Can a graded reader corpus provide 'authentic' input?

Rachel Allan

In addition to their intended purpose, graded reader texts can be made into a corpus appropriate for use with lower-level learners. Here I consider using such a corpus for data-driven learning (DDL), to make this approach more accessible to intermediate level students. However, how far does grading the corpus in this way compromise the authenticity of the language learners are exposed to? The simplified nature of such corpora may limit learners' exposure to lexical chunks, which are fundamental to the acquisition of natural and fluent language. This paper compares lexical chunks in graded corpora and the British National Corpus, examining frequency, type, and composition, to evaluate the 'authenticity' of graded input. Despite some differences, it is argued that the scale and type of lexical chunks are sufficient to provide input that reflects authentic language, suggesting that graded readers may offer an acceptable balance of accessibility and authenticity.

Introduction

Graded readers are a useful way of motivating learners to read extensively, through the accessibility they provide by limiting the number of headwords. This accessibility also makes them a valuable resource when made into a corpus, a database of texts, for learners not yet able to manipulate an authentic corpus. A graded corpus gives such learners the opportunity to analyse and explore language in new ways, such as through data-driven learning, as described below.

Data-driven learning

Data-driven learning (DDL) refers to the use of a corpus of texts with concordancing software, to find answers to linguistic questions. The learner inputs the target word or words into the software and all examples from the corpus are returned, usually in a keyword in context (KWIC) format, with the target word in the middle of the line (see Figure 1). These lines can then be sorted in a variety of ways that may help to reveal patterns in meaning and usage. The learner is viewed as a 'research worker whose learning needs to be driven by access to linguistic data', whose role is to 'identify–classify–generalize' that data (Johns 1991: 4). Learners, then, interact with the concordance and find answers to their questions about the target word by looking for patterns in it, categorizing them and deriving their own hypothesis, rather than relying on a teacher's intuition or research.

ELT Journal Volume 63/1 January 2009; doi:10.1093/elt/ccn011

Advance Access publication March 12, 2008

At a theoretical level, DDL is appealing in many ways. Although it is not a communicative approach, it is in harmony with many of the other current themes in language teaching pedagogy, being learner-centred, using authentic language input, and encouraging learners to 'notice' linguistic features. It can be viewed as a task-based approach with language as topic (Sheehan 2005), and one in which 'learning by doing interacts thoroughly with learning by reflection' (Little 1996: 210), a feature encouraging autonomous behaviour. Although not communicative, it can certainly be collaborative, encouraging peer learning. Not only does it use authentic input, but it is a 'pedagogical application of a research method' (Mishan 2004: 222)—an authentic task in its own right. For lexical learning, it is particularly helpful in that it gives learners multiple exposures to words in context, offering potential for deepening word knowledge through the information provided about collocations, contextual behaviour, and register. It would appear to be a valuable explicit 'focus on form' technique; not one to be used to excess, as Sheehan (op. cit.) notes, but a useful tool all the same.

Despite this, DDL seems not to have attracted a wide following, at least in general English language teaching contexts. There have been few empirical studies on its effectiveness (e.g. Allan 2006; Cobb 1997). Research has focused mainly on its use in ESP contexts, and the impression from the literature is that its use is more widespread, and perhaps more profitable here. Contributing to the success of DDL in this context may be the smaller, more specialized corpora used, which increase their accessibility and relevance to the learners. Working with a corpus containing the quantity and range of authentic texts required to reflect general language use, such as The British National Corpus (BNC 2001), has the opposite effect. While more proficient learners may be able to cope with this, those at an intermediate level, situated at B1 or B2 of the Common European Framework of Reference (Council of Europe 2001), for example, are unlikely to be able to deal with the peripheral linguistic content of a search from the BNC or other large corpus. The B1 learner, for example, is described as being able to deal with 'high frequency everyday or job-related language' (Council of Europe 2001: 26), while the B2 learner can deal with more complex and lower frequency language provided 'the topic is reasonably familiar' (op. cit.). As the sample concordance lines from the BNC show in Figure 1, there is quite a high proportion of language on topics which are quite unfamiliar and far from everyday.

Concordance

1 streets. Despite French government threats to prevent the **deal** *going* through, and criticisms by Belgium and Italy, few
2 art dealer, Andrew Ciechanowiecki LONDON. A great **deal** *has* been said and written over the past two years
3 in dollars, receiving the face value in local currency. The **deal** *means* the French bank will get some of the money it
4 of stellar energy is nuclear. Normal stars contain a great **deal** *of* hydrogen, the lightest and most abundant substance
5 exploiters of the indigenous peoples; there was a great **deal** *of* theological questioning and guilt and struggle around

FIGURE 1
Sample lines from BNC concordance of 'deal'[1]

The five random lines taken from a search on 'deal' from the BNC in Figure 1 illustrate the problem. Items like 'criticism', 'face value', 'guilt', and

'struggle', may well not be known by a B1 or even a B2 learner. In addition, there is some quite highly specialized language—'stellar', 'nuclear', 'abundant', 'exploiters', 'indigenous', and 'theological'— which neither the B1 or B2 learner is likely to know. These examples are not exceptional. In any given concordance line of 15–20 words, a learner with a vocabulary of 2000 words is likely to meet two or three lower frequency, potentially important content words. A further problem is presented by the length of sentences in most authentic text—note that there are no complete sentences in the lines in Figure 1; this makes the cut-off nature of the concordance lines more difficult to deal with.

Grading the corpus

It has been suggested that grading the corpus, using 'limited and manageable' text sources (Gavioli and Aston 2001: 244), might be a way of overcoming this problem. As I have indicated, one way of doing this is to make up a corpus of graded reader texts which contain a limited number of headwords. This enables us to adjust the ratio of known to unknown words for learners with a more limited vocabulary, making them more able to work with the data, without the need for filtering through the teacher. If we look at some sample lines from corpora made up of graded readers, their increased accessibility is clear—see Figure 2. The corpora are described in detail below; both are made up of Penguin graded reader texts, with the B1 corpus using level 4 texts and the B2 corpus using level 5 texts.

Concordance

1 would still allow them to broadcast the show, but the **deal** *meant* that they were losing control of their most
2 with the big entertainment company Time Warner. The **deal** *meant* that she was now the boss of her own
3 proved her skills as a businesswoman when she signed a **deal** *with* a value of $60 million with the big entertainment
4 on Chaos was clear, too: this book will help you to **deal** *with* the problem. Events in the years after 1987
5 world and nobody can really explain it. So how can we **deal** *with* these changes? Handy tells us to forget about

Concordance

1 he was sincere. We shared this interest. I knew a great **deal** *about* Italian wines myself, and bought large amounts
2 'He is rather unusual, perhaps. He has travelled a great **deal,** *and* seen much of the world. I suppose he is clever,
3 reported anywhere." Soon the two men shook hands. The **deal** *was* done. Wilson rgrow found Joe Roy Spicer in the
4 'of the highest importance' could he possibly have to **deal** *with?* I feared that the continued weight of misfortune
5 he answered. It was a difficult situation. The only way to **deal** *with* it was to use the direct method of shock. 'Where

FIGURE 2
Concordance lines on 'deal' from B1 (top) and B2 corpora[2]

There is no highly specialized or very infrequent vocabulary present. At times the context gives very clear clues as to the meaning of the word, for instance, that when a deal is done, people shake hands (B2 line 3). Sentences are shorter; the lines include some complete sentences, and those that are cut off can be more easily predicted. Difficulties may still arise, for example in B2 line 4, from issues of style and register, but overall it is clear that the lines are much more manageable for learners.

However, how far can it be assumed that a graded corpus like this reflects authentic language? Do the language patterns learners need to know still emerge? As DDL automatically draws attention to common collocations,

a useful way of examining this is to compare the occurrence of lexical chunks in graded and authentic corpora. 'Lexical chunk' is here used to refer to a continuous sequence of two or more words that frequently occur together, which 'display pragmatic integrity and meaningfulness regardless of their syntax or lack of semantic wholeness' (O'Keeffe, Carter, and McCarthy 2007: 78). In other words, although a chunk may be a fragment, it has an internal coherence of some kind. Included within this definition are sentence frames, linkers, fixed and semi-fixed expressions, and collocations. Lexical chunks are considered fundamental to achieving native-like fluency (e.g. Pawley and Syder 1983; Sinclair 1991) so even text that is simplified should contain such chunks to provide useful input. This makes it important to find out how far such items are filtered out in the grading process.[3]

The corpora

B1 and B2 corpora were made up of simplified texts from graded readers. The Penguin series of graded readers were chosen because of the variety of genres and topics covered. The majority of the books are works of fiction, both historical and contemporary, but a limited number of works of non-fiction are included. See Figure 3 for a detailed breakdown:

Text type	B1	B2
Contemporary fiction	176,933	445,917
Classic fiction	266,963	566,020
Non-fiction	76,321	53,199
Total words	520,217	1,065,136

FIGURE 3
Composition of graded corpora

The authentic corpus used was the written only portion of the BNC, comprising around 93 million words of a wide variety of text types. There was, therefore, a variation in text types, as the whole written BNC was used, rather than a sub-corpus of fictional texts which would reflect the text types in the graded corpora. This was because my aim was to compare lexical chunks from the graded corpora with general, baseline data, to see whether they are representative of chunks generally occurring in authentic texts. The Keywords function of Wordsmith Tools (Scott 1996) was used to find words whose frequency was unusually high in the graded corpora in comparison with the BNC. This showed in general a greater emphasis on personal pronouns, proper nouns (names of characters), and on conveying action, intent, and emotion using verbs and adjectives, as could be expected with the high proportion of fiction.

How many word clusters recur?

The first thing I considered was how many word clusters recurred within the two graded corpora, to see if this was comparable with an authentic corpus. Recurrent word clusters were identified using the Wordlist facility in Wordsmith Tools (op. cit.) which can generate lists of word strings of any given number of words that occur in the corpus a specified number of times. Of course, not all of these clusters would count as lexical chunks according to the definition given above, but they help paint a general picture of text composition.

Working on the premise that a minimum of ten samples of a word cluster would be required to make it recognizable as a chunk in DDL, all

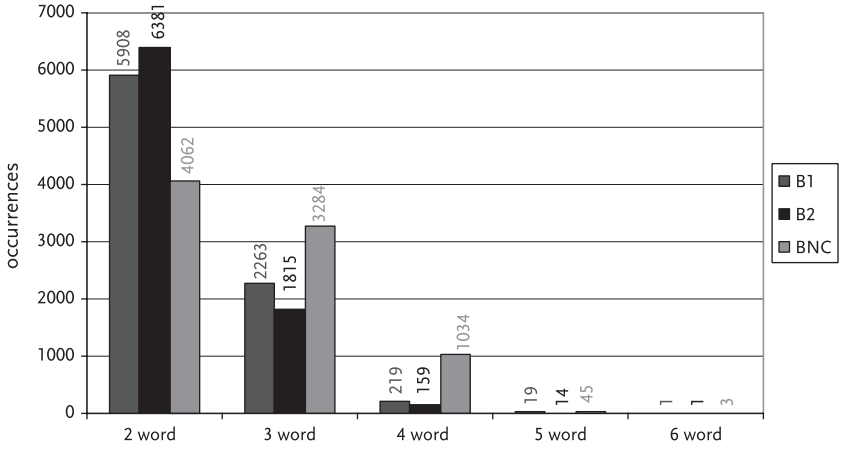

FIGURE 4
Distribution of word strings occurring at least 20 times per million words (normalized data from 93 million-word BNC)

the clusters of two words or more occurring in the B1 corpus at least ten times were counted; in the B2 corpus those occurring a minimum of 20 times were identified, to normalize the data for comparative purposes. Clusters occurring in the BNC were similarly counted and the patterns compared. Figure 4 shows the occurrence of two-, three-, four-, five-, and six-word clusters, using normalized data based on 20 occurrences per million words.

The B1 and B2 corpora show a similar number of occurrences of clusters at each level, and the BNC, when scaled down proportionally, shows a similar pattern overall, as seen in Figure 4. As we would expect, there is quite a dramatic drop in the number of clusters occurring as they become longer. There are, however, differences between the graded corpora and the BNC. There are fewer two-word clusters occurring in the BNC compared to the two graded corpora. This is likely to have arisen from the greater lexical variation in the BNC, which means that two-word clusters which include lower frequency items will not show up with sufficient frequency to be recognized—particularly when data is normalized in proportion to a corpus of only one million. Furthermore the drop-off rate is much more gradual in the BNC, with many more four-, five-, and six-word clusters appearing. These discrepancies may well arise from comparing corpora of significantly different sizes. Nevertheless, it certainly appears that word clusters occur in the graded corpora with sufficient density to be identifiable.

What kind of chunks occur?

My next step was to look at some of those clusters that could be called chunks, according to the definition above. Two- and three-word clusters occurring in the graded corpora mainly consisted of preposition + article, subject + verb, subject + verb + complement, noun phrase + of. These showed relations of time and place, other prepositional relations, interpersonal functions, and linking functions. In other words, they are what we would expect, and correspond to what we might find in an authentic corpus. (See Carter and McCarthy 2006: 829.) As such, through the graded corpora learners will be exposed to plenty of chunks that are representative of the most commonly used authentic language. However, we need to look to the longer word clusters to find more 'useful' chunks—useful to the

learner in that they may be more idiomatic and less easily formed by knowledge of syntax alone.

Chunks from the B2 corpus

Figure 5 shows some examples of four- and five-word chunks occurring in the B2 corpus at a frequency of ten per million, again following the premise that ten samples would create an identifiable chunk in DDL. These represent a random selection of fairly cohesive chunks, and are very loosely categorized into groups.

Clearly, most of these phrases would need to be seen in their broader contexts to make sense of them—and DDL would facilitate this. The range of chunks shown suggests that learners would be getting exposure to structures formed either principally or partially by lexis.

Idiomatic phrases	Collocations	Interpersonal phrases	Sentence builders
face to face with	took a deep breath	seems to me that	there was no sign
red in the face	put the phone down	can I help you	as soon as he had
made up my mind	head in his hands	I don't want you to	as a matter of fact

FIGURE 5
Sample four- and five-word chunks in B2 corpus

Chunks from the B1 corpus

While the B1 corpus illustrates a more limited range, there still seem to be plenty of useful chunks for learners, as Figure 6 shows. They reflect strong collocations in everyday usage, but there is less evidence of idiomatic language here. The interpersonal phrases are of a more functional nature than those in B2, above. There are also many discourse markers, particularly connected with time, which are essential for organizing narratives. Exposing learners to these and raising consciousness of them might be helpful in assisting B1 learners to make the move from short turns to longer narratives in their own speech.

Collocations	Interpersonal phrases	Narrative markers	Sentence builders
knock on the door	what do you	for the first time	there was no sign of
a piece of paper	mean/think/want	by the end of	one of the most important
eyes filled with tears	what's the matter	a few minutes later	couldn't think of anything
	what are you doing		

FIGURE 6
Sample four- and five-word chunks in B1 corpus

Do 'graded' chunks represent 'authentic' exposure?

Having established that potentially useful chunks were occurring in the graded corpora at frequencies more or less comparable with an authentic corpus, the final question was whether the same chunks occurred with similar frequency levels. In other words, were learners getting exposure to the more frequent chunks through the graded corpora, as well as the more frequent words? To get some insight into this, I looked at the frequencies of chunks around some specific words in the B1 and B2 corpora and compared these with the BNC.

'Mind' chunks

To start with, chunks around the target word 'mind' were found in the three corpora, as shown in Figure 7. The graded corpora show a similar pattern, with 'change your mind' and the functional use of mind, in phrases such as 'if you don't mind/do/would you mind' coming at the top of the list, although the rankings are reversed. The B1 corpus then has a rapid drop off, with only a few instances of 'make up your mind' and 'out of your mind'. The same pattern is reflected, but with more occurrences in the B2 corpus. The BNC displays a different emphasis. Although these chunks do occur near the top of the frequency list, a number of other chunks are present which do not show up in the graded corpora (indicated in bold in Figure 7). The most notable of these is 'bear/be borne in mind'; there are many more instances of this chunk, which does not occur at all in the B1 corpus and only once in the B2 corpus.

B1	B2	BNC
change(d) ~ mind (22)	do/did/would (n't) mind (43)	**bear/borne in mind**
do/did/would (n't) mind (13)	change(d) ~ mind (41)	**in the mind of**
made up my mind (5)	make/made up ~ mind (32)	make/made up ~ mind
out of ~ mind (3)	out of ~ mind (11)	out of ~ mind
	bear in mind (1)	**back of ~ mind**
		have/had in mind
		do/did/would (n't) mind
		change(d) ~ mind

FIGURE 7
Chunks containing 'mind' in the three corpora

This raises the question of how far 'bear' has been filtered out of the graded corpora. A search on 'bear' in the B2 corpus shows that the most frequent use is in the context of 'could not/cannot bear'. This ranks third in frequency in the BNC, following 'bear in mind' and 'brought to bear'. Thus, learners' exposure to a common chunk is somewhat restricted by the corpus in this case, although whether this is due to the grading or composition of the corpus is unclear. The fictional nature of the graded corpus, with its emphasis on language of emotion, may explain the emphasis on 'couldn't bear' here.

Returning to 'mind', another chunk high on the frequency list of the BNC is 'out of * mind'—the asterisk is a 'wild card' representing another word—which again occurs with a very low frequency in the B1 and B2 corpora. 'Out of' is used in many contexts in the graded corpora, in prepositional phrases for 'bed', 'town', 'business', 'prison', 'jail', and more idiomatically with 'control', 'breath', and 'sight', for example. This time, a search on 'out of' in the BNC shows that this corresponds to frequent usage, with all of these uses more common than 'out of * mind' (though the length of the cluster is also a contributing factor here).

'State' chunks

The different composition of the corpora is highlighted in a search on clusters around 'state', in Figure 8. The B2 corpus shows that the use of 'state' in the sense of 'condition' in descriptive phrases is prevalent (there are insufficient occurrences in the B1 corpus to generalize), whereas in the BNC

'state' is principally used in the sense of 'government'—reflecting the use of current affairs materials in the BNC.

FIGURE 8
Chunks containing 'state' in B2 and BNC

B2	BNC
a state of (18)	secretary of state
in a state (17) BNC rank 8	of the state
the state of (17)	the secretary of
state of mind (8) BNC rank 33	the state of
a state bank (7) BNC rank 577	a state of

'Deal' chunks

Searches on other words show the graded corpora to follow very similar patterns to the BNC. Taking, for example, the word 'deal', we can see that exactly the same five chunks appear at top frequencies, although in a slightly different order, as shown in Figure 9. Again there are too few occurrences in B1 to draw any conclusions.

FIGURE 9
Chunks containing 'deal' in B2 and BNC

B2	BNC
a great deal (53)	a great deal
great deal of (30)	to deal with
to deal with (22)	great deal of
a good deal (20)	deal with the
deal with the (9)	a good deal

These examples suggest that although some chunks in common usage may be screened out in the grading process and due to text genre, occurrences of chunks in the B2 graded corpus may reflect authentic language use quite closely. The size and grading of the B1 corpus, however, does affect access to commonly used chunks, as seen in the case of 'deal' and 'state'. However, for the teacher looking for a way into DDL with lower-level learners it seems that graded corpora may offer a reasonable balance of accessibility and authenticity in the data it provides.

Conclusion

The picture presented here is, of course, only a very small tip of the iceberg. Clearly, a much more detailed analysis would be needed to define the true limitations of the graded corpora. However, even a small snapshot like this provides an argument for using graded corpora in DDL. The advantages of using a smaller, more accessible corpus do not seem to be outweighed by its limitations. The data may not be authentic, but it does contain authentic features. Learners are less likely to be overwhelmed by the data, and more likely to be able to understand it and draw conclusions from it. Learning can be staged in a way that it cannot be through an authentic corpus, with learners introduced to a limited number of senses and uses of a particular word or phrase, just as they would be through a learner dictionary or textbook; moving through levels of learner corpora would allow them to deepen knowledge gradually.

Of course, the limitations of the corpus do, inevitably, restrict the learners' exposure to some very frequent chunks, as shown. To a large extent, this appears to be bound up with text type in the graded corpora, and the

predominance of fiction. It is difficult, perhaps impossible, to find graded texts which reflect the range included within an authentic corpus. However, if learners and teachers are persuaded of the value of using graded corpora like these, publishers may respond and make them commercially available, graded to different levels and screened for lexical chunks, to ensure that the common ones appropriate to the level do indeed occur. Dictionaries and grammars have become increasingly computer-based in recent years, CD-roms now coming as standard with them. As learning becomes more autonomous and technology-driven, it is not difficult to imagine learners wanting to add a user-friendly corpus and concordance package to their set of learning resources.

Finally, there are a couple of more general observations to be made. First, this study uses data-driven learning as a means of identifying chunks. This indicates that DDL is a research tool that is just as valuable for teachers as learners in identifying linguistic features. Second, there can be a great deal of resistance, particularly initially, to working with concordance lines because of their decontextualized nature. Our natural instinct is to read a meaningful, complete text and this is clearly the main function of graded readers. This investigation has indicated that useful authentic chunks are present, and obviously learners will be exposed to them by simply reading the texts. However, if we want them to explicitly notice specific chunks, DDL provides an effective means of doing this.

Revised version received July 2007

Notes

1 Data cited herein have been extracted from the British National Corpus, distributed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved.

2 These concordance lines are drawn from a variety of Penguin graded reader texts at levels 4 (B1) and 5 (B2), donated for research purposes by Pearson Education Ltd., and are reproduced here with permission, all rights reserved. (Titles cited: Evans, D. *Management Gurus* and *Women in Business*; Poe, E. A. *Tales of Mystery and Imagination*; Bronte, C. *Jane Eyre*; Grisham, J. *The Brethren*; Thornley, G.C. (ed.). *Outstanding Short Stories*.)

3 According to the series editors of the Penguin graded readers, authors and adapters of the readers 'are provided with extensive briefing notes, which include sections on structure, content, and style', as well as 'structural guidelines which specify the grammar that may and may not be used at each level and British and American wordlists for each level' (personal communication). There is no explicit indication of lexical chunks or collocations that should or should not be used.

References

Allan, R. 2006. *Data-Driven Learning and Vocabulary: Investigating the Use of Concordances with Advanced Learners of English*. CLCS Occasional Paper 66. Dublin: Trinity College.

BNC. British National Corpus, Version 2 (BNC World). 2001. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk/.

Carter, R. and **M. McCarthy.** 2006. *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Cobb, T. 1997. 'Is there any measurable learning from hands-on concordancing?' *System* 25/3: 301–15.

Council of Europe, 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Gavioli, L. and **G. Aston.** 2001. 'Enriching reality: language corpora in language pedagogy'. *ELT Journal* 55/3: 238–46.

Johns, T. 1991. 'Should you be persuaded: two examples of data-driven learning' in T. F. Johns and P. King (eds.). *Classroom Concordancing*. Birmingham: ELR.

Little, D. 1996. 'Freedom to learn and compulsion to interact: Promoting learner autonomy through the use of information systems and information technologies' in R. Pemberton, E. S. L. Li, W. W. F. Or, and H. D. Pierson (eds.). *Taking Control: Autonomy in Language Learning*. Hong Kong: Hong Kong University Press.
Mishan, F. 2004. 'Authenticating corpora for language learning: A problem and its resolution'. *ELT Journal* 58/3: 219–27.
O'Keeffe, A., R. Carter, and **M. McCarthy.** 2007. *From Corpus to Classroom*. Cambridge: Cambridge University Press.
Pawley, A. and **F. Syder.** 1983. 'Two puzzles for linguistic theory: Nativelike selection and nativelike fluency' in J. Richards and R. Schmidt (eds.). *Language and Communication*. London and New York: Longman.
Scott, M. 1996. *Wordsmith Tools*. Oxford: Oxford University Press.
Sheehan, R. 2005. 'Language as topic: learner-teacher investigation of concordances' in C. Edwards and J. Willis (eds.). *Teachers Exploring Tasks*. Basingstoke: Palgrave Macmillan.
Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

The author

Rachel Allan has worked as an English language teacher in Asia and Europe, and is currently working in teacher education and research into vocabulary acquisition at the Applied Language Centre, University College Dublin, Ireland. The research reported was carried out as part of the Daedalus Vocabulary Acquisition Project.
Email: Rachel_Allan@alc.ucd.ie