged a more top-down processing of text. Students’ scrutiny of the concordance output prompted one student to ask: Why are there so many occurrences of focus in the present perfect? This kind of comment which I have termed a ‘triggered query’ because it is activated by something the student has alighted on in the corpus data, unprompted by the teacher (Flowerdew 2008b), echoes Swain’s
Applying corpus linguistics to pedagogy

(1998) concept of ‘noticing’. Swain (1998: 66) remarks that there are several levels of ‘noticing’, one of which is that: “Learners may simply notice a form in the target language due to the frequency or salience of the features themselves”. An examination of the wider context of the present perfect forms of focus revealed that this tense was used when previous research was introduced, to set up a critical evaluation of this work signalled by however. This discourse-based function of however is therefore being used as a key signalling item in Swales’ (1990) CARS (create a research space) model, opening up a gap for the author’s own research e.g. Much of this cross-cultural work to date, however, has focussed on East Asian versus Anglo comparisons, with little attention given to the issue of cross-cultural differences within the East Asian region.

This type of browsing is thus in the spirit of Bernardini’s philosophy as the ‘learner as traveler’ (Bernardini 2004). Although the type of serendipitous learning advocated by Bernardini (2000, 2002) has been mildly criticized as ‘incidentalist’ (Swales 2002), an example such as the one above illustrates that this ad hoc browsing can encourage students to process corpus data in a much more top-down way. In fact, both Granger (1999) and Hahn (2000) emphasise that the teaching of tenses should be approached from a discourse-based perspective and that a corpus is an ideal medium for achieving this.

Another account of searches extending from bottom-up to top-down processing is reported in Lee & Swales (2006). Their innovative corpus-informed EAP course, entitled “Exploring your own discourse world”, required students to compile their own corpora after working with specialized corpora and conduct more genre-based enquiries. For example, using the BNCweb, students were sensitized to the different discourse environments in which for instance and for example are found.3

... for instance is used a lot more frequently in the social sciences and humanities (where it often introduces casual, non-essential exemplifications of points, mainly for emphasis or color), whereas in the natural sciences for example is clearly favored (being used to illuminate and clarify a difficult or complex point through the exemplification).

(Lee & Swales 2006: 67)

The pedagogic applications reviewed above testify to the fact that traditional classroom corpus-based explorations which tended to centre on a ‘vertical reading’ have now been complemented by a more discourse-based approach which requires ‘horizontal reading’ for the analysis of linguistic patternings in relation to their communicative and cultural embedding (Braun 2005), and one could also add here in relation to the practices of different academic disciplines (see Flowerdew, in press, for further examples of corpus-based discourse approaches to writing
instruction). In fact, Swales has now modified his position and acknowledges this
more top-down orientation, as reported by Lee (2008).

It can be seen that utilizing a more top-down approach to processing cor-
pus data provides more co-text, and hence more contextual information on the
 corpora under investigation by shedding light on different practices of different
academic disciplines, as revealed by differences in lexico-grammatical patterning.
However, whether the starting point should be with a bottom-up or top-down ap-
proach is not an easy question to answer and very much depends on the nature of
the query and composition of the corpus. Starting with the moves (which could
be coded in the corpus) may be appropriate for those genres which have clearly
defined move structures, such as law cases with four obligatory moves: facts/sta-
ting history of the case; presenting argument; deriving ratio decidendi; pronounc-
ing judgment (cf. Bhatia et al. 2004), but difficult to implement for those genres
which are mixed or which display embedded moves (Flowerdew 2004). Biber et
al. (2007b: 241) compare these two different approaches, noting that which one is
adopted depends on the primary basis of the analysis:

Functional analysis is primary in top-down approaches; functional distinctions
are determined on a qualitative basis, to determine the set of relevant discourse
types and to identify specific discourse units within texts. In contrast, linguistic
analysis is primary in bottom-up approaches; a wide range of linguistic distribu-
tional patterns are analysed quantitatively, again being used to determine the set
of relevant discourse types and to identify specific discourse units within texts.
(Biber et al. 2007b: 241)

3. Corpus data are decontextualised and may not be directly transferable

Corpus data have been viewed as decontextualised such that the findings may not
be directly transferable, lock stock and barrel, to pedagogy. This issue is discussed
below with reference to pedagogic applications in the field of ESP.

3.1 The issue of contextualisation in corpus data

Widdowson’s (2004) arguments on the decontextualised nature of corpus data are
well-rehearsed in the literature (see Flowerdew 2008a; Braun 2005; Kaltenböck &
Mehlmauer-Larcher 2005; McEnery et al. 2006), but it is worth reviewing them
again briefly. Both Aston (1995) and Widdowson (1998, 2002) have drawn atten-
tion to the decontextualised nature of corpus data, with Widdowson commenting
that corpus data are but a sample of language, as opposed to an example of authen-
tic language, because it is divorced from the communicative context in which it was
created: “the text travels but the context does not travel with it” (keynote lecture, 29 July 2002).

Whether Widdowson is correct or not would seem very much to depend on what is being transferred. Charles (2007: 295) disagrees with Widdowson on the issue of decontextualisation and maintains that one of the advantages of the type of corpus work described in Section 2.1 above is that “… it allows students to gain a greater sense of contextualization than is possible to achieve through the use of paper-based materials”. While it is undoubtedly true that more top-down corpus enquiries, by their very nature, provide more contextualization, the question of the practices of different academic and professional disciplines needs to be taken into account, as uncovered by the corpus-based enquiries of for instance and for example in Lee & Swales (2006), which show just how finely nuanced differences can be (also see Hyland 2000, 2002 for research studies in this area).

3.2 ‘Pedagogic processing’ of corpus data

Widdowson maintains that it may not be expedient to transfer corpus data directly to pedagogic materials on account of the cultural or contextual inappropriacy of the corpus data (see Cook 1998; Widdowson 1991, also cited in Seidlhofer 2003, for a discussion on the issue of prescription vs. description, regarding the transfer of corpus data to pedagogy). Widdowson therefore advocates adopting some kind of ‘pedagogic processing’, as do other corpus linguists such as Braun (2007) and McCarthy (2001) in order to transform samples of language into pedagogically-accessible examples. This aspect of pedagogic mediation of corpus data is discussed from the perspective of the “what” and the “how” below.

3.2.1 The “what” of pedagogic processing

Section 3.1 has shown that variation across disciplines needs to be considered in the transfer of corpus data to pedagogy. Another aspect that needs to be considered concerns pragmatic appropriacy. Flowerdew (2008a) advises caution on exploiting a corpus of reports, in which consultancy companies are advising external clients, for student report writing which requires them to write internally to university authorities. The student writing is similar to the corpus of reports in respect of the rhetorical Problem-Solution pattern. However, it would not be registerially appropriate for students to transfer the pattern: grammatical metaphor noun (indicating a solution to a problem) + will + verb (signalling mitigation of a problem) (e.g. Implementation of barriers will reduce noise) to their own report writing in view of the different contextual features. Students would need to modify the ‘frame’ (see Biber et al. 2004 and Stubbs 2004 for further examples of frames) derived from the corpus of reports by supplying mitigation devices to attenuate the
phrase to make it socio-culturally appropriate for writing to university authorities. Thus, they would need to expand the original frame with the addition of a prefacing phrase such as “we would like to suggest that…” and replace will reduce by the more rhetorically appropriate would reduce. Corpus consultation has therefore to be conducted with great care and it is not surprising that Widdowson (1998) sees the need for some kind of ‘mediating process’ whereby students authenticate the corpus data to suit the socio-cultural and linguistic parameters of their own writing in light of considerations relating to differences across disciplines and pragmatic appropriacy.

3.2.2 The “how” of pedagogic processing
Having established in the previous sub-section that some type of pedagogic processing may be necessary with some types of data, there still remains the question of how this can be achieved.

In order to integrate the type of pedagogic processing Widdowson is referring to so as to enable students to authenticate the corpus data for their own contextual writing environment, Flowerdew (2008b) has adopted student peer response activities, which draw on Vygotskian socio-cultural theories of co-constructing knowledge through collaborative dialogue and negotiation (see O’Sullivan 2007 who gives a very insightful exposition on the role of cognitive and social constructivist theories to foster corpus consultation literacy). In these peer-to-peer interaction groups, weaker students were intentionally grouped with more proficient ones to foster productive dialogue through ‘assisted performance’, thus drawing on another aspect of socio-cultural theory. In this scaffolding-type of activity, more proficient students were able to offer their insights and interpretations on the corpus data, thus assisting the weaker students to gradually develop more independence. The author reports some success with this approach of incorporating group discussion activities revolving around the corpus data as a form of pedagogic mediation, resulting in consciousness-raising of register awareness, not only for the task in hand, but also what might be appropriate phraseologies for other contexts. Peer discussion also raised issues of what could be transferred from corpus data, i.e. the use of nominalisations such as implementation, which led to further discussion as to whether the gerund, implementing, would also be acceptable, and what would not be appropriate for the context, i.e. the frame “It is recommended that…”, which students mentioned sounded too authoritative. Students were therefore encouraged to engage in “collaborative metatalk” (Swain 1998:68) to “use language to reflect on language use” (ibid.). Gavioli & Aston (2001:242) also advocate spoken interaction among students in corpus consultation as “different learners will often notice different things in concordances, and draw different conclusions”. Suggestions for other types of pedagogic mediation of corpora have been given by Braun...

Pedagogic mediation of corpora could well be assisted through the incorporation of contextual information in written texts to aid the transfer of corpus data to pedagogy. Following Burnard (2004), Krishnamurthy & Kosem (2007) advocate encoding the corpus with metadata to aid subsequent analyses. Although various speech corpora such as the Michigan Corpus of Spoken Academic English, MICASE, have been marked up with metadata categories such as the gender, age range, academic position / role of the interlocutors, these are lacking in corpora of writing. Corpora of business writing are especially context-sensitive and could benefit from the inclusion of such metadata.

However, it should be noted that sometimes the co-textual environment can provide clues to contextual information. In the business letters written by students, the structure and use of appreciate* was found to be particularly problematic across a wide range of students, with learners confusing the active and passive forms, e.g. *I would be much appreciated if …, and omission of the object in the active, e.g. *I would appreciate if…. The Business Letters Corpus referred to earlier proved invaluable for alerting students to the correct structure. What students were unsure of, though, was in which situations the active and passive forms were most appropriate. Here, frequency counts and the co-text in the environment of appreciate* provided valuable clues. The frame …appreciate it if … occurred 105 times, whereas there were only 9 instances of the frame It…appreciated if…, thus suggesting some kind of marked use. In fact, scrutiny of the co-textual environment, i.e. the ‘extended unit of meaning’ revealed that the passive frame would be used when the power relations between the addresser / addressee were quite distinct and when a big favour was being asked. This example thus demonstrates that corpora may not be completely devoid of context, which can sometimes, in part, be recovered from the co-textual environment.

4. Corpus-based pedagogy is usually associated with an inductive approach which may not be appropriate for all students

Both Gavioli (2005) and Meunier (2002) have noted the drawbacks of an inductive approach, in which students extrapolate the rules, or patterning, from examples:

Despite their advantages, DDL activities have some drawbacks…. The various learning strategies (deductive vs. inductive) that students adopt can lead to problems. Some students hate working inductively and teachers should aim at a combined approach (see Hahn 2000 for a combined approach).

(Meunier 2002: 135)
In common with Meunier (ibid.), I also believe that an inductive approach may not appeal to students on account of their different cognitive styles (Flowerdew 2008b). Field-dependent students who thrive in cooperative, interactive settings and who would seem to enjoy discussion centering on extrapolation of rules from examples may benefit from this type of pedagogy. However, field-independent learners who are known to prefer instruction emphasizing rules may not take to the inductive approach inherent in corpus-based pedagogy. It is interesting to note that Vannestål & Lindquist (2007: 343) state that some of the students in their inductive corpus-based grammar course commented that “…they preferred the more traditional way of reading about grammatical rules in the book and did not feel that they learned anything by doing corpus exercises”.

Another reason as to whether an inductive or deductive approach is adopted would very much seem to depend on the nature of a particular enquiry. If the enquiry is based on a grammar rule (for example, the difference between for and since in time expressions; see Tribble & Jones 1990), then the differences are quite clear-cut. However, if the enquiry focuses on an aspect of phraseology, students may find it difficult to extrapolate the tendencies associated with patterns in language (Hunston & Francis 2000), as they may be confronted with conflicting examples which do not follow a particular pattern in all cases.

One area that posed difficulty for my students was that of ergativity. As noted by Celce-Murcia (2002), overpassivisation of ergative verbs is an aspect that poses particular problems for advanced learners:

> With the verbs ‘increase’ and ‘decrease’ [the ergative] tends [my italics] to be used when the inanimate subject is objectively or subjectively measurable (rather than an animate agent/dynamic instrument object — both of which favor active voice — or a patient subject — for the passive voice.)

(Celce-Murcia 2002: 146)

Students found it difficult to work out from a close reading of concordance lines the correct choice of verb in the following sentence because of the probabilistic nature of language when viewed syntagmatically:

> With a very crowded schedule, students’ level of motivation was decreased / has decreased.

Vannestål & Lindquist (2007) have commented on the difficulty students have in interpreting corpus data and this aspect seems to be a particularly thorny issue when phraseology comes into play. It would seem then that it is in order to supply prompts or hints to enable students to work out the tendencies of phraseological patterns. For example, in the case of the use of the ergative students could be given
a prompting question such as: “Do you notice any difference in the subjects for was decreased and has decreased”?

In tackling corpus-based enquiries, Carter & McCarthy (1995) have formulated the ‘3 Is’ strategy:

- Illustration: looking at data
- Interaction: discussion and sharing observations and opinions
- Induction: (making one’s own rule for a particular feature)

However, based on the difficulties my students have encountered with inducing phraseological tendencies, I would like to elaborate on the above model, by proposing a ‘4 Is’ formulation, adding ‘Intervention’ as an optional stage between Interaction and Induction. This would allow the inclusion of hints such as the one mentioned above. Although in the literature on language teaching deductive and inductive approaches are usually seen as polarities, the above discussion has shown that clues and prompts can be used to mediate the inductive ↔ deductive continuum. For this reason, the following dynamic paradigm for corpus investigations is proposed, which allows for finer-tuning of corpus queries.

![Dynamic paradigm for corpus investigations](image)

Implementing a more delicate approach to corpus queries would help to reduce some of the difficulties associated with interpretation for students, especially when they are engaged in working out phraseological tendencies. As pointed out by Gardner (2007) it is this combinatorial nature of lexis and grammar which poses problems:

…it is likely that only the most advanced language learners can take advantage of the intricate semantic relationships between words that are revealed through concordancing. Certainly, such an approach to language training presupposes that learners will know most of the words (cotext) that surround a key word or phrase in context (KWIC), and that they can connect their meanings — an assumption that seems unreasonable for many groups of language learners (children, beginning L2 learners, learners with low literacy skills etc.)

(Gardner 2007: 255)
Corpora are useful for phraseological enquiries (cf. Granger & Meunier 2008, Meunier & Granger 2008) as the language which falls between lexis and grammar is often not easily retrievable from grammars or dictionaries. However, some intervention in the form of clues or hints may be needed to enable students to connect meanings. Conversely, while hard-and-fast grammar rules may be easier for students to glean from corpora, a corpus, or indeed a particular sub-corpus, may not be the best, or most efficient, resource for consultation. This issue is the focus of the following section.

5. Which corpus and which online resource?

Chambers (2005) and Chambers & O’Sullivan (2004) have underscored the importance for students of having the ability to select appropriate electronic resources:

The concept of literacy now includes not only the knowledge and skills which are traditionally associated with that concept, but also the ability to select, evaluate and use the electronic tools and resources appropriate for the activity which is being undertaken.

(Chambers & O’Sullivan 2004: 158)

In this respect, Davies (2004) reports on a program on student use of three main corpora for examining syntactic variation in Spanish, noting that sometimes the students’ intention was to use a corpus that was not the most appropriate for the research question they had formulated.

In my own class of report writing referred to earlier in the article, students wanted to know which of the verb collocations below was the most appropriate for survey.

We plan to do / carry out / conduct a survey on the use of computers.

Students considered the 7-million word sub-corpus of reports to be ideal for searching the noun survey and expected that it would show correct verb + noun collocations. Although the corpus data displayed useful verbs to collocate with the noun survey, these were not easy to discern. There was a lot of ‘noise’ as students were required to read through quite a number of concordance lines to identify appropriate verb + noun collocations for their context of writing, as evidenced by the results shown in Figure 5.

This problematic example above then gave me the opportunity to remind students of another program, JustTheWord. The screenshot below shows this to be a more appropriate online tool to use, with the cluster feature of particular use as the collocations are grouped semantically. In Figure 6 below, a glance at Cluster 1
Applying corpus linguistics to pedagogy

<table>
<thead>
<tr>
<th>Words</th>
<th>Left sort</th>
<th>Right Sort</th>
<th>Show PoS?</th>
<th>Frequency Sorted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response rate to a survey from See contexts</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>And hcfa distributed a survey to See contexts</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Response rate to a survey of See context</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Response rates to a survey form See contexts</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Thinking about conducting a survey to See contexts</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>$150,000 to undertake a survey and See contexts</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1998 report on a survey by See contexts</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>2 we sent a survey to See contexts</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Acquisition venter/footnote33/sent a survey on See contexts</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Addition to mailing a survey of See contexts</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>And employment funded a survey of See contexts</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>And francis used a survey to See contexts</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

**Figure 5.** Search for *a survey*

**Figure 6.** Search for *survey* in *JustTheWord* collocations program
confirmed students’ initial intuitions, but some were surprised to find that the verb *do* in *Cluster 3* was acceptable. An examination of the concordance lines for this collocation revealed, though, that it was mainly used in an informal setting in speaking, as in the following: *I mean, I haven’t done a detailed survey on anything.*

One misconception held by students was that the Business Letters Corpus would be useful for consulting for any aspect of their letter writing. The utility of this corpus for answering business-related language queries such as the structure and use of phrases with *appreciate* has been illustrated earlier in this article. For other problematic areas, though, such as topic-comment (e.g. *For the training program, it will start on…*), it would have been more appropriate to consult a local reference grammar targeting common errors of Hong Kong students.

It is noteworthy that which resource (corpus, grammar, dictionary etc.) is the most appropriate for a particular query has not been explored much to date. Kennedy (2008) notes that a corpus might not be the most efficient way for students to discover the differences in use between *tall, high, upright* and *vertical* when the differences are made explicit in good dictionaries, but such insightful observations are few and far between in the literature. This is an important area that Bernardini (2002, 2004) has flagged for future development.

Here are two sets of typical examples, one from published journal articles and one from student dissertations. What do you notice about the use of *it seems* in the two sets of examples? Can you suggest why they are different?

<table>
<thead>
<tr>
<th>Published articles</th>
<th>Student dissertations</th>
</tr>
</thead>
<tbody>
<tr>
<td>• It seems clear that as insider holding proportions increase, capitalization ratios decrease.</td>
<td>• It seems that different studies have shown different results.</td>
</tr>
<tr>
<td>• It seems likely that the eighties and nineties will be known as decades of large scale disaggregation.</td>
<td>• It seems that the practice of employing local staff by multinationals is increasing.</td>
</tr>
<tr>
<td>• It seems quite probable that consumers would not recognize such relatively small degrees of difference.</td>
<td>• It seems that some individual training courses are below their full capacity.</td>
</tr>
</tbody>
</table>

Now look at the following examples of *it seems that* from published journal articles. How is it used differently from student dissertations?

- It seems that consumers are more likely to use price tactic and switch stores only when certain brands and product categories are promoted.
- It seems that the issue of privatization could become an object of a national referendum.

**Figure 7.** Concordance task for *it seems* in published articles and student dissertations (from Hewings 2002)
Neither should it be forgotten that corpora of learner writing are another valuable resource in corpus-based pedagogy (see Pravec 2002 for a review), either to inform materials (cf. Granger 2004; Gilquin et al. 2007; Mukherjee 2006) or for exploitation by the learners themselves (Hewings & Hewings 2002; Mukherjee & Rohrbach 2006; Seidlhofer 2000). For example, Mukherjee & Rohrbach (ibid.) propose individualising the corpus analysis in order to compare variation in individual learners’ output. Having learners build corpora of their own writing to compare with a reference corpus would thus increase the relevance of corpus-based pedagogy by individualising it. The corpus-based materials of Hewings & Hewings (2002) and Hewings (2002) on the use of metadiscoursal anticipatory *it* in professional business writing, i.e. published journal articles from the field of Business Studies, also incorporate the findings from learner corpora (MBA dissertations written by non-native speakers). Asking students to compare and discuss the differences of *it seems*... in concordance lines selected from the two corpora, as shown in Figure 7 overleaf, would serve to alert students to particularly problematic areas for post-graduate writers, which students might not appreciate if they were just exposed to working with expert or professional corpora.

6. Conclusion

This article has reviewed four inter-related issues concerning the application of corpus linguistics to pedagogy and ESP in particular. It can be seen that very recent pedagogic endeavours have adopted a much more discourse-based, top-down approach to analysis (or worked from a bottom-up to a more top-down analysis), a development that was advocated by Flowerdew (1998) over a decade ago. It has also been illustrated that corpus pedagogy has progressed beyond looking at truncated concordance lines, and is now encompassing Sinclair’s ‘units of meaning’, outlined in the introduction of this article.

However, the issue of contextualization still remains problematic and it is envisaged that in future more attention will be paid to the mark-up of written text with contextual features, as is the norm for spoken corpora nowadays. It has been shown, though, that corpora are not completely devoid of context, and that the co-textual environment may provide useful contextual clues. Although there are a few accounts in the literature regarding the ‘pedagogic mediation’ of corpus data, these are few and far between, indicating this is an area for further discussion and expansion. Finally, it has been suggested that more attention needs to be paid to the types of enquiry corpora are best suited for. The increasing availability of other online resources, such as grammars, thesauri, dictionaries etc., will make it easier for students to toggle between a multitude of online resources to decide which is
the most relevant and useful look-up tool. Learner corpora, it is argued, are also of value here. However, the above can only be accomplished with strategy training, not only of students but also of teachers, as called for by Frankenberg-Garcia (2006). There is therefore still much to debate and develop in the application of corpus linguistics to pedagogy, a field first founded with the pioneering work of Tim Johns (1991a, 1991b) in the early nineties.

Notes

* This is a revised and extended version of a paper given at the 8th Teaching and Language Corpora Conference, Lisbon, Portugal on 6th July 2008, and also an invited lecture given at the Hong Kong Association for Applied Linguistics on 5th March 2007.

I wish to thank the two anonymous reviewers for their very helpful and constructive comments on an earlier draft of this paper. Any shortcomings, naturally, remain my own.

1. I use ‘corpus-based’ in this article to refer to any hands-on pedagogic applications of corpora. See Tognini-Bonelli (2001) for a discussion on her definitions of ‘corpus-based’ vs. ‘corpus-driven’. See also Lee (2008) for additional details on ‘corpus-informed’ and ‘corpus-supported’ linguistics.


4. Information on MICASE can be found at http://quod.lib.umich.edu/m/micase/ (accessed July 2008).

5. JustTheWord is an online collocations program which interfaces with the 100-million-word BNC.

References


Kennedy, G. 2008. “Phraseology and language pedagogy”. In F. Meunier & S. Granger (Eds.), *Phraseology in Foreign Language Learning and Teaching*. Amsterdam/Philadelphia: John Benjamins, 21–41.


**Author’s address**

Lynne Flowerdew
Hong Kong University of Science and Technology
Language Centre
Clear Water Bay Road
Kowloon
Hong Kong, SAR

lclynne@ust.hk