Learner corpora: the missing link in EAP pedagogy

Gaëtanelle Gilquin, Sylviane Granger & Magali Paquot

[DRAFT VERSION]

Abstract

This article deals with the place of learner corpora, i.e. corpora containing authentic language data produced by learners of a foreign/second language, in EAP pedagogy and sets out to demonstrate that they have a valuable contribution to make to the field. Following an initial brief introduction to corpus-based analyses of academic writing, the article zooms in on learner corpora, describing some of the findings that emerge from corpus studies of L2 learners’ EAP writing. The next section examines the use of corpora in EAP materials design and shows that the few existing corpus-informed EAP tools tend to be based on native corpora only. The article then reports on a collaborative corpus-based project between the Centre for English Corpus Linguistics (Université catholique de Louvain) and Macmillan Education, which aims to describe a number of rhetorical functions particularly prominent in academic writing. The analysis of learner corpus data and their comparison with data from native corpora have highlighted a number of problems which non-native learners experience when writing academic essays, e.g. lack of register awareness, phraseological infelicities, semantic misuse. In this article, we illustrate how these findings were used to inform a 30-page academic writing section in the second edition of the Macmillan English Dictionary for Advanced Learners.

Keywords
corpus linguistics – learner corpora – academic writing – materials design – phraseology – dictionary

1. Introduction

Computer corpora have secured a key role in most language-related fields, from lexicography to language teaching through natural language processing and literary criticism. The corpus wave has spread to the EAP field but a look at the literature and pedagogical materials shows that EAP researchers and materials writers mainly use native corpora. Learner corpora, that is corpora containing data produced by L2 learners - both foreign and second language learners - are seldom analyzed, which is regrettable as they hold tremendous potential for EAP studies. L2 learners admittedly share a number of difficulties with novice native writers but they have also been proven to have their own distinctive problems which a careful corpus-based investigation can help uncover. The aim of this article is to show that analyzing learner corpus
data is an effective way of “operationalizing writing difficulties” (Bitchener & Basturkmen, 2006, p. 14).

The article is structured as follows. In sections 2 and 3 we highlight the respective contribution of corpora in general, and learner corpora in particular. Sections 4 and 5 focus on EAP materials design, with, in section 4, a brief survey of corpus-based EAP materials, followed by the presentation, in section 5, of one concrete achievement in the field of pedagogical lexicography, the integration of learner-corpus-informed materials into the new edition of the Macmillan English Dictionary for Advanced Learners. Section 6 concludes the article.

2. The contribution of corpora to EAP

J. Flowerdew (2002) identifies four distinct major research paradigms for investigating English for Academic Purposes (EAP), namely (Swalesian) genre analysis, contrastive rhetoric, ethnographic approaches and corpus-based analysis. While the first three approaches to EAP place emphasis on the situational or cultural context of academic discourse, corpus-linguistic methods focus more on the co-text of selected lexical items in academic texts. This co-textual approach has enabled corpus-linguistic studies to make two significant contributions to the field of EAP: detailed descriptions of its distinctive linguistic features, and more specifically its highly specific phraseology, and careful analyses of linguistic variability across academic genres and disciplines.

Corpus linguistics is concerned with the collection and analysis of large amounts of naturally occurring spoken or written data in electronic format, “selected according to external criteria to represent, as far as possible, a language or language variety as a source of linguistic research” (Sinclair 2005). Computer corpora are analysed with the help of software packages
such as WordSmith Tools 4 (Scott 2004), which includes a number of text-handling tools to support quantitative and qualitative textual data analysis. Wordlists give information on the frequency and distribution of the vocabulary – single words but also word sequences – used in one or more corpora. Wordlists for two corpora can be compared automatically so as to highlight the vocabulary that is particularly salient in a given corpus, i.e. its keywords or key word sequences. Concordances are used to analyse the co-text of a linguistic feature, i.e. its linguistic environment in terms of preferred co-occurrences and grammatical structures. More sophisticated tools are currently being developed to help researchers explore large corpora. For example, the Sketch Engine provides “word sketches”, i.e. “one-page automatic, corpus-based summaries of a word’s grammatical and collocational behaviour” (Kilgarriff et al. 2004). Frequency is a key issue as corpus-based studies aim to provide automated descriptions of what is frequent and typical in the corpus under examination. The research paradigm of corpus linguistics is thus ideally suited for studying the linguistic features of academic discourse as it can highlight which words, phrases or structures are most typical of the genre and how they are generally used.

Corpus-based studies have shed light on a number of distinctive linguistic features of academic discourse as compared with other genres. Biber et al. (1999) have shown, for example, that nouns, nominalisations, derivational suffixes and linking adverbials are particularly frequent in academic prose while private verbs, that-deletions and contractions occur very rarely. Studies of vocabulary in academic prose have stressed the importance of a sub-technical vocabulary that is common to a wide range of academic texts and disciplines and that is typically used to serve organisational or rhetorical functions prominent in academic writing, e.g. introducing a topic, hypothesizing, exemplifying, explaining, evaluating, concluding (cf. Thurstun and Candlin, 1998; Luzón Marco, 2001). Several studies have pointed to the existence of an EAP-specific phraseology characterized by word combinations...
that are essentially semantically and syntactically compositional, e.g. *in the presence of*, *the aim of this study*, *the extent to which*, *it has been suggested*, *it is likely that* (Biber et al., 1999). These studies have also shown that the phraseology of academic discourse is highly conventionalized and that novice writers differ from professional writers in their use of EAP-specific lexical bundles (Cortes, 2002).

With the recent development of specialised genre-based corpora (cf. L. Flowerdew, 2002, p. 96), the field of academic discourse research has witnessed a rapid increase in the number of studies on variability within academic texts. Studies have investigated the similarities and differences between different genres within the same academic discipline (e.g. Conrad, 1996). Others have described differences in the same genre across several disciplines (e.g. Hyland, 2000; Fløttum et al., 2006) and even sub-disciplines (e.g. Ozturk 2007). Some studies have also compared the use of linguistic features across text sections (e.g. Biber and Finegan, 1994; Martinez, 2003).

A number of these variationist studies have also focused on the phraseological preferences of academic prose and have shown that phraseological patterns may differ across genres and disciplines. They have also suggested that phraseological patterns correlate closely with the communicative purposes that they serve in different genres or disciplines (Groom, 2005; Charles, 2006) and with the rhetorical functions that they perform in specific text sections (Gledhill, 2000). Thus, these studies support Hyland’s (2002) plea for as much specificity as possible in the teaching of EAP at university but, as will be discussed in section 4, their pedagogical implementation is not without its problems.

### 3. The contribution of learner corpora to EAP

The distinctive, highly routinized, nature of EAP proves undeniably problematic for many (especially novice) native writers, but it poses an even greater challenge to non-native writers.
Until recently, it was quite difficult to form a precise picture of learner EAP writing, but learner corpus research has the potential to offer a major breakthrough as researchers now have access to large databases of learner data and powerful methods of analysis.

Learner corpora are a relatively new addition to the wide range of existing corpus types (cf. McEnery et al., 2006 for a survey of corpus resources). Their specificity resides in the fact that they contain data from foreign or second language learners. More than any other, this type of corpus needs to be compiled on the basis of strict design criteria in order to control the wide range of variables that affect learner language, both learner variables (age, proficiency level, mother tongue background, etc.) and task variables (field, genre, topic, etc.) (Granger, 2002). Learner corpus data offer a number of significant advantages over other types of learner data: the corpora are usually quite large and therefore give researchers a much wider empirical basis than has ever been available before; they can be submitted to a wide range of automated methods and tools which make it possible to quantify learner data, to enrich them with a wide range of linguistic annotations (e.g. morpho-syntactic tagging, discourse tagging, error tagging) and to manipulate them in various ways in order to uncover their distinctive lexico-grammatical and stylistic signatures. One method of analysis, Contrastive Interlanguage Analysis (CIA) (Granger, 1996; Gilquin, 2000/2001), has played a key role in identifying L2-specific features. This methodology, which is very popular among learner corpus researchers, involves two types of comparison: comparisons of learner language and one or more native speaker reference corpora (L2 vs. L1), and comparisons of different varieties of learner language (L2 vs. L2). It has been applied to a wide range of linguistic features – orthographic, lexical, grammatical, phraseological, stylistic, pragmatic – and has brought to light interesting patterns of overuse, underuse\(^1\) and misuse which are

---

\(^1\) It is important to note that the terms “overuse” and “underuse” are descriptive, not prescriptive, terms: they merely refer to the fact that a linguistic form is found significantly more or less in the learner corpus than in the reference corpus.
helping to fill in some gaps in our hitherto somewhat patchy knowledge of the different stages of interlanguage development.

Interest in learner corpora is growing fast and has already generated a range of stimulating studies which highlight the potential of this new resource for the EAP field. Milton & Tsang (1991) were among the first to deplore the lack of sufficient evidence to quantify students’ problems in written expression and advocate the use of learner corpus data to compensate for it: “Without a reliable index of the degree of difficulty that our students have with the various dimensions of written English such as its lexis, syntax, pragmatics and semantics, we are left to make do with approximations based on impressions, anecdotes and manual counts of small samples” (p. 216). Similarly, L. Flowerdew (2001, p. 364) insists that “insights gleaned from learner corpora need to be employed to complement those from expert corpora for syllabus and materials design”. She shows how careful investigation of learner corpus data can help uncover three areas of difficulty in learner EAP writing: collocational patterning, pragmatic appropriacy and discourse features. Errors such as “we have performed a survey” or “a questionnaire has been conveyed to the public” reveal that students have some knowledge of key EAP verbs but are not familiar with their lexico-grammatical patterning. Learner anomalous use of modal verbs, modal adjuncts, boosters, hedges, etc. causes pragmatic inappropriacy. Hyland & Milton’s (1997) investigation of expressions of doubt and certainty shows that Cantonese learners use a more restricted range of epistemic modifiers and have considerable difficulty conveying the appropriate degrees of qualification and confidence. Finally, as regards discourse features, quite a number of studies have focused on the use of connectors in EAP writing and revealed patterns of overuse, underuse and misuse (Milton & Tsang, 1991; Granger & Tyson, 1996; Altenberg & Tapper, 1998; L. Flowerdew, 1998). Other studies highlight the more general problem of learners’ tendency to adopt an overly spoken style in their EAP writing. Using a fully automatic method of investigation,
Granger & Rayson (1998) show that learners overuse many lexical and grammatical features typical of speech, such as frequent use of first and second person pronouns or the use of short Germanic adverbs (also, only, so, very, etc.), and significantly underuse many of the characteristics of formal writing, such as a high density of nouns and prepositions. In addition, many of the typical EAP words (issue, advocate, belief, argument) are typically underrepresented while general and/or vague nouns (people, thing, problem) are overrepresented.

One important finding emerging from learner-corpus-based studies in general and EAP in particular is that some of the linguistic features that characterize learner language are shared by learners from a wide range of mother tongue backgrounds whilst others are exclusive to one particular learner population. The shared features can be assumed to be developmental whilst the latter are presumably due to transfer from the learners’ mother tongue. Although further work is needed to consolidate the results, most studies point to a marked influence of the learners’ L1, in particular as regards (semi-) prefabricated language (cf. Nesselhauf, 2005; Paquot, in press).

All these studies show that insights gained from learner corpus research have huge potential for EAP research. However, the overwhelming majority of corpus-based EAP studies are exclusively based on native corpora. In the call for proposals for this special issue of JEAP, Hyland (2006, p. 248) quotes a proportion of 40 % of corpus-based articles in 2005 and 2006 issues of JEAP. A close look at the content of the articles shows that in the overwhelming majority of the cases, corpus-based in fact means native-corpus-based. Analyses of L2 learners are not absent but they tend to focus on the writing process rather than the product and/or use traditional non-corpus-based methods to analyze learner productions.
Admittedly, some corpus-based studies focus on novice native writing and it can be assumed that novice native writers share a number of difficulties with non-native writers. For example, Neff et al. (2004) show that excessive visibility of the writer is common to native and non-native student writing. As we will show below, however, the overlap between novice native and non-native writing is far from perfect, and quite a few difficulties appear to be specific to learners. This issue of the degree of overlap between novice native writers and non-native writers has far-reaching methodological and pedagogical implications and is clearly in need of empirical studies.

The use of non-native speakers’ writing corpora has been advocated by several linguists and a number of descriptive studies exist on important EAP topics like metadiscourse (Hewings & Hewings, 2002. Martinez, 2005. Ådel, 2006). However, as pointed out by L. Flowerdew (2001, p. 366), these studies have had little pedagogical impact: “not many of the findings have been applied directly to pedagogy and tend to remain at the level of implications”. The following two sections will be devoted to this important issue.

4. Corpora and EAP materials design

While materials designed to help students improve their academic writing skills are legion (e.g. Bailey, 2006; Hamp-Lyons & Heasley, 2006), few are corpus-informed, relying instead on materials writers’ perceptions of what good academic prose is or should be. Because of this lack of empirical support, many of these tools provide misleading information and unsound advice, as comparisons of published EAP materials and actual usage reveal (cf. Paltridge, 2002).

A careful examination of some of the major EAP resources reveals that the few that are corpus-informed tend to be based on native corpora only. Thus, both Thurston & Candlin’s (1997) Exploring Academic English, which proposes a fully concordance-based
approach to EAP, and Coxhead’s (2000) Academic Word List, which has given rise to several
textbooks (e.g. Huntley, 2006; Schmitt & Schmitt, 2005), rely exclusively on data from native
corpora. Although such tools are very useful, they are arguably less well-suited for non-native
learners, despite Thurstun & Candlin’s (1998) claim that their own materials are equally
appropriate for native and non-native writers. This is because, as mentioned above, learner
writing is characterised by errors and infelicities which are often quite different from those
found in native writing, even novice native writing. By relying solely on native corpus data,
EAP materials ignore these and thus fail to provide non-native learners with the type of
information that is arguably most vital to them. For example, Coxhead’s (2000) Academic
Word List does not include the 2,000 most common English words, with which non-native
writers may still have considerable difficulties, especially in cases where their use in
academic writing differs from their habitual use (see also Paquot, 2007).

What L2 learners need is EAP resource books addressing the specific problems they
encounter as non-native writers. By showing in context the types of errors learners make, as
well as the items they tend to underuse or overuse, learner corpora make such an approach
possible. Yet, hardly any materials writers up to now have taken up the challenge of using
learner corpus data. Milton’s (1998) WordPilot is one of the few concrete pedagogical
applications resulting from the analysis of learner writing and its comparison with a reference
corpus of native writing (see Tseng & Liou, 2006 for another example). Starting from attested
difficulties for Hong-Kong learners, WordPilot proposes a series of remedial tasks and tools,
such as proofreading exercises, an interactive grammar or lists of words or expressions
generally underused by learners. It should be emphasised, however, that this program is a
CALL (Computer-Assisted Language Learning) application (this is the case of Tseng & Liou,
2006 too) and that, once we move on to more traditional resources, positions are more
conservative and innovations, less common, with many editors “far more comfortable with
rehashes of what has gone before than with something different (and refreshing)” (Harwood, 2005, p. 152).

There are several explanations for the relatively modest role that corpora have played in EAP materials design. One of them is the difficulty of dealing with the fragmented picture of academic writing that emerges from native corpus analyses. As already suggested in section 2, corpus-based research has demonstrated that “[t]he discourses of the academy do not form an undifferentiated, unitary mass but a variety of subject-specific literacies” (Hyland, 2002, p. 389). Yet, as Hill (2005) points out, “in most academic institutions it is simply not possible to provide all the students who need EAP support with the same level of specificity”. For financial or logistic reasons, courses often have to focus on English for General Academic Purposes, rather than English for Specific Academic Purposes (see L. Flowerdew, 2002 on this distinction). One way to get round this difficulty is to provide students with specially designed corpora of the literature in their own field (or get them to collect such corpora) and encourage them to scan concordance output to discover regularities of patterning, using the data-driven learning technique advocated by Johns (1994). But while the benefit of such an approach has been underlined by some linguists (e.g. Bowker [2003] in the field of ESP translation training), others are more cautious and speak of “a danger of over-generalization on the part of learners” (Sripicharn 2004, p. 243). In addition, learners need to be sufficiently trained to use these corpora efficiently, which requires “an apprenticeship oriented toward the development of ‘corpus research’ skills” (Kennedy & Miceli, 2001, p. 88), for which teachers may not always have the time.

Learner corpora, too, reveal variability which may be difficult to incorporate in EAP materials. As we saw in section 3, despite some shared problems, many of the difficulties that non-native writers experience are L1-specific. This is true both for lexico-grammatical errors, which may be due to transfer, but also for rhetorical infelicities, which may reflect different
conventions in the learner’s mother tongue (see e.g. Cargill & O’Connor, 2006 on the structural differences between Chinese and Western research articles). While it is interesting from a descriptive point of view, however, L1-orientation is often unrealistic from the perspective of publishers who, for obvious commercial reasons, often prefer generic tools, which can be sold to students all over the world, to L1-specific tools, whose market is usually more limited. As a consequence, EAP materials tend to focus on problems which are shared by many learner populations, leaving it to the teacher to tackle L1-specific issues.

Finally, the creation of “corpus-based pedagogical enterprises” (Swales, 2002, p. 151) is made even more complex by the fact that there is no direct link between corpus findings and pedagogical relevance. As emphasised by Granger (forthcoming), whether characteristics of learner language uncovered by corpus research are “selected for pedagogical action or ignored depends on a variety of features, including learner needs, teaching objectives and teachability”. Thus, some of the attested errors or infelicities may be more crucial for advanced learners than for beginners. Another issue is that of the type of language being targeted. Some learners may have native (or native-like) writing as a target, while others may consider English as a Lingua Franca the ideal target. These factors, combined with the discipline and L1 specificities mentioned above, result in a great diversity which may be quite difficult to reconcile in a holistic pedagogical approach.

While we have mainly focused on EAP resource books up to now, it should be noted that learners have another tool at their disposal in order to improve their writing skills, namely the monolingual learners’ dictionary (MLD). In the last decade, MLDs have taken “more proactive steps to help learners negotiate known areas of difficulty” (Rundell, 1999, p. 47), to the point that they are today conceived of as comprehensive writing tools. They now include productively oriented information in areas such as syntactic behaviour, prevention of errors, phraseology and collocation. Some of the recent editions include writing sections dealing with
various aspects of academic prose, thus adopting a more EAP-oriented approach (see, for example, the *Longman Exams Dictionary* [Major, 2006] or the *Collins COBUILD Advanced Dictionary of American English* [Sinclair, 2007]). Unlike in the EAP textbook industry, the use of corpora is well established in lexicographical practice and MLDs have made good use of corpora of academic writing (although the focus is on English for General Academic Purposes, for the same commercial reasons as described above for resource books), as appears for example from the integration of the Academic Word List into the *Longman Exams Dictionary*. However, the corpora used are still largely native corpora, and learner corpus data have only been integrated into error notes. Yet, we strongly believe that learner corpora can be exploited to improve other aspects of the dictionary. In the next section, we report on such an enterprise, which resulted from the close collaboration between the Centre for English Corpus Linguistics (Université catholique de Louvain) and Macmillan Education.

5. An EAP collaborative project

In the second edition of the *Macmillan Dictionary for Advanced Learners* (MED2, Rundell, 2007), learner corpus-based information is used not only to compile error notes but also to guide the selection and content of grammar sections that focus on aspects of English grammar, spelling and punctuation that are still problematic at an advanced level. In addition, learner corpus data have informed an extended writing section focusing on twelve rhetorical or organisational functions particularly prominent in academic writing: (1) adding information; (2) comparing and contrasting: describing similarities and differences; (3) exemplification: introducing examples; (4) expressing cause and effect; (5) expressing personal opinions; (6) expressing possibility and certainty; (7) introducing a concession; (8) introducing topics and related ideas; (9) listing items; (10) reformulation: paraphrasing or clarifying; (11) reporting and quoting; (12) summarizing and drawing conclusions (Gilquin et
al., 2007, p. IW1-IW29). These writing sections will be the focus of the last part of this article.

5.1. Data and methodology

While we are aware of the importance of discipline- and L1-specific features in EAP, we also recognise the demands of the pedagogical market and the realities of the classroom, and hence situate our research in a more general context, investigating features which are common across disciplines and learner populations. We selected EAP-specific words according to the corpus-driven method described in Paquot (2007), a method that targets discipline-independent words by extracting items which appear in a wide range of academic texts. The list was then completed by words and phrases which did not emerge from our corpus analysis but are commonly mentioned in EAP materials. This selection yielded a total of about 350 EAP markers.

A thorough analysis of these markers in a native expert corpus was first necessary as detailed descriptions of their use are not widely available. In the context of the development of the British edition of MED2, we took British native language as our target. We used the 15-million word academic sub-corpus of the British National Corpus as a comparable corpus.2 Not only does this corpus include samples of academic texts from a wide range of disciplines (including humanities, arts, social science, medicine and natural science), but it is also one of the few available academic corpora that is large enough to give access to lexical patterning and other phraseological phenomena. It represents professional, or expert writing. Using expert writing as a norm against which to compare learner writing is controversial. Researchers such as Lorenz (1999) criticize this type of norm, arguing that it is “both unfair and descriptively inadequate” (ibid. 14). What these researchers recommend, instead, is the use of a corpus of native-speaker student texts, that is, novice writing. However, a novice

---

2 For more information on the British National Corpus, see http://www.natcorp.ox.ac.uk.
norm may not always be desirable, especially in the context of pedagogical applications. As Leech (1998, xix) puts it, “[n]ative-speaking students do not necessarily provide models that everyone would want to imitate”. The question of the norm can be settled by taking into account the aim of the comparison, as Ädel (2006, p. 206-207) does:

On the one hand, it can be argued that in order to evaluate foreign learner writing by students justly, we need to use native-speaker writing that is also produced by students for comparison. On the other hand, it can also be argued that professional writing represents the norm that advanced foreign learner writers try to reach and their teachers try to promote. In this respect, a useful corpus for comparison is one which offers a collection of what Bazerman (1994: 131) calls ‘expert performances’.

Since the aim of MED2 is to help learners improve their writing skills, a professional control corpus such as the BNC is arguably preferable to a non-professional corpus.

The learner corpus that was used within the frame of the MED2 project is the International Corpus of Learner English (Granger et al., forthcoming), which now totals 3.5 million words and includes 6,085 essays written by learners from 16 mother tongue backgrounds belonging to different language families (e.g. French, Chinese, Norwegian, Turkish). Each of the 6,085 essays is accompanied by over 20 learner and task variables (see section 3), all of which have been stored in a database and can be used by researchers as queries to compile sub-corpora that match certain criteria. For the purposes of this project, 16 sub-corpora were compiled on the basis of one learner variable, i.e. the mother tongue variable, and three task variables, i.e. all texts are untimed argumentative essays potentially written with the help of reference tools.

The method used is based on Contrastive Interlanguage Analysis as described in section 3. Native and learner data were compared to highlight distinctive features in learners’
use of lexico-grammatical and discourse patterns. A major advantage of corpus-based research is that it allows researchers not only to check how lexical items are used in naturally occurring data, but also to highlight linguistic features, and more particularly, phraseological patterns that are used by EFL learners but which are not (or very rarely) found in native professional writing (cf. De Cock, 2003). Learner corpora were then compared with each other so as to distinguish between linguistic features found in the writing of learners from a wide range of mother tongue backgrounds and those found exclusively in the writing of one or two learner populations. Only linguistic features shared by at least half of the learner populations under study are discussed in the writing sections. In addition, we used spoken data, extracted from the spoken component of the BNC, to situate learner writing against both academic prose and speech and to assess EFL learners’ awareness of register differences.

5.2. EAP writing functions

Analysing these learner corpus data and comparing them with data from native corpora has highlighted a number of problems which non-native learners experience when writing academic essays. These include problems of frequency, register, positioning, semantics and phraseology. In what follows, we illustrate each of these problems in turn. Then, we show how these problems have been dealt with in MED2.

Learner writing often exhibits patterns of frequency different from those found in native writing. Thus, while in native English, *for instance* is much less common than *for example* (2,410 vs. 9,233 hits in the academic component of the BNC), in learner English it is used almost as frequently as *for example*, which results in a massive overuse of the expression. *For example* too tends to be overused by learners. This is likely to be related to their underuse of a whole range of alternative expressions which can be used for exemplification, such as *X is an example of Y* or *an example of Y is X*, illustrated by (1) and (2).
1. Self-regulation is an example of what we earlier called “corporatism”. (BNC EBM 526)

2. An example of a treaty which apparently gave procedural rights to individuals is the Anglo-Irish Agreement on Northern Ireland. (BNC EF3 1531)

Because learners, as a rule, have a limited repertoire of expressions at their disposal to fulfill a particular rhetorical function, they tend to rely on a few items only, which they use over and over again, to the detriment of other, perhaps less salient expressions (see Paquot, in press on the function of exemplification in learner writing).

Problems of register confusion also regularly arise among non-native writers. Particularly striking is learners’ tendency to use expressions which are more typical of speech than of writing. This is visible, for example, in their overuse of adverbs expressing a high degree of certainty, such as really, of course or absolutely, which are characteristic of speech rather than writing (cf. Figure 1). By contrast, hedging adverbs (e.g. apparently, possibly, presumably), which are common in academic writing, occur much less frequently in learners’ essays.

Several studies have underlined learners’ preference for the sentence-initial position of connectors (cf. Granger & Tyson, 1996 for French-speaking learners and Field & Yip, 1992 for Cantonese learners). To give but one example, learners tend to use however at the beginning of the sentence, as in (3), whereas in native writing, it is more often found inside the sentence, cf. (4). Our analyses show that, in the case of however at least, this tendency is
common to learners from a large number of mother tongue backgrounds, which points to a
universal feature of interlanguage.

(3) **However**, it is extremely difficult to get them away from the fast food corner.
    (ICLE-GE)

(4) There is, **however**, another way in which a useful advance might be made.
    (BNC A1A 1491)

A word such as *though*, by contrast, is often used by learners as an adverb in sentence-final
position, e.g. (5). While this use does occur in native academic writing, it is more typical of
speech. In writing, native speakers favour the conjunctive use of *though*, mostly found in the
middle of the sentence, as exemplified in (6).

(5) There is no doubt about astrology being important in today’s society, **though**.
    (ICLE-SW)

(6) We do not know her name, **though** it is unlikely that the sculptor did not have a
    model. (BNC A04 1465)

The numerous phenomena of over- and underuse should not hide the fact that items
are also often misused by learners. The incorrect use of *on the contrary*, for example, brought
to light by linguists like Crewe (1990) or Lake (2004), is confirmed by our corpus data and
appears to be valid across several L1s. Sentence (7) is representative of the way *on the
contrary* is usually used in ICLE. While it normally expresses a direct denial of what has been
asserted before, meaning that the opposite is true, learners tend to use it simply to describe
differences, as a synonym of *by contrast or on the other hand*.
(7) Some of us overuse this trait, some, on the contrary, use it very seldom.

(ICLE-PO)

Finally, phraseology also appears to be a major stumbling block for learners. To take the example of the word conclusion, our study indicates that learners often use it in the (unidiomatic) expression as a conclusion, illustrated in (8). This expression is very rare among native writers, who prefer in conclusion, cf. (9).³

(8) As a conclusion I would like to say that imagination will never die simply because all the adults are big children who love everything fairy and unreal and who never forget about tales and miracles existing in the world. (ICLE-RU)

(9) In conclusion the challenges posed by an ageing population need radical rethinking and immense cultural changes in attitudes. (BNC HXT 545)

Another difference, which also emerges from the comparison of (8) and (9), is that in learner writing in conclusion or as a conclusion are often followed by an expression including the personal pronoun I, cf. I would like to say in (8), which is hardly ever the case in native writing.

The writing sections in MED2, besides showing how several important rhetorical functions are expressed in native academic writing, specifically address the five types of problems discussed above. Our treatment of these problems is mainly explicit (see Rundell, 1999 on the distinction between explicit and implicit treatment of learner corpus-based information), in that we draw learners’ attention to error-prone items and we provide them

³ Interestingly, Mukherjee & Rohrbach (2006) make the same remark about German-speaking learners, but note that, in their data, as a conclusion only appears at a later stage in the language acquisition process, which suggests that phraseological problems may particularly affect more advanced learners.
with negative feedback in the form of “Be careful!” notes and “Get it right” boxes. Numerous authentic examples are provided to illustrate all the points we make. Problems of frequency, register confusion and atypical positioning are dealt with by means of graphs like the one in Figure 1 above. Such graphs help the reader visualise the differences between learners’ behaviour and that of native writers. In addition, in cases where learners tend to use the same few items over and over again, we present them with alternative ways of expressing the function, thus giving them a chance to widen their lexical repertoire. Figure 2 shows a “Be careful!” note which warns the reader against the excessive use of *for example, for instance* and *e.g.* to express the function of exemplification. It is followed by a series of expressions which, though common in native academic writing, are neglected by learners.

[insert figure 2 here]

Typical semantic errors are illustrated in “Get it right” boxes, such as the one in Figure 3 for *on the contrary*. These boxes start with an authentic learner error and explain why a particular item is not appropriate in a given context and how it can be corrected. They also give a clear description of what the item means and how it should be used.

[insert figure 3 here]

Finally, unidiomatic collocations recurring in learner writing are highlighted (e.g. *according to me*, Gilquin et al., 2007, p. IW15) and some of the most common collocates of certain words are offered in special boxes, thus combining an explicit and implicit approach. In Figure 4, some typical adjectives and verbs used with the word *conclusion* are listed and examples are provided.
It should be emphasized that the information found in the writing sections may be helpful for novice native writers, whose writing presents some of the problems displayed by non-native writing. Thus, a study of the LOCNESS corpus, a corpus of native student essays (see Granger, 1996), reveals a number of shared problems, such as the semantic misuse of *i.e.* or the lack of register-awareness, with an overuse of speech-like lexical items like *maybe, so, really, PRO is why, I think and first of all* (see Gilquin & Paquot, 2006). However, the sections were written with the non-native writer in mind, and many of the features described are learner-specific. The overuse of *for instance* or the erroneous use of the expression *as a conclusion*, for example, are limited to non-native students. Other learner-specific features include the overuse of items such as *I would like/want/am going to talk about, certainly, to my mind, from my point of view, as far as I am concerned*; the semantic misuse of *on the other side, as* (in lieu of *such as*); the erroneous use of *according to me*; and lexico-grammatical errors such as *a same, possibility *to, despite *of or discuss *about. Interestingly, in a number of cases, the data also reveal a continuum between expert writing, novice writing and learner writing, as illustrated in Figure 5 for *maybe.*

### 6. Conclusion

The fast-growing literature on corpus-based EAP is a clear indication that corpora are beginning to gain a strong foothold in the field. This article has proposed that the hitherto largely native speaker orientation of these studies can usefully be complemented with an L2 perspective derived from the careful analysis of learner corpora. Although this two-pronged
approach was advocated by scholars as far back as the early 1990s, it has remained relatively marginal, failing to have a major impact on EAP materials design.

Our investigation of rhetorical functions has shown that the use of learner corpus data, and their systematic comparison with native corpora, can bring to light a wide range of learner-specific patterns, not limited to grammatical or lexical errors, but also including clumsy wording, over-reliance on a limited set of linguistic items and under-representation of a wide range of typical EAP writing patterns. While we have shown here how these findings can be integrated into a learner’s dictionary, it is our belief that other writing resources, such as textbooks or electronic writing aids, can equally benefit from the use of learner corpus data.

Finally, our study has also provided some insight into the relation between novice writing and non-native writing. Of the three categories of difficulties identified by L. Flowerdew (see section 3), only the first – lexico-grammatical patterning – is exclusive to L2 learners; the other two – pragmatic appropriacy and discourse patterns – display only partial overlap. This is an important and drastically under-researched area which should figure prominently on the EAP agenda.

Acknowledgements

We gratefully acknowledge the support of the Communauté française de Belgique, which funded this research within the framework of the ‘Action de recherche concertée’ project entitled ‘Foreign Language Learning: Phraseology and Discourse’ (No. 03/08-301). Gaëtanelle Gilquin also acknowledges the support of the Belgian National Fund for Scientific Research (FNRS).

References


**Figure legends**

Figure 1. Relative frequency of *really, of course and absolutely* in native academic writing, non-native academic writing and speech (Gilquin et al., 2007, p. IW17)

Figure 2. “Be careful!” note for exemplification (Gilquin et al., 2007, p. IW10)

Figure 3. “Get it right” box for *on the contrary* (Gilquin et al., 2007, p. IW9)

Figure 4. “Collocation” box for *conclusion* (Gilquin et al., 2007, p. IW29)

Figure 5. Relative frequency of *maybe* in native expert writing, native novice writing, non-native writing and speech