Repeated Language in Academic Discourse: The Case of Biology Background Statements

Diane Pecorari, Mälardalen University

Abstract
Repetition in language use has been approached from several rather diverse angles, including pre-fabricated multi-word lexical units and intertextuality of types ranging from quotation to patchwriting (Howard, 1995) to plagiarism. This paper suggests that such divergent approaches to the question of repetition have commonalities which can inform EAP practice, and reports the results of an investigation into repetition in a specific element in biology research articles.

1. Introduction

Bolinger's oft-quoted dictum that 'speakers do at least as much remembering as they do putting together' (1976: 2) put the question of repetition in language use on the applied linguistics research agenda. Work on repetition in language can be placed in two broad groups. First, a sizeable and rapidly growing body of research has investigated fixed or semi-fixed chunks of language. These chunks have been labelled variously as lexical phrases (Nattinger & DeCarrico 1992), formulaic sequences (Schmitt 2004), prefabs (Erman & Warren 2000), and lexical bundles (Biber, Johansson, Leech, Conrad & Finegan 1999). It is not merely the terminology that distinguishes these approaches; they take different features into consideration as well. For example, Biber et al. (1999) define lexical bundles in terms of frequency and distribution, while others take into account factors such as transparency of meaning, and whether a unit is recognized as conventional by native speakers. The existence and prevalence of these units have been explained as the result of them being stored as single units (Peters 1973; Wray 2002). It is in this sense that multi-word units are remembered rather than put together, and so it is this first category which relates most closely to Bolinger's remark.

The second broad category of repetition differs from the first in that it is more conscious, and the source of the repeated language is not the mental lexicon but a specific earlier text. One of a number of subsets of

Nordic Journal of English Studies, vol. 7 No. 3, December 2008 9-33

this category is quotation, which involves the direct and intentional repetition of an earlier formulation, usually with attribution. Although quotation can occur in spoken language (e.g., Tagliamonte & D'Arcy 2004), it is an especially common feature in written academic discourse. Quotation, as a sub-set of citation, has been a topic of investigation within the sociology of science and bibliometrics, and more recently within English applied linguistics (see Swales 1986 and White 2004 for reviews). Within English for Academic Purposes (EAP), research has investigated frequency, forms and functions of reports of other sources, including, but not limited to, quotation (e.g., Charles 2006; Dubois 1987; Pecorari 2006; Salager-Meyer 1999; Thompson 2000).

The received view in academia is that quotation is ordinarily the only legitimate means of incorporating language from an existing text into a new one; or, stated differently, that when language is repeated from an earlier text, writers should signal that fact. That widely held principle notwithstanding, it is equally well known that not all repeated language use *is* signalled. When *de facto* quotation is not acknowledged, the result is often viewed as plagiarism, a deceptive act of wrongdoing which is treated by the academic community with unreserved scorn.

However, not all unacknowledged repetition constitutes prototypical plagiarism (i.e., the unattributed repetition of language with intention to deceive; Pecorari 2003).[1] Sometimes it can best be classed as *patchwriting* (Howard 1995; 1999). Patchwriting occurs when inexperienced writers (believe that they) lack a sufficiently skilled authorial voice and draw on the language of other, more proficient writers, to produce a text which has the superficial features of the unfamiliar register within which they are working. Patchwritten texts typically stitch together elements from several sources, possibly with alterations to word choice or structure. Writers exhibit varying degrees of awareness of their reproductive strategies, and particularly of the extent to which those strategies are unconventional (Pecorari 2008). Writers may see copying with changes as a necessary survival skill (Currie 1998;

[1] This compact definition may give the misleading impression that plagiarism is a clearcut and uncontested construct. While this is not the case, space does not permit a more nuanced discussion of plagiarism.

Spack 1997), an unavoidable flaw given practical constraints (Pecorari 2008), or even as a token of their virtue, because changes to the source text reflect a good-faith effort to write autonomously (Angelil-Carter 2000; Hull & Rose 1989).

The strategy of repeating language from a source is defended as acceptable by some writers on two potentially overlapping grounds. The first will be familiar to many EAP specialists who have heard students lament the difficulty of expressing ideas accurately (and in a foreign language) while observing the demand for originality in academic writing: 'there are only so many ways to say the same thing.' The second is a view that, despite the strongly worded strictures against it, some sorts of unacknowledged repetition are acceptable in some contexts.

This point was made by a Turkish physicist in a public episode which exposed sharply divergent views of acceptable source use practices. In 2007, *Nature* reported on of 'a massive plagiarism scandal' (Brumfiel: 8) resulting in 70 allegedly plagiarised research articles being withdrawn from a database. The originator of the database, a Cornell University academic, said that the response should not be 'overly draconian' but characterized the plagiarism as 'dishonest and sloppy' (p. 8). But in a letter to the editor the following month, one of the scientists implicated characterized the accusations as 'upsetting and unfair':

> It's inappropriate to single out my colleagues and myself on this issue. For those of us whose mother tongue is not English, using beautiful sentences from other studies on the same subject in our introductions is not unusual. I imagine that if all articles from specialist fields of research were checked, similarities with other texts and papers would easily be found. . . . Borrowing sentences in the part of a paper that simply helps to better introduce the problem should not be seen as plagiarism. Even if our introductions are not entirely original, our results are — and these are the most important part of any scientific paper. (Yilmaz 2007: 658)

Three themes are worth highlighting in this response. First, the strategy of repeating language is implied to be particularly necessary for scholars 'whose mother tongue is not English.' Second, this practice is said to be 'not unusual.' Third, it is suggested that the repetition is unimportant because it was limited to the parts of the articles which provided background information ('our introductions') and not the findings, which are 'the most important part' of an article. The same themes emerge clearly from Flowerdew and Li's (2007) study of the

writing processes of second-language academic writers in Hong Kong. They concluded that

> students' language re-use goes well beyond formulaic expressions and technical terminology which are characteristics of the scientific research article, yet the students believe that their textual practices do not constitute plagiarism, which, to them, primarily means the stealing of others' work. (p. 440)

Participants in my investigation of source use in postgraduate student writing would likely agree with that. Erden, a biology student, commented on the line he drew between appropriate and inappropriate borrowing:

> Copying a whole paragraph without giving citation, it is a plagiarism. But taking a note in one paper, just one sentence comes directly or, yeah, one or two sentence or one explanation, it is I think acceptable. (Pecorari 2008: 115)

Ingrid, another biologist, had also drawn conclusions about what is appropriate:

> You read through all these articles and you find out that they actually write, they have quite similar introduction, the introduction are quite similar in all the different articles. . . . And I just can't see whose words are theirs and I just . . . put them because that's kind of like common knowledge that most people know, it's easy to find it, you can find it anywhere. (Pecorari 2008: 116)

Ingrid apparently believed, based on similarities in article introductions she read, that the authors repeat language from other sources ('I just can't see whose words are theirs'). Because the kind of information reported in an introduction can be found 'anywhere,' the need to show where it had come from was less pressing. As a result, Ingrid did indeed 'just. . . put them': a literature review section in her thesis was made up almost entirely of language repeated from her sources, often without any citation at all.

As this brief review suggests, these two types of repetition have been treated quite separately. In extreme cases, separate handling may be entirely appropriate. A writer who reads the phrase *a focal point for* and re-uses it, and the case of a student who downloads an essay from an internet site and submits it for academic credit very different indeed. However, similarities appear in the middle ground, where novice writers

mine published texts for the idiomatic-sounding words and phrases that more experienced writers have stored away in their mental lexicons. It is the broad purpose of this paper to argue that these two areas of repetition, which have hitherto been treated separately, have similarities which can usefully be explored. More specifically, this paper offers an initial step in that exploration by addressing three questions about repeated language:

1. What multi-word units serve a specific function in biology texts?
2. Is the number of realizations of the same idea limited? That is, is it true that there are only so many ways to say the same thing?
3. Is there evidence that some unattributed repetition of language is conventional in some kinds of academic writing?

2. Methods

These questions were investigated using a small corpus of academic writing designed for the purpose. The premise for the third question, that there may be some circumstances in which unattributed repetition of language is conventional, required sensitivity to context. Specifically, such suggestions in the literature relate to what could be termed background information, rather than specific research findings, and have come so far from the sciences (possibly due to the highly unconventional nature of direct quotation in the sciences, e.g., Dubois 1988; Pecorari 2006). This suggested that the corpus should consist of portions of texts which could be regarded as background from a single area within natural sciences.

To address the second question it was necessary to collect a number of realizations of a similar idea. An appropriate candidate emerged from an earlier investigation (Pecorari 2003), in the course of which it was found that research articles in biology often include a brief characterization of the organism under investigation, as in examples (1)-(3) below (numbers in parentheses at the end of examples identify the source article). Such statements appear to work as Step Two of Move One in the 1990 iteration of Swales' CARS model, 'making topic generalizations' (141). They were thus well suited to the present investigation.

(1) *Candida albicans* is a dimorphic opportunistic fungal pathogen that can grow in a yeast or a filamentous phase depending on the environmental conditions. (7)

(2) *C. albicans* is a dimorphic pathogenic fungus that causes superficial and systemic infections in man. (8)

(3) *C. albicans* is the most important human fungal pathogen, causing various forms of superficial and systemic infections in the human host. (13i)

Statements were gathered from the SpringerLink[2] database, which includes international peer-refereed journals in biology. The text of articles was searched for 'albicans is,' omitting *Candida*, as either the full form, as in (1) above, or the abbreviation (2 and 3) may be used. Filters were applied to return results only for research articles written in English.[3] A total of 265 journal articles containing the search string were found. Each article was then searched for the string. A number of hits did not fit the criteria and were eliminated, either because some other species ending in *albicans* was referred to, or because they were not part of background statements (e.g. when 'is' was part of another word, such as 'isolates').

Some articles, on the other hand, contained more than one '*C. albicans* is…' statement. Unsurprisingly, the majority of CA statements came in the introduction or the abstract. However, a few appeared in the discussion sections, when the authors recapitulated their own findings in light of previous knowledge. These and all other statements which fit the basic criterion of presenting background information were included in the corpus, regardless of where in the article they appeared. In total, this process yielded 156 statements from 114 articles, resulting in a small corpus of 2948 words, which was then analyzed with the *AntConc* concordancing software (Anthony 2008).

---

[2] Springer Verlag is gratefully acknowledged for permission to use these articles.

[3] The database classed as 'articles' some texts which are not prototypical research articles, such as a collection of conference abstacts.

3. Findings

Of the three research questions, the first relates to the kind of repeated language which is unconscious and mediated through the mental lexicon, while the other two have to do with the deliberate use of repeated language. Each will be taken up in turn below.

3.1. Repeated language mediated through the mental lexicon

What multi-word units are used? As noted above, multi-word units have been described using a range of labels and attributing to them a diverse range of characteristics. The unit termed 'lexical bundles' by Biber et al. (1999) offers the practical advantage that it is distinguished on the basis of frequency and distribution, disregarding more subjective elements, and is therefore the unit of analysis used here. The criteria set for three- and four-word lexical bundles by Biber et al. (1999: 992-993), that they must occur at least ten times per million words, and in at least five different texts, had to be adapted to the size of this corpus size.

Here strings were counted as lexical bundles if they occurred five or more times in the corpus, and in at least two different texts. This—like any threshold—is to some extent arbitrary. It is also conservative, since to qualify as a lexical bundle, a group of words must be relatively much more common in this corpus than in Biber et al.'s. However, to exclude the possibility of strings being identified as lexical bundles when in fact their co-occurrence was due to other mechanisms of repeated language, absolute frequency of occurrence was important as well as relative frequency. The search string, '*albicans* is,' preceded by either '*Candida*' or '*C.*,' was not counted toward the length of each string since the research design insured that it occurred in every statement in the corpus.

In total five four-word bundles and twenty-two three-word bundles were found (some of the latter occurring as part of the former). No five-word bundles were found. (Biber et al. relax the frequency criterion for bundles of five words and more, but for the reason noted above this was not done here.) There were 151 instances of the three-word bundles (i.e., tokens), a figure which is equivalent to over 51,000 per million words, or rather close to the 60,000 per million words that Biber et al. (1999) found for their academic corpus. They (p. 994) found 5,000 four-word bundles

per million words, while the normalized figure for the present corpus would be significantly higher, over 10,000.

The value of these numerical findings is limited given the small size and specific composition of the present corpus. However, an interesting qualitative feature deserves comment. Biber et al. found that 'in academic prose there are almost no lexical bundles representing complete structural units. Instead, most bundles span two structural units, such as noun phrase + beginning of a prepositional phrase' (1999: 999). Given this functional role, one could speculate that lexical bundles may be relatively high in functional words, and therefore low in specialist terminology. Biber et al. do not address that point directly, but that impression is given by their lists of frequent bundles in academic register. There are many which appear on the face of it to be register-neutral (e.g., 'the end of the,' 'in addition to the,' p. 999) and others which could be described as belonging to general academic discourse (e.g., 'in the present study,' 'in the case of,' pp. 999-1000), but subject-specialist lexis is in short supply. By contrast, as shown in Table 1, a significant number of the lexical bundles found in this study (perhaps 16 of the 27, depending on the criteria applied) contain words related specifically to the subject area (e.g., 'pathogen,' 'dimorphic').

This finding is, of course, readily explained by the fact that the corpus used in the present study is much more specialized than the academic corpus in Biber et al. (1999). It demonstrates, though, that while generic multi-word units are a pervasive part of everyday speech and writing, discipline-specific bundles also exist, a fact which has important implications for the EAP classroom

Table 1. Three- and four-word lexical bundles

Frequency	Bundle	Length
15	the most common	3 words
11	most frequently isolated	
11	opportunistic fungal pathogen	
11	the most frequently	
8	systemic infections in	
6	a dimorphic fungus	
6	an opportunistic fungal	
6	an opportunistic pathogen	
6	and systemic infections	
6	in immunocompromised patients	
6	infections in immunocompromised	
6	most common cause	
6	of superficial and	
6	superficial and systemic	
6	the most frequent	
5	a range of	
5	common cause of	
5	in immunocompromised individuals	
5	member of the	
5	responsible for the	
5	species such as	
5	such as C	
10	the most frequently isolated	4 words
6	an opportunistic fungal pathogen	
5	and systemic infections in	
5	most common cause of	
5	of superficial and systemic	

Another perspective can shed additional light on the formulaicity of these texts. This is the set of structural and lexical similarities which Hoey argues are the result of lexical priming (2005). Language users are primed to use—or avoid—certain words in the environment of others, or in certain grammatical roles, or with certain pragmatic functions, and so

on. Each individual has his or her own set of primings, but they overlap, explaining why some instances of language use are recognizably familiar and idiomatic. Because shorter groups of words can be combined with others, and these combinations may prefer certain grammatical roles, Hoey says 'some sentences. . . are made up of interlocking collocations such that they could be said to reproduce, albeit with important variations, stretches of earlier sentences' (2002, cited in Hoey, 2005, p. 5). Hoey illustrates this phenomenon with a sentence from a Bill Bryson book which begins:

> In winter, Hammerfest is a thirty-hour ride by bus from Oslo. . . (p. 5).

He then identifies the pattern found in this example as:

> SMALL PLACE is a NUMBER-TIME-JOURNEY—(by vehicle)—from LARGER PLACE

and offers further examples from his corpus:

> 1 Ntobeye is a two-hour ride by four-wheel drive vehicle from the vast refugee camp at Ngara.
> 2 The village is a four-hour drive from London.
> 3 Pamuzindo is an hour's drive from Harare. (Hoey, 2005: 18)

The sense of an underlying structural template found in these examples is present in the CA statement as well. In all but 25 of the 156 statements, the search string ***albicans* is** is followed by a noun phrase (20 are followed by adjective phrases and in five **is** is part of a passive verb phrase).

This large pattern contains smaller ones. The most common head is **pathogen(s)**, which occurs in this position 43 times. Most of the NPs have both premodifier and postmodifier. In 25 cases the premodifier characterizes the prevalence of the organism (**most common, commonly occurring, important)**. Of those, all but eight also have a postmodifier, and eight postmodifiers describe *C. albicans* as occurring in humans:

(4) *Candida albicans* is the most common opportunistic pathogen of humans. (195)

(5) *Candida albicans* is the most frequently isolated fungal pathogen in humans. (213)

(6) The yeast *Candida albicans* is an important fungal pathogen in man. (166ii)

This structure, found in 5% of the statements in the corpus, can be described as:

Candida albicans is (a) PREVALENT pathogen FOUND IN HUMANS

It could be argued that this pattern is an artefact of the research design, since a sentence starting *NP is* is likely to be followed by another NP or an AdjP in subject predicative position, since within the NP the sequence premodifier-head-postmodifier is expected, and since the words chosen to fill those slots are a function of the specialized nature of the corpus.

This argument does not explain, however, the strong preference for NP instead of other apparently equally appropriate realizations with AdjP. Example (6), for instance, could be reworded as **an important, pathogenic fungus in humans**, but while **fungal pathogen** occurs 24 times in the CA corpus, **pathogenic fungus** occurs only four. This preference may be the result of those two words being primed to occur in one configuration more strongly than the other. This argument would also overlook the importance of the very existence of the CA statements. The fact that in a single database there were 151 sentences consisting of ***Candida albicans* is + subject predicative** suggests that it is indeed a pattern within the discourse community of biology.

3.2. Deliberate use of repeated items

Are there only so many ways to say it? When novice academic writers are confronted with the challenging demands of reformulating ideas from an authoritative text, without distortion, in an original form—i.e., the task of paraphrasing—they sometimes protest that originality of expression is an unachievable objective because there are only so many ways to say the same thing. This corpus, consisting of 156 iterations of very similar ideas, allows the truth of that idea to be examined.

The CA statements do in fact draw on a rather narrow pool of lexis. The 2948 words consist of 642 types, of which *hapax* and *dis legomena* account for 465 types, or 72%. In other words, 28% of the types in the corpus occurred at least three times. By comparison, a sample of precisely the same length drawn from a corpus of research articles in

biology, but without the additional constraints applied here, was found to contain 945 types, of which 742 (79%) were *hapax* or *dis legomena*. Excluding ***Candida/C. albicans*** **is**, 18 lexical words have an average frequency of one occurrence per ten statements (Table 2). Unsurprisingly, then, the corpus of CA statements uses a rather restricted number of lexical items, even when compared to other biology writing.

Table 2. Frequency of lexical words

Frequency	Word	Frequency per 100 words
50	most	1.7
48	pathogen	1.63
39	infections	1.32
39	yeast	1.32
36	fungal	1.22
33	species	1.12
31	opportunistic	1.05
23	cause	.78
23	isolated	.78
23	patients	.78
21	dimorphic	.71
20	common	.51
19	human	.64
19	immunocompromised	.64
19	systemic	.64
18	frequently	.61
18	responsible	.61
17	fungus	.58

As noted above, by far the most common pattern among the CA statements was the search string followed by an NP as subject predicate. The NPs had 14 heads (Table 3) as well as various combinations of determiners and modifiers. Premodifiers were more numerous than heads, both in types and in tokens (since there was often more than one premodifier). They cover four broad semantic categories: the prevalence

of the organism, its characteristics, its effect, and where it is found (Table 4).

Table 3. Heads of NPs

Head	Frequency[4]	Head	Frequency[4]
agent	7	member	3
cause(s)	14	organism	7
colonizer	1	part	2
commensal	4	pathogen(s)	43
fungus/i	14	problem	1
infections	1	species	18
isolate	2	yeast	14
Total		131	

The postmodifiers were longer, and as a result both more diverse and more difficult to classify. However, the two largest categories correspond to the last two in Table 3, describing the effects of Candida albicans and where it is found (examples 7 and 8, respectively).

(7) the most common etiological agent of both superficial and deep-seated candidiasis. (69)

(8) a commonly occurring pathogen in the human population, and in particular in patients undergoing cancer chemotherapy. (52)

The range of purposes accomplished by the CA statements was, therefore, rather narrow. The writers classified CA in terms of how often it is found, where it occurs, what it does, and its inherent characteristics. The range of lexis used to do these things is predictably limited.

[4] Frequency as head of NP in subject predicative position, not in the entire corpus.

Table 4. Premodifiers in NPs

Premodifying expressions	Tokens
FREQUENCY	86
most frequently isolated, reported, (one of the) (most) frequent	24
(one of) (among) (the) (second, third, fifth) most common, most commonly isolated, commonest, commonly occurring	23
major, main, primary	15
important, most important, clinically important, medically important	11
predominant, predominating, dominant	7
best studied, most prevalent, typical, ubiquitous, pervasive, normal	6
CHARACTERISTICS	79
fungal	27
opportunistic	26
polymorphic, dimorphic	18
Candida, candidal	3
yeast, yeast-like	2
sentinel, invasive, most adapted	3
EFFECT	20
causative, causal, etiologic/al	10
pathogenic	7
infectious, harmless, symptomatic	3
WHERE IT ARISES	14
human	6
commensal	3
nosocomial	3
genitourinary, oral	2

Despite these shared characteristics in the statements, though, it cannot be said that the writers found only one way, or even just a few ways, of describing the organism. The statements described a limited number of characteristics of *C. albicans*, but they selected smörgåsbord-style from that range, choosing (presumably) the most relevant descriptors which for their own studies, and that created differences in the statements.

This can be illustrated with the subset of statements which described CA as a cause of something. Two head nouns and two premodifiers are directly related to causation, *cause* and *agent*, and *causative* and *etiological*. Thus four options are immediately available:

(9) *Candida albicans* is a cause. . .

(10) *Candida albicans* is an agent. . .

(11) *Candida albicans* is a causative agent. . .

(12) *Candida albicans* is an etiological agent. . .

As the ellipsis in the sentences above indicates, both *cause* and *agent* are likely to demand a prepositional phrase as postmodifier, yielding many more possibilities:

(13) cause of intravascular catheter related infections (101iii)

(14) cause of nosocomial blood stream infection (106)

(15) cause of infections in immunocompromised patients (142ii)

In addition, the claim that CA is a cause of something can be qualified (a common cause; one of the most common causes) a number of ways. Further variety can come from using a head noun unrelated to causation, with a premodifier such as 'causative.' Still greater variety is created by introducing additional propositions unrelated to the effects of CA. The result is that even when the objective is as narrow as identifying a single organism as the cause of infection and related complications, many different formulations are possible:

(16) *Candida albicans* is the main cause of systemic fungal infections for which there is an urgent need for novel antifungal drugs. (260i)

(17) *Candida albicans* is the most common etiological agent of both superficial and deep-seated candidiasis. (69)

(18) As a pathogen, *Candida albicans* (*C. albicans*) is the causative agent in 60%–80% of Candida infections, including 85%–90% of VVC. (46)

Even in this very circumscribed context, it is therefore not true that there are strictly limited ways available to say the same thing. However, they do give an impression of similarity which might lead a novice writer to speculate that recycling language from other texts is a common practice. That is the final question to be addressed here.

3.3. Unattributed repetition of language

Is background language 'borrowable'? Must all parts of an academic text be equally original? Or is it the case, as some writers have suggested, that some unattributed repetition of language may be appropriate? Background information, like the CA statements are precisely the sorts of texts those writers appear to mean. Did the writers who produced the CA statements use reproductive strategies? In other words, are the similarities created by the writers drawing phrases directly from other texts?[5]

It must be acknowledged at the outset that there are two obstacles to answering this question. The first is that sources for the CA statements, if they exist, could come from elsewhere than the database used here. It is therefore possible to establish if repetition *has* occurred, but not the reverse. The second problem is determining whether similarities are due to repetition, or coincidental. That could be done in principle either by demonstrating that copying is statistically more likely, or more intuitively. However, despite efforts in that direction (e.g., Turrell, 2004) there is no reliable statistical threshold which could be applied to these cases. The intuitive option, on the other hand, would result in different individual judgements. Extreme cases might be clear, but shorter chunks, or ones with some superficial differences, will be less so. It is therefore possible that individual readers may differ in terms of how they evaluate the likelihood that copying took place. These two caveats should be

[5] It is important to emphasise the limited scope of this question, and the fact that it does not include whether plagiarism exists in the corpus. While many consider unattributed repetition of language to be plagiarism, as noted above, others would not. In addition, determining the presence of plagiarism requires attention to contextual factors which are not within the scope of this investigation.

borne in mind in looking at the findings of repeated language in the corpus.

The n-gram function in *AntConc* (Anthony 2008) was used to identify occurrences of strings of a given length, *n*, occurring a stipulated number of times (in this case, twice). The maximum length of *n* was raised progressively until the longest repeated string in the corpus had been identified. In addition, some very similar but not identical strings were identified, for example when otherwise similar strings used two different synonymous words. Although it was not possible to account for all such cases systematically, those that were found are presented below.

Three groups of very long, identical strings (39, 25 and 23 words, respectively) were identified. All three are long enough to suggest that their similarity is due to one instance being copied from another.

(19a) Although in most studies *Candida albicans* is still the most frequent cause of candidemia, there is an increase in the isolation of non-albicans *candida* strains, such as *C. parapsilosis*, *C. krusei*, *C. tropicalis*, and *C. glabrata* [1]. (186i)

(19b) Although in most studies *C. albicans* is still the most frequent cause of candidemia, there is an increase in the isolation of non-albicans *Candida* strains [2–6] (186ii)

(20a) Although *Candida albicans* is the most frequently isolated yeast from clinical specimens, the emergence of non-*albicans* species has clearly been a recent concern. (208i)

(20b) Although *C. albicans* is the most frequently isolated yeast from clinical specimens, the emergence of non-*albicans* species has clearly been a recent concern. (208ii)

(21a) Although *C. albicans* is by far the predominant isolate in this condition other non-albicans species such as *C. tropicalis* and *C. glabrata* (syn. *Torulopsis glabrata*) are frequently isolated both from the acrylic denture surfaces and the palatal mucosa [4]. (152)

(21b) Although *C. albicans* is by far the predominant isolate in this condition other nonalbicans species such as *C. tropicalis* and *C. giabrata** (syn. *Torulopsis glabrata*) are frequently isolated both from the acrylic denture surfaces and the palatal mucosa [4]. (203) (**sic*)

However, in two of the three groups (examples 19 and 20), both occurrences of the string came from the same article. In the third group (21), the two articles each had five authors, of whom the first, second, third and fifth were the same people. It thus appears that the answer to

the question, 'is background language "borrowable"?' is at most a qualified 'yes.' Long strings of repeated text are *not* common in this corpus, occurring only six times among the 156 CA statements. Further, in each case the 'borrowing' was not from a different author, but either from an earlier part of the same text, or from the writers' earlier text. There is no evidence, therefore, that verbatim copying of long chunks of text is a common practice in this disciplinary community.

Other cases were less obvious, though. Several cases were found of language which was similar but not identical either because synonyms were used, or because one included detail which the other omitted. Thus, (22a) uses 'human microflora' while (22b) has 'body microflora,' and the phrase 'ranging from superficial to systemic mycoses' is found only in (a).

(22a) *Candida albicans* is an opportunistic fungal pathogen that may be present as a normal component of the human microflora. It is responsible for a variety of diseases in the immuno-compromised or immuno-suppressed hosts ranging from superficial to systemic mycoses (Cotter and Kavanagh 2000). (85i)

(22b) *Candida albicans* is an opportunistic fungal pathogen that may be present as a normal component of the body microflora. It is responsible for a variety of diseases especially in immunocompromised and immunosupressed hosts (Cotter & Kavanagh 2000). (212)

A second group is similar, in that the text in (23b) has extra detail ('the gastrointestinal and urogenitary tracts'), as well as small differences in wording: (e.g., 'healthy humans' versus 'many healthy people'). In addition, only one has a citation. The papers from which (23a) and (23b) come have three and seven authors, respectively, and the last in each is the same person. In (22), however, there is no overlapping authorship, and the authors have institutional affiliations in different countries.

(23a) The yeast *Candida albicans* is a member of the microflora on mucosal surfaces of healthy humans, but it can also cause serious infections, especially in immunocompromised patients. (247)

(23b) The yeast *Candida albicans* is a member of the microflora on the mucosal surfaces of the gastrointestinal and urogenitary tracts in many healthy people, but it can also cause opportunistic infections, especially in immunocompromised patients (Odds 1988). (252)

In these last two pairs the similarities are strongly suggestive that either the later in the group copied from the earlier, or that both drew

their descriptions from a common source (the question of which was the case is not germane to this investigation, which is concerned with the frequency of repeated language and not its sources). Other cases (e.g., 24-26) have fewer words in common (words which appear in all statements in each example are underlined). In addition, it is possible that those that are in common are 'generic' chunks. This makes it difficult to establish whether copying or coincidence is the explanation for their similarity.

(24a) *Candida albicans* is the most frequently isolated fungal pathogen in humans and is responsible for a wide variety of . . . (116)
(24b) *Candida albicans* is the most frequently isolated fungal pathogen in humans. (213)
(24c) Fungal infections have become increasingly significant due to the growing population of immunocompromised patients and *C. albicans* is the most frequently isolated fungal pathogen of nosocomial infections [24]. (48iii)

(25a) In humans *Candida albicans* is the most frequently isolated opportunistic fungal pathogen. (184i)
(25b) *Candida albicans* is the most frequently isolated opportunistic fungal pathogen. (184ii)
(25c) *Candida albicans* is the most important opportunistic fungal pathogen. (261ii)

(26a) *Candida albicans* is a dimorphic pathogenic yeast capable of producing alternate morphological forms (yeast or mycelium) in response to environmental changes. (166i)
(26b) It is well known that *Candida albicans* is a dimorphic pathogenic yeast capable of producing yeast (Y) or mycelial (M) forms in response to environmental conditions. (255)

The difficulty of understanding the origins of some similar but not identical language is illustrated by the nine statements in Table 5, the structure of which could be represented as:

(APPOSITIVE) *Candida/C. albicans* is (X) (and) CAUSES Y (in Z) (REFERENCE).

Some slots (those indicated by parentheses) are filled in only some of the statements, and that, together with the various ways of realizing the slots in uppercase letters, results in different, but strikingly similar, statements. For example, the first two differ from each other only in the presence of

the appositive phrase 'the dimorphic yeast' and a reference. Others, while conforming to the pattern, are less similar, making it difficult to say whether the similarities are the result of coincidence or repetition.

Table 5. Structural similarities in selected CA statements

Candida albicans is	an opportunistic fungal pathogen	capable of causing	a range of superficial and systemic infections	in the immuno-compromised host. (219)
The dimorphic yeast *Candida albicans* is	an opportunistic fungal pathogen of humans	and is capable of inducing	a range of superficial and systemic diseases	in the immunocompromised host [1]. (229i)
The yeast *C. albicans* is	an opportunistic fungal pathogen	capable of causing	serious systemic infections	in immuno-compromised individuals (e.g. patients undergoing chemotherapy, neutropenia). (28iii)
C. albicans is	the most important human fungal pathogen,	causing	various forms of superficial and systemic infections	in the human host. (13i)
The yeast *Candida albicans* is		responsible for	a range of superficial and systemic diseases	in the immunocompromised patient. (228)
Candida albicans is	a medically important yeast	that causes	a spectrum of superficial and systemic infections	in human hosts. (264)
The yeast *Candida albicans* is	an opportunistic fungal pathogen	and causes	a range of diseases	in susceptible individuals (Pfaller et al. 1998). (146)
Candida albicans is	the main opportunistic fungal pathogen of humans	that has increasingly been found to cause	systemic infections	in immunocompromised patients (Beck-Sagué and Jarvis 1993). (260ii)

Candida albicans is	notable because, although it commonly enjoys a commensal relationship with humans and other animals it may become pathogenic,	causing	a range of superficial and deep-seated infections.	(162)

The difficulty of understanding the origins of cases like these presents an obstacle to answering the third research question, whether it is conventional, or at least acceptable, to appropriate the language used to present 'background' information. Given that only isolated examples of manifestly repeated language were found, there is not sufficient evidence to say that repetition is a common strategy. On the other hand, the similarities found in statements like the ones in Table 5 make it difficult to refute claims that such borrowing *is* permissible. The similarities are, however, such that it is understandable that some writers have formed the view that repeating background information *is* common practice.

4. Discussion and conclusions

This paper has shown that repeated language has a real presence in writing in biology. Three questions related to repetition in language use, albeit from three rather diverse perspectives, were addressed. The first, what lexical bundles are used in biology writing, will ultimately require further research on a larger and less specialized corpus. However, the evidence of the present study is that some lexical bundles are discipline-specific. EAP practitioners should be aware of this as yet another language feature existing on the boundary between our discipline and our students', and to be negotiated in collaboration with them.

The second question was whether unoriginal ideas (a common feature of RA introductions), must be expressed in unoriginal language. If 'unoriginal' means language identical to a source, or virtually so, then the evidence presented here suggests that the answer is 'no.' The same idea can be formulated in a number of ways which are grammatical, idiomatic and responsive to the rhetorical demands of the larger text.

The third question asked whether source-dependent writing in biology articles is as common as it is in some student writing, and in novice writers' accounts of their practices (Flowerdew & Li, 2008; Pecorari, 2003). Here the answer has been seen to be a qualified 'no.' Long chunks of text identical, or nearly so, to other works on the same topic were found to be unusual. When they occur, there tend to be mitigating circumstances, with the duplicate language occurring in two portions of the same text, or in two articles with one or more authors in common.

Relatively more common, though, were chunks of language which are sufficiently similar to others in the corpus that repetition is a plausible hypothesis, but insufficiently similar to rule out other explanations, such as nesting lexical bundles. It is in these cases, more difficult to classify, that the rather separate types of repetition examined here merge. Is it simply that the origins of these cases are difficult to identify, or is there a sort of repetitive language use which is neither deliberate copying nor simple formulaic language use? Are these unclear units the result of second-language writers reaching into published texts for the idiomatic forms of expression they know exist but cannot access through the mental lexicon? If so, is the use of such 'loan-chunks' acceptable to the wider academic community? If it is not, what other options are open to non-native speakers of English trying to disseminate their research to a global community? These are questions for future investigation.

There are several implications for the EAP teacher from this work; unfortunately, they are somewhat contradictory. Novice writers have an apparent tendency to recycle formulaic language (Flowerdew & Li 2008; Pecorari 2003), and source-repetitive writing strategies allow novice writers to expand their personal lexicon (Villalva 2006). Biology RAs have been shown to consist in part of a stock of lexical bundles and standard formulations. It is tempting, therefore, to suggest that EAP instructors encourage writers to harvest language from published texts.

Set against that is the fact that some EAP writers who have used this strategy—such as the Turkish scientists mentioned in the introduction to this paper—have been subjected to accusations of academic misconduct. Writers who use this apparently beneficial and widespread strategy can find that it backfires more violently than any other unsuccessful writing strategy possibly could.

This creates a dilemma for the EAP specialist. In promoting the use of language from texts, she may encourage students to do something for which they will be sanctioned later. In discouraging it, she may block avenues for learning. If she remains quiet on the subject, she neither exploits the pedagogical benefits nor warns writers about potentially dangerous strategies. If she addresses the issue and draws on her experience and intuition to distinguish between what is acceptable and what is not, those views may not be acceptable to other academics, since there is not consensus in the academic community about what types of repeated language are appropriate, and, perhaps more dangerously, little recognition of the diversity of views that exist (Pecorari, 2008).

This is not a question which EAP teachers can solve; before we can help our students discover what sorts of practices are acceptable in their disciplines, there must be some sort of consensus within and across the disciplines. There is an urgent need, therefore, for an academy-wide conversation. In the absence of answers, the EAP teacher can perhaps best serve her students by making them aware of the questions, discussing this issue in the classroom, and thus starting the conversation at a local level.

References

Angélil-Carter, Shelley. 2000. *Stolen Language? Plagiarism in Writing*. London: Longman.

Anthony, Laurence. 2008. *AntConc 3.2.2*.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Harlow, UK: Longman.

Bolinger, Dwight. 1976. Meaning and memory. *Forum Linguisticum 1*: 1-14.

Brumfield, Geoff. 2007, 6 September. Turkish physicists face accusations of plagiarism. *Nature*.

Charles, Maggie. 2006. Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines. *English for Specific Purposes, 25*: 310–331.

Cortes, Viviana. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*: 397-423.

Currie, Pat. 1998. Staying out of trouble: Apparent plagiarism and academic survival. *Journal of Second Language Writing, 7*: 1-18.

Dubois, Betty Lou. 1988. Citation in biomedical journal articles. *English for Specific Purposes, 7*: 181–193.

Erman, Britt and Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text, 20*: 29-62.

Flowerdew, John and Li, Yongyan. 2007. Language re-use among Chinese apprentice scientists writing for publication. *Applied Linguistics, 28*: 440–465.

Hoey, Michael. 2005. *Lexical priming: A new theory of words and language*. London: Routledge.

Howard, Rebecca Moore. 1995. Plagiarisms, authorships, and the academic death penalty. *College English, 57*: 788-805.

Howard, Rebecca Moore. 1999. *Standing in the shadow of giants*. Stamford, CT: Ablex.

Hull, Glynda and Mike Rose. 1989. Rethinking remediation: Toward a social-cognitive understanding of problematic reading and writing. *Written Communication, 6*: 139-54.

Hunston, Susan and Gill Francis. 2000. *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Nattinger, James R. and Jeanette DeCarrico. 1992. *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Pecorari, Diane. 2006. Visible and occluded citation features in postgraduate second-language writing. *English for Specific Purposes, 25:* 4-29.

Pecorari, Diane. 2003. Good and original: Plagiarism and patchwriting in academic second-language writing. *Journal of Second Language Writing, 12*: 317-345.

Pecorari, Diane. 2008. *Academic writing and plagiarism: A linguistic analysis*. London: Continuum.

Peters, Ann M. 1973. *The units of language acquisition*. Cambridge: Cambridge University Press.

Salager-Meyer, Françoise. 1999. Referential behavior in scientific writing: A diachronic study (1810–1995). *English for Specific*

Purposes, 18: 279-305. Schmitt, Norbert (Ed.). 2004. *Formulaic sequences: Acquisition, processing and use*. Amsterdam: John Benjamins.

Spack, Ruth. 1997. The acquisition of academic literacy in a second language: A longitudinal case study. *Written Communication, 14*: 3-62.

Swales, John M. 1986. Citation analysis and discourse analysis. *Applied Linguistics, 7*: 39-56. Tagliamonte, Sali and Alex D'Arcy. 2004. He's like, she's like: The quotative system in Canadian youth. *Journal of Sociolinguistics 8*: 493-514.

Thompson, Paul. 2000. Citation practices in PhD theses. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective: Papers from the third international conference on Teaching and Language Corpora*. Frankfurt: Peter Lang. 91-101.

Turrell, T. 2004. Textual kidnapping revisited: The case of plagiarism in literary translation. *International Journal of Speech, Language and the Law, 11*: 1-26.

Villalva, Kerry Enright. 2006. Hidden literacies and inquiry approaches of bilingual high school writers. *Written Communication, 23*: 91-129.

White, Howard D. 2004. Citation analysis and discourse analysis revisited. *Applied Linguistics, 25:* 89-116. Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Yilmaz, Ihsan. 2007, 11 October. Plagiarism? No, we're just borrowing better English. *Nature*.