Using online corpora to develop students' writing skills

Alex Gilmore

Large corpora such as the British National Corpus and the COBUILD Corpus and Collocations Sampler are now accessible, free of charge, online and can be usefully incorporated into a process writing approach to help develop students' writing skills. This article aims to familiarize readers with these resources and to show how they can be usefully exploited in the redrafting stages of writing to both minimize the teachers' workload and encourage greater cognitive processing of errors. An exploratory investigation comparing the use of these two online corpora in Japanese university writing classes is then described. This suggests that the participants in the study were able to significantly improve the naturalness of their writing after only a 90-minute training session and that the majority of students found these online resources beneficial, although there was a marked preference for the COBUILD Corpus and Collocations Sampler.

Introduction

> Writing is easy. All you do is sit staring at a blank sheet of paper until drops of blood form on your forehead.
>
> (Gene Fowler, American journalist and biographer)

Writing can be a slow, painful process even in our mother tongue, but when it is in a second language the problems (and the pain) are magnified. Given the amount of conscious effort involved in the writing process, learners in ESOL classrooms understandably expect feedback on their work and may feel discouraged if it is not provided (Hedge 1988). The difficulty for the teacher, however, is in balancing the needs of individual students for meaningful feedback with the unfortunate reality of ever-increasing workloads. In Japanese universities, for example, it is not uncommon for teachers to have at least three concurrent writing classes, with 20 or so students per class. This means around 60 essays to mark for each assignment (assuming that only one draft is requested), but since the process writing approach often demands rewriting of initial drafts, this number can easily double or triple.

Providing feedback on sentence-level errors

Investigations over the last 30 years into the benefits of providing students with feedback on sentence-level errors in their writing have been inconclusive, with researchers arguing strongly both for and against it (Truscott 1996, 1999, 2004; Ferris 1999, 2004; Chandler 2004; Hyland and Hyland 2006). The reason for the lack of concrete answers to this issue

ELT Journal Volume 63/4 October 2009; doi:10.1093/elt/ccn056

Advance Access publication November 11, 2008

lies in failures 'to investigate questions surrounding error correction in L2 writing in a sustained, systematic, replicable manner that would allow for comparisons across either similar or different contexts and student populations' (Ferris 2004: 55). However, as Ferris (ibid.) quite rightly points out, teachers cannot afford to wait for answers to these important questions and must fall back on their own intuitions and experience instead. Personally, I find it very hard to believe that the 'scaffolding' provided through error feedback would not benefit those students who are willing to actively engage in the process of redrafting, and, judging from post-course feedback I have received from learners in my own writing classes, it is certainly appreciated.

Teachers have a number of options available when it comes to providing feedback on students' sentence-level errors in written work, and these can be placed on a cline, in terms of their comprehensiveness, with each choice having associated advantages and disadvantages.

Feedback method	Advantages	Disadvantages
1 Complete reformulation of errors by teacher.	Students receive accurate and comprehensive feedback, which specifically addresses their language needs.	Time consuming for teacher. Does not encourage cognitive processing of errors by students so there may be no long-term benefits. The quantity of corrections may discourage students.
2 In-class peer feedback.	Reduces teacher's workload. Provides a wider audience for students' work, which can have a motivating effect. Encourages greater cognitive processing of errors by students and promotes learner independence. Encourages collaboration and negotiation of meaning in the classroom.	Students require training in how to give constructive feedback, which takes time away from actual writing practice. May be perceived as less valuable feedback by students themselves. Time-consuming in-class activity. Feedback can be (a) wrong or (b) less helpful than teachers' comments.
3 Selective feedback by the teacher on specific issues or target language of current concern.	Reduces teacher's workload. Feedback can be tailored to ongoing themes in the class.	Less comprehensive feedback provided, which may not address students' own particular concerns.
4 Minimal marking (marking codes, underlining problem areas, etc.).	Reduces teacher's workload. Encourages greater cognitive processing of errors by students.	May not provide sufficient support for less proficient students to correct errors by themselves.
5 No feedback on errors.	Reduces teacher's workload. Increases the amount of time available for actual writing practice, which should benefit students' writing fluency.	Provides no support or encouragement for students to correct errors. Goes against students' desire for feedback and may cause frustration.

TABLE 1
Advantages and disadvantages of various feedback options for written work

As Table 1 shows, teachers face difficult decisions on how to best utilize limited time and resources both inside and outside the classroom. A greater focus on accuracy is likely to reduce the amount of actual writing practice students get and affect their fluency, while less attention to mistakes may deprive them of the tailored feedback they need to develop their interlanguage. In addition, practical realities can often outweigh any pedagogical considerations, with teachers simply too busy to provide more individual feedback even when they believe it would be beneficial.

Using online corpora in the classroom

One teacher-friendly way to encourage students to focus more on error correction, while at the same time providing them with the support they need, is to train them in methods to query online corpora such as the British National Corpus (BNC) (http://www.natcorp.ox.ac.uk/) or the COBUILD Concordance and Collocations Sampler (http://www.collins.co.uk/Corpus/CorpusSearch.aspx). These are large collections of texts (books, newspapers, journals, transcribed speech, etc.), produced by native speakers of English, which are stored electronically and can be accessed using search software. Users type in a query word or phrase to generate 'concordance lines' (randomly selected lines of text containing the target language) which are extracted from the corpus. The key characteristics of these two free online corpora are summarized below in Table 2.

BNC	COBUILD
■ 100 million word collection	56 million word collection
■ British English	British and American English
■ Written/spoken English	Written/spoken English
■ Not possible to search subcorpora	Option to search subcorpora (British; USA; spoken)
■ Collocation information not available	Collocation information available
■ Up to 50 randomly selected concordance lines displayed	Up to 40 randomly selected concordance lines displayed
■ Information on the source text for concordance lines available	No information on the source text for concordance lines available
■ Generally slower query response times (approximately 10 seconds in my trial)	Generally faster query response times (approximately 4 seconds in my trial)
■ Less user-friendly: keyword or phrase not highlighted or positioned centrally in concordance lines.	More user-friendly: keyword or phrase highlighted and positioned centrally in concordance lines.

TABLE 2
Key characteristics of BNC and COBUILD online corpora

It seemed to me that these freely available online resources could usefully be incorporated into the redrafting stages of a process writing approach, by highlighting problematic areas in students' essays and then allowing them to use the corpora to generate their own hypotheses on how to make their writing more natural—an inductive approach, known as data-driven learning, most commonly associated with its originator, Tim Johns (for example Johns 1986). This way of dealing with error correction is more in line with constructivist theories of learning from developmental psychology, which see individuals as active participants in the construction of their own personal meaning from the experiences they have (Williams and Burden 1997). With each learner's interlanguage system in its own unique stage of

development, inductive approaches, which encourage students to find their own solutions to their own particular problems, are more likely to create the conditions necessary for language acquisition to occur. Of course, as classroom activities, they are also more time consuming, but the increased cognitive work they require should also lead to greater learning gains (for example Cobb 1997).

The remainder of this paper describes an exploratory investigation into the use of online corpora to develop students' writing in Japanese university English for Academic Purposes classes, looking at both the effects on the naturalness of their redrafted essays as well as the learners' own reactions to the approach.

The investigation

Method

The aim of the investigation was two-fold:

a to determine whether training learners in the use of online corpora would have any noticeable effect on the 'naturalness' of their redrafted essays;
b to explore learners' reactions and preferences regarding the BNC and COBUILD online corpora.

Forty-five second-year intermediate-level Japanese university students, enrolled on a compulsory academic writing course, were asked to write a factual report based on the theme of 'obsession'. Sentence-level, lexical, and grammatical problems in students' first drafts were highlighted by the teacher, by underlining, and the essays were then returned for redrafting. Prior to any corrections being made, learners received a 30-minute introduction on how to use online corpora (see Appendix 1) and then spent 1 hour in the computer room, comparing the usefulness of the BNC and COBUILD corpora to clarify problems with their writing. Students were then asked to produce second drafts of their essays outside of class, correcting problem areas identified in their first drafts by referring to one or both of the online corpora introduced. Sentences identified as problematic in the first drafts were isolated and compared with the revised versions in the second drafts and blind-rated for 'naturalness' (i.e. raters were not told which draft version the sentences had come from) by four native-speaker teachers (see Appendix 2). Finally, students were also asked to comment on

a the usefulness of online corpora for improving their writing;
b their preferences for either the BNC or the COBUILD Concordance and Collocations Sampler.

Results

A total of 350 lexical or grammatical problems were identified in the 45 texts analysed, with a range of 1–17 issues occurring in each student's essay. From the changes made by students between the first and second drafts, 214 (61.14 per cent) were rated as more natural, 114 (32.57 per cent) as equivalent, and 22 (6.29 per cent) as less natural by the native-speaker raters. Examples of the kinds of modifications made are illustrated below:

More natural:
First draft: He became popular in the USA not only Japan.
Second draft: He became popular not only Japan but also in the USA.

Equivalent:
First draft: Human body burns energy to keep life-maintenance.
Second draft: Human body burns energy for keeping life-maintenance.

Less natural:
First draft: Underage smoking was prohibited in Japan, so she couldn't avoid fired.
Second draft: Underage smoking was prohibited in Japan, so she couldn't evade displacement.

Student feedback on the activities was generally very positive with 95 per cent of respondents believing that online corpora were a useful resource to aid them in redrafting their essays. The reasons cited typically mentioned the autonomy corpora allowed, the ease with which numerous concordance lines could be accessed and the fact that the examples shown were 'real English':

YM I think corpora is useful because we can check our mistake by ourselves and it can help us in many ways.

NM Corpora are very useful for me because I can get many example sentences very quickly from these sources.

MY I think it is very useful for me because I can know the native speaker's sentences. In my dictionary there are many sentences but they are not natural sentences.

The students who found the corpora to be less useful generally emphasized the difficulty in either accessing the information they needed to correct their mistakes or understanding the concordance lines generated:

AM It's a little difficult to understand when the sentence has words I don't know.

JI We couldn't know the right answer. We didn't know which is mistake.

In terms of preferences, 84.5 per cent of students preferred the COBUILD Concordance and Collocations Sampler to the BNC, typically stating that it was more user-friendly and faster:

SN Because the searching word were put on the same position, it's very easy to find.

KU BNC is too slow. For search one sentence, it spends 1–2 minutes. Also COBUILD is more convenience than BNC.

The 15.5 per cent of students who preferred the BNC tended to emphasize its size (which at almost twice that of COBUILD means that 'hits' are more likely with less frequent lexis):

KS I prefer BNC because it has more words than COBUILD.

Discussion

Since around 61 per cent of changes made to students' first drafts, with the support of online corpora, resulted in more natural language, we can safely say that this is an approach worthy of further investigation. These results concur with those of other researchers who have already demonstrated that learners are able to make corrections based on concordance evidence (for example Todd 2001; Gaskell and Cobb 2004). Of course, in this

particular study, there was no control group so it is impossible to say to what degree the improvements seen can be attributed to the training given. It could be that simply by highlighting problem areas by underlining, students are able to significantly improve the naturalness of their writing with the help of more traditional reference sources (grammar books or dictionaries)—I suspect that this is not the case though. It is also worth mentioning that the reported results are somewhat distorted by subjects who made no effort to improve their second drafts. These students simply printed out their first drafts again and handed in identical work, thus increasing the number of 'equivalent' ratings.

The very high approval ratings seen from participants in this investigation provide further support for the use of online corpora in the classroom. It was clear, however, that some students, particularly those of lower proficiency, found both the selection of keywords for their searches and the interpretation of the resulting decontextualized concordance lines difficult. Because the whole sentence or clause containing mistakes had been underlined, it was not always obvious to learners exactly what to search for, and the wrong choice could easily produce misleading information. For example, with the phrase 'Since then, he started to go . . . ' from Appendix 1, a search for 'started' could produce concordance lines such as these (retrieved from the COBUILD corpus):

the first time. People who simply couldn't get **started** without our help. But we desperately need to

but also very subtle. After that I **started** buying albums by Herbie, and I got a few

did you hear that Salif Keita's wife has **started** a musical movement? They're called the Griot

the spool, stating 4X. As in the old days one **started** with silkworm gut approximately 11

I said I'd rather starve, that's when the band **started** working [p] I worked at a recycling plant

Samples like these might lead students to conclude that 'started' is always followed by an '-ing' clause, causing them to rewrite the sentence as 'Since then, he started going . . . ' in their second drafts. The most obvious solution to this problem is to increase the amount of support provided by, for example, circling the keyword/phrase to search with, in addition to underlining the clause or sentence containing the error:

'Since then, he started to go . . . '.

Concordance lines could also be edited by the teacher so that only clear examples are displayed, although this would be time consuming and impractical with large classes.

The clear preference by students for the COBUILD Concordance and Collocations Sampler over the BNC is understandable, given its more user-friendly characteristics. Searches tend to be faster since it is a smaller corpus and, most importantly, the search word is easy to locate because it is positioned centrally and displayed in boldface: for learners with less

proficient scanning skills, this is a great help when faced with a series of decontextualized concordance lines. The ability to search British, American, and speech subcorpora, and to investigate common collocations, is also a useful feature of the COBUILD corpus. These comments, of course, only relate to the specific use of the free versions of these corpora, available online, with non-native speakers of English. The BNC, now available on DVD as an XML Edition, is an excellent resource for language teachers and researchers.

Conclusion

The approach to error correction suggested here will clearly appeal to some types of learners more than others; for example those who are more visually oriented, more analytic and logical, or less tolerant of ambiguity. Nevertheless, the results of this exploratory investigation do suggest that online corpora may well have a valuable role to play in the redrafting stages of classes adopting a process writing approach. After only a 90-minute introductory session, students appeared to be able to use these resources effectively to improve the naturalness of their writing and the vast majority of them found the training useful. For busy teachers, online corpora can reduce their workloads by providing learners with the support they need to make corrections autonomously, without the necessity of lengthy explanations in the margins. Underlining problem areas in students' work is quick to do and frees up time to concentrate on more global issues of cohesion and coherence, which the corpora cannot easily highlight.

Final revised version received July 2008

References

Chandler, J. 2004. 'A response to Truscott'. *Journal of Second Language Writing* 13/4: 345–8.
Cobb T. 1997. 'Is there any measurable learning from hands-on concordancing?'. *System* 25/3: 301–15.
Ferris, D. 1999. 'The case for grammar correction in L2 writing classes: a response to Truscott (1996)'. *Journal of Second Language Writing* 8/1: 1–11.
Ferris, D. 2004. 'The ''Grammar Correction'' debate in L2 writing: where are we, and where do we go from here? (and what do we do in the meantime?)'. *Journal of Second Language Writing* 13/1: 49–62.
Gaskell, D. and **T. Cobb.** 2004. 'Can learners use concordance feedback for writing errors?'. *System* 32/3: 301–19.
Hedge, T. 1988. *Writing*. Oxford: Oxford University Press.
Hyland, K. and **F. Hyland.** 2006. 'Feedback on second language students' writing'. *Language Teaching* 39: 83–101.
Johns, T. 1986. 'Micro-concord: a language learner's research tool'. *System* 14/2: 151–62.
Todd, R. W. 2001. 'Induction from self-selected concordances and self-correction'. *System* 29/1: 91–102.
Truscott, J. 1996. 'The case against grammar correction in L2 writing classes'. *Language Learning* 46/2: 327–69.
Truscott, J. 1999. 'The case for ''the case against grammar correction in L2 writing classes'': a response to Ferris'. *Journal of Second Language Writing* 8/2: 111–22.
Truscott, J. 2004. 'Evidence and conjecture on the effects of correction: a response to Chandler'. *Journal of Second Language Writing* 13/4: 337–43.
Williams, M. and **R. Burden.** 1997. *Psychology for Language Teachers: A Social Constructivist Approach*. Cambridge: Cambridge University Press.

The author

Alex Gilmore is currently working as a visiting lecturer at Kyoto University in Japan, where he teaches English for Academic Purposes. He has a Cambridge Diploma in Teaching English as a Foreign Language to Adults, as well as an MA and a PhD from Nottingham University. He is also a teacher trainer on the Cambridge Certificate in English Language Teaching to Adults (CELTA) course. His research interests include, among other things, technologies in ELT, materials design, and classroom-based research.
Email: alexgilmore@mac.com

Appendix 1
Training materials used in the investigation

Using online corpora to improve your writing

A What are online corpora?

Corpora are basically large collections of texts (books, newspapers, journals, transcribed speech, etc.) stored electronically and accessible using search software. If you know how to use these resources, they can help you to identify problems in your writing and to express yourself in the same way as English native speakers do. Two of the most useful online corpora are:

i The British National Corpus (BNC): http://www.natcorp.ox.ac.uk/
ii The COBUILD Corpus and Collocations Sampler: http://www.collins.co.uk/Corpus/CorpusSearch.aspx

B How can online corpora be used to improve drafts?
In order to understand how to use online corpora to improve your writing, let us look at some genuine student writing errors:

i 'Since then, he started to go . . . '
ii '. . . but we cannot make it worth.'
iii 'My confidence changed . . . '
iv 'and she died for a car accident'

With a partner, try to decide what the problems are with each of these phrases.

C Answers
In (i), a search using the keywords *since + then* in COBUILD gives the following example sentences (known as concordance lines) from the corpus:

Episcopal church services has increased by 23 **since** then. [p] What difference would it make if

Constitutional Committee in 1991 and have **since** then served on panels dealing with a wide

as a grade A8 administrator in May 1993 and **since** then have worked in Directorate-General 1A

oldest members. They were elected in 1920 and **since** then their relationship has been a close

18 months ago after a bloody military coup. **Since** then thousands of Haitian refugees have been

From these examples, we can see that *since then* is typically used with the present perfect tense in native speakers' texts (have + past participle) and we can conclude that (i) should be rewritten as 'Since then, he has started to go . . . '

In (ii), a search using the keywords *make it worth* in the BNC (notice that the BNC does not require a '+' sign between words) gives the following concordance lines:

A6A 394 This alone can make it worth a reporter's time to come along.

ADA 1557 I'll make it worth your while, honey.

ADB 891 No-one would have planned a system like it, but, the argument goes, its powers to examine and revise legislation are scarcely great enough to make it worth the bother of finding something to replace it.

ASH 386 Benefits such as good facilities can make it worth travelling farther; it's amazing how a well-designed yard can cut down on working time.

AT4 1263 He would, if you make it worth his while.

From these examples, we can see that *make it worth* does not come at the end of a sentence. It often occurs in the pattern 'make it worth + possessive pronoun + while', for example, 'make it worth your while'. Another pattern we can see is 'make it worth + verb-ing', for example, 'make it worth travelling'. We can therefore conclude that the phrase 'make it worth our while' might be more appropriate.

In (iii), a search for *confidence + changed* in COBUILD produces the following:

Lookup Error: No matches.

This means that there are no examples of this pattern in the whole COBUILD corpus, which is made up of 56 million words! We can therefore conclude that this is not a natural expression. So how would a native speaker write this idea? A search using the keyword *confidence* produces the following example concordance lines:

I was tearful all the time and I lost my **confidence**. I couldn't sleep and I suffered from

lost and bewildered. How can I regain my **confidence**? [p] [f] A [f] No one can go through the

helped me out a lot. This place has built my **confidence** up for just getting out there and going for

I should be reaching and I have got my **confidence** back. [p] Aberdeen as a team have been

and that was a considerable boost to my **confidence**. [p] I owe him a favour but he'll have to

From these examples, we can see that there are many ways to describe how confidence changes in English, depending on whether it increases or decreases.

In (iv), a search using the keywords *car accident* in the BNC produces the following example concordance lines:

CN3 815 To take one example, a man was killed in a car accident.

CEK 1948 Her husband dies in a car accident alongside another woman and driven by grief and jealousy, she investigates his secret life and becomes

G15 2972 For many years, Marek wrote, he had believed his mother when she said his father had been killed in a car accident.

HH0 2263 In fact I cause a car accident by obstructing someone's driveway.

HWL 7 I covered the mouthpiece and said: 'Salome's been involved in a car accident.

We can see from these examples that 'was killed in . . .' (passive construction) or 'somebody died in . . .' are more appropriate structures to use.

D Identifying errors in your own writing
Now look at your own essay and, using an online corpus, try to identify the errors in your own writing.

Appendix 2
An extract from the rating criteria used for estimating 'naturalness'

Student writing samples

Name of rater:

Date:

The following extracts are taken from university students' academic essays. Please indicate which version you consider to be more natural by placing a cross next to it, for example:

I started to associate with my girlfriend a year ago. __
I started going out with my girlfriend a year ago. X
No difference __
If you do not consider one to be more natural than the other, please put a cross next to 'No difference X'.
KU (052015)

1 In 2003, Best basketball player retired. __
In 2003, the best basketball player retired. __
No difference __

2 He was a legendary basketball player. __
He was legend. __
No difference __

3 When he was 1 year in high school student, he couldn't join the basketball club. __
When he was a freshman in high school student, he couldn't join the basketball club. __
No difference __

4 But one year later, he became the body of club. __
But one year later, he became a captain. __
No difference __