Corpus, concordance, classification: young learners in the L1 classroom

Author for contact:

Alison Sealey
English Department
University of Birmingham
Edgbaston
BIRMINGHAM
B15 2TT

a.j.sealey@bham.ac.uk

+44 121 414 5667

Second author:

Paul Thompson
Department of Applied Linguistics
School of Languages and European Studies
University of Reading
Reading RG6 6AA
+44 118 3786472

Short title: ‘Corpus, concordance, classification’

Corpus, concordance, classification: young learners in the L1 classroom

Abstract

This article reports on an ESRC-funded project which investigated the use of corpus-based activities in a primary-school context, with children aged 8-10 years. The study aimed to explore the contributions that could be made by a corpus - comprising language written for a child audience – and a modified version of the associated software, in helping these young children (all L1 English speakers) to learn about language. Activities were devised which complied with educational policies in England, so the interactions recorded often involved classification of linguistic items. The article presents a qualitative analysis of these interactions, identifying aspects of the approach which prompted metalinguistic discourse. It suggests that, in contrast with textbooks and other reference resources, this approach may provide a flexible route into metalinguistic understanding, which maintains links with authentic discourse.

Keywords

Metalinguistic awareness; corpus linguistics; concordancing; young learners

Introduction

This article reports on an ESRC-funded project[1] which investigated the use of corpus-based activities in a primary-school context, and the contributions that such resources can make in helping children to learn about language. Current education policy in England, including the National Curriculum and National Literacy Strategy (NLS) (DfEE and QCA 1999; DfES 2004), requires the explicit teaching of grammar, in which traditional approaches often involve textbook exercises, using invented sentences. By contrast, the study reported here explores the potential of using a purpose-built corpus for English primary school children's learning about their first language. Corpus linguistics seeks to identify the patterns which emerge when the sentences people have actually written (and the utterances they have actually spoken) are collected together and submitted to computer-assisted analysis. This means that intuition about the way language works – on the part of both teachers and learners – can be replaced by empirical data, and applications of corpora in language teaching have developed significantly in recent years. However, up to now these have been mainly at the tertiary level, and mainly with learners of additional languages (see, for example, Hunston 1995; 2002; McEnery *et al.* 1997; Sinclair 2004a; Wichmann et al. 1997). The corpus used in this project was a sub-set of the British National Corpus (BNC), which contains 100 million words of running text, the majority from a wide range of types of writing in English. The texts categorised as having been written for a child audience were identified and extracted, giving us a corpus of approximately 800,000 words, from 40 texts (including stories, history books, a *Brownie* annual and so on). To investigate language patterns, we used the concordancing program, *WordSmith Tools* (Scott 1999). The research investigated the children's reactions to this kind of 'data-driven learning' (Johns 1990), as well as to the software interface, which, in consultation with them, was adapted to facilitate a user-friendly on-screen presentation of concordance lines, queries and collocations.

1. The classroom-based investigations

For the first phase of fieldwork, access was negotiated with two primary schools in a southern county in England to groups of six children in each (one from a Year 4 class of 8-9 year-olds, and the other from two Year 5 classes of 9-10 year olds), whose levels of literacy made them suitable for participation in the study. The ethical procedures agreed within the University of Reading were adhered to, and the conduct of the study was informed by the British Association for Applied Linguistics Recommendations on Good Practice in Applied Linguistics (BAAL 1994). Along with their parents, the children selected gave informed consent to their involvement in the study, as did the class teachers.

Detailed recordings were made of sessions during which the two researchers withdrew participating children to an area suitable for small group work and recording. These comprised six 40 minute sessions with each group in the first phase of fieldwork, and three 50 minute sessions in the second phase, when children from the same two schools participated. In the second phase, both groups were from Year 5 classes, and six of the children were new to the project, while two had taken part in the first phase and were therefore able to act as ‘experts’ in the process of corpus-based language investigations.

Recordings were made using two mini-disc recorders and two video cameras. The former were used to ensure that, as far as possible, all relevant talk was recorded, and the latter so as to facilitate the allocation of turns (children’s voices often sounding very similar) and the recovery of any contextual or paralinguistic information during the process of transcription. The recorded data thus consist of the talk generated by activities involving the corpus, concordancing and the interface, including prompted reflection immediately after completing these, and summative interviews at the end of Phase 1. Throughout both phases of the fieldwork, the children were prompted to reflect on and evaluate their experience of using concordance lines (particularly as paper printouts) and the corpus (particularly on computer). The concordancing program *WordSmith Tools* was being revised at the time, and its author, Mike Scott, took account of feedback from the children in this study, incorporating into the new edition a facility to highlight patterns in the output by means of colour. (See Section 2.4 below.)

The questions guiding the research included: ‘How do primary school pupils respond to corpus-based teaching and learning activities?’ and ‘What kinds of metalinguistic knowledge, understanding or misconceptions are the children prompted to articulate by the presentation of texts in a corpus format (such as concordance lines)?’ As will be apparent, this was conceived as an exploratory study, with no attempt made to ‘test’ the children’s metalinguistic knowledge, or to evaluate this approach to teaching in comparison with others. This left us able to explore a wide range of issues as they were raised by the children, rather than focusing on predetermined categories.

Each session was planned with reference to at least one teaching objective from the NLS, and the programme of Phase 1 sessions was designed to progress from familiarising the children with the concept of a corpus to increasingly greater independence on their part in investigating patterns of grammar and vocabulary. Although the children had access in Phase 1 to the corpus and concordancing program directly, via the laptop computer, it was in the Phase 2 sessions that the majority of activity consisted of on-screen investigation of the properties of the corpus. Activities in Phase 1 made extensive use of paper copies of corpus-derived output, with the concordance lines edited by the researchers so as to highlight specific patterns. Lines of investigation suggested by the children themselves were incorporated into later sessions as the project progressed.

The classroom discussions were transcribed, with links established to the sound files, so that these could be readily consulted during the process of analysis as necessary. The transcription conventions used were minimal, as the primary interest is in the content of the utterances (see Appendix 1 for transcription conventions). The interactions were then coded according to an analytical system based on a set of categories derived from the data with reference to the research questions. These included the kinds of topic discussed (e.g. ‘word classes’, ‘colour coding’) and, where relevant, kinds of contribution (e.g. ‘assertion’, ‘justification’, ‘curiosity’), although frequently it is necessary to analyse fairly extensive portions of dialogue to appreciate how knowledge or understanding is being articulated.

2. Identifying patterns in language

In order to be capable of metalinguistic understanding, learners need to begin to appreciate that there are both links and distinctions between message and medium. Extensive research has been done on how this process occurs, what indicates its development, differences between bi- and monolinguals, the role of explicit instruction, especially in literacy, and so on, and debate continues about all of these issues (see, for example: Benelli et al. 2006; Edwards and Kirkpatrick 1999; Francis 2002; Galambos and Goldin-Meadow 1990; Gombert 1992; Karmiloff-Smith et al. 1996; Tunmer et al. 1984). However, whatever approach is involved, a component of metalinguistic awareness is the recognition that language has distinctive properties, patterns which can be discerned as those properties are identified and grouped to form linguistic categories. Corpus research is helping to clarify the operation of such categories, often challenging conventional systems and identifying new ones, and, in particular, drawing attention to the links between systems which have traditionally been conceptualised as relatively separate from each other – especially the links between lexis and grammar (e.g.: Hunston and Francis 2000; Sinclair 1991; Sinclair 2004b; Stubbs 2001). The ability to classify linguistic items with reference to the properties that they have in common may thus be regarded as a fundamental component of metalinguistic awareness. In the following analysis of our classroom data, we demonstrate various ways in which such classifications emerged as the children interacted with and discussed the corpus and concordance lines generated from it. The research also generated data on a range of other metalinguistic phenomena, but limitations of space preclude presenting those here.

2.1 The identification of word classes: distinguishing nouns from verbs

According to the NLS for primary schools, traditional parts of speech or word classes are specified in teaching objectives such as ‘to … identify, … classify, … understand and use the term …’. While these teaching objectives were used to classify word classes in several teaching sessions, classification also arose more indirectly as exemplified in Extract 1 below. Here, children were completing a worksheet devised both to familiarise them with the appearance of text in concordance format, and to contribute to the NLS objective ‘to spell words with the common endings: *-ight,* etc’. Figure 1 illustrates the worksheet used:

Figure 1 about here

There were several accompanying tasks, one of which was to answer the question ‘What’s the difference between *light* in Line 10 and *light* in Line 11?’ (the first two lines of the abbreviated version shown in figure 1). The children discussed this in sub-groups of three, some drawing fairly readily on metalinguistic terminology and concepts, acquired presumably in their regular English lessons – although the links between the two are clearly not yet secure. Before the following extract, one boy has read the two lines aloud, and the question is then discussed.

~~Data~~ Extract 1: School A, Session 1

Note: Speaker identities are indicated by sex (B = Boy, G = Girl) and school (A or B) and a number to distinguish them within the group. Thus ‘BA4’ is Boy 4 at School A etc. The researchers are R1 and R2. Turns are numbered sequentially within each recorded session.

BA1 [311]: that one
that one would be used as a verb
and this
where is it
and this one would be a

BA3 [312]: er that one’s
that one’s used as an adjective
no that one’s used there as a verb

BA1 [313]: that’s a verb

BA3 [314]: that’s a verb
a doing thing
and that one

BA1 [315]: and that one

BA3 [316]: no
what’s a doing thing called?

BA1 [317]: a doing thing is a
a noun

BA3 [318]: okay so that one’s a noun and that will be a verb
noun
verb
noun verb I think

BA1 [319]: ok
so one is a noun and one is a verb

BA3 [320]: actually
actually that would be a adjective
it’s describing a light

BA1 [321]: *a beautiful golden light*

BA3 [322]: actually *golden* will be a describing thing so that
that
that the er

BA1 [323]: that one’s
look that one’s the
that one’s the verb and that one’s the noun

BA3 [324]: what’s a noun?

is noun the doing thing?

This extract exemplifies two phenomena which we encountered several times. Firstly, many of the children in our study were familiar with labels, such as ‘noun’, ‘verb’ and ‘adjective’, and had some idea of what distinguishes each from the others, but they were not so sure about which kind of word should be accorded which label. Secondly, their resources for determining ascription to a class were limited to ‘notional’ definitions such as ‘doing word’ and ‘naming word’. The presentation of some concordance lines in this context does not, at first sight, differ greatly from the kind of approach which many teachers would routinely use, of posing the question, ‘What’s the difference between the uses of *light* in these two sentences?’, and then supplying some invented examples from intuition. However, two apparently minor differences lie at the heart of our approach, and we shall claim below, from the evidence of our research, that they are potentially very wide reaching.

2.2 The corpus as a source of authentic language

The issue of authenticity has been much discussed, both in the language teaching literature (and, as the ‘performance’ side of the competence/performance distinction, in linguistics itself) (e.g.: Carter 1998; Cook 1998~~; Cook,~~ 2001; Stubbs 2002; Widdowson 2000). Space does not permit a detailed account of the arguments, but briefly they concern questions of two main kinds. Firstly, should descriptive linguists be seeking to account for an underlying, idealised but rule-governed system, of which any instance of spoken utterance or written sentence is a potentially imperfect example, or should they start with empirical evidence and derive rules from the patterns observed? If the former, intuitive assessment of invented examples may well be adequate; if the latter, empirical evidence assumes much greater significance. Secondly, whatever one’s position on this more academic question, should language teachers simplify and idealise language-in-use so as to make examples more accessible and ‘pointed’ for learners, or should they present them with language such as they will encounter beyond the classroom?

Although the issues are slightly different when the focus is not the acquisition of a foreign language but the language awareness of young learners, the question is similar: is it potentially helpful for learners to see linguistic patterns exemplified in contexts such as those where they may actually meet them, rather than abstracted and idealised into sentences which have a distinctly artificial ring to them?

In another of the group discussions about the different patternings of *light*, one of the children, GA2, reads aloud the first concordance line from where it starts: ‘in in in line ten *deep sleep only to be woken by a beautiful golden light*’. From this context she concludes (437), ‘so number ten means er light you turn on.’ She continues this turn by reading aloud the other sentence fragment, ‘*and the daughter had to bring in logs of oak and pine and light the stove*, so it means light up the stove.’ These authentic examples of the same word behaving in different contexts as both a noun and a verb are arguably more suitable ways of presenting this idea than invented examples would be: the concept of word class behaviour is illustrated ‘in vivo,’ and the learner accesses it there rather than in the ‘laboratory conditions’ of the artificial discourse so typical of grammar textbooks.

2.3 The corpus as a repository of evidence about language

Another feature of the approach, which these brief extracts from one of the first teaching sessions exemplify, is the option for the learners of alternating between 'vertical' and 'horizontal' reading of concordance lines. In this activity, the column of '-ight' words down the centre of the page encourages appreciation of a pattern which is not so readily seen with words presented in isolation. At the same time, though, they are not lifted completely out of context, as in a spelling list, and the option remains to switch to horizontal reading, as the children did unprompted when asked to contrast two uses of the same word. We specifically asked the children from Phase 1 about the processability of corpus lines. Some disliked the decontextualisation, with BA1 reflecting, about the concordance line format, 'you don't really understand what the beginning is and you don't even know' [A1/591], whereas BA3 disagreed: 'er well you can really tell, because it like kind of tells you er that it's a noun because it doesn't have the whole sentence, but you can you can still know it's a noun by er listening to some, to the little paragraph' [A1/598]. (It was established that by this he meant the concordance line.)

A corpus also makes visible and publicly available a store of facts about language that can otherwise seem to learners to be held only in their teachers' heads. It is non-judgemental, a source of information about patterns that is quite separate from the source of behavioural authority – the teacher – and yet nor is it a static arbiter like a dictionary or grammar book. Once we had explained what our corpus was, the children began to suggest queries to put to it, in terms which help to suggest how they conceptualised this new resource. For example, BA2's claim, quoted above, that 'you can really tell, because it like kind of tells you' [A1/598], attributes a sort of agency to the corpus output, an interpretation reinforced by his choice of 'listening' for the means by which 'the little paragraph' (i.e. the concordance line) yields up its meaning (' … you can still know it's a noun by er listening to some, to the little paragraph'). A similarly interactive relationship with the computer data is suggested by BA1's proposal, later in the same session, 'I want to ask it for *–ing* or *–ed* words' [A1/629].

In a much later session, when the children were accessing the corpus directly via the computer, rather than paper materials based on it, BB2 responds to GB2's complaint that 'I don't get this,' by pointing out that they can find what they are looking for from the machine: 'There. Look it shows you, look there. Click on that. Right,' [B7/188-189]. Indeed, the children's familiarity with computers in general seemed to stand them in good stead for carrying out queries and getting the program to do what they wanted it to. This is exemplified by the following extract, from a session designed to familiarise the children with the software; they were following written instructions to discover how the words *boy* and *girl* are patterned differently in the corpus. At this point, they are sorting concordance lines for *boy* by the first word to the left of each line:

~~Data~~ Extract 2: School B, Session 7

BB2 [105]: *bespectacled*
BB3 [106]: *bespectacled*
BB2 [107]: press F-six again
GB2 [108]: sorry
BB2 [109]: choose one-L
GB2 [110]: L

BB3 [111]: one at the end
look
Later
BB2 [126]: no it doesn’t go up
oh it does
GB2 [127]: you’ve got to press F-six again choose one-L [xx] to the left
BB2 [128]: on there
BB3 [129]: that’s right
BB2 [130]: so it does the same except
BB2 [324]: there’s about
ooh that’s three-hundred-and-sixty-three
BB3 [325]: oh well
that means boys are more popular
BB2 [326]: yeah
GB2 [327]: hang on the girls are five-hundred-and+
BB2 [328]: no that’s the boys

Again, the use of authentic language from a corpus meant that some of the traditional distinctions between different aspects of ‘knowledge about language’ - and particularly the division of such knowledge into ‘word level,’ ‘sentence level,’ and ‘text level’ - as the National Literacy Strategy is organised, are difficult to sustain. Corpus linguistics, as mentioned earlier, points to fields of influence within language that are apparently emergent from the interaction of grammar and lexis. In the task these children were doing, a quantitative investigation of the relative frequency of two nouns (*boy* and *girl*) was combined with a further opportunity to reinforce the concept of nouns and adjectives, and particularly of their distributional patterns. So by sorting the lines they found alphabetically, by the word immediately preceding the target noun, the role of adjectives in describing nouns (with which they already had some familiarity) was made visible at the same time as they explored the semantics of noun phrases headed by *boy* and *girl* respectively.

In early acquaintance with the approach, children asked about the corpus as a potentially useful resource for them when we, as teacher/researchers, were not there. For example, GB2 reflects after the ‘–ight’ activity that the words listed are common words, to which R1 replied that ‘one of the things you can do with the corpus is find out which words are more common …’ [B1/188]. GB2 and another girl responded independently to this, GB2 musing, ‘yeah I suppose because if there’s more w-, meaning then you’re going to use them more’ [B1/190], which cuts across GB1’s question, ‘is it a website where you can -+’ [B1/189]. A similar question occurs to GA4 in her first session (Phase 2): ‘Can you er can you get the corpus er up on your com-, your computer at home?’. Unfortunately, unless and until a corpus suitable for young learners is developed such interest is likely to be frustrated.

2.4 Colour-coding and classification

There are different views among corpus linguists about the pre-coding of corpus data. Some researchers believe in minimising the categories assumed *a priori*, to allow for the emergence of patterns which might otherwise be missed (Sinclair 1991). The BNC, however, has been tagged for ‘part of speech’ (POS), using the extensive CLAWS tag-set. During the first phase of fieldwork, we introduced the children to concordance lines whose words appeared in different colours, according to the tag with which they had been classified; (the tags themselves were not visible). This colour-coded output stimulated lines of inquiry about the number of different word classes coded, particularly categories with which the children were unfamiliar. They were fairly sure about categories such as noun, verb and adjective, as this brief exchange illustrates:

~~Data~~ Extract 3: School B, Session 5

BB1 [489]: what are the brown ones?
BB1 [493]: the brown ones look like *cut*
GB1 [495]: like *kiss*
GB1 [497]: and *went, got, went*
GB2 [498]: I'm definitely sure they're verbs I think

Function words posed more problems, and, as we have reported elsewhere (Sealey and Thompson 2004), the notion of a broad category of 'dull' – i.e. grammatical or non-lexical – words emerged, when one group looked at colour-coded output for the first time.

For Phase 2, when the children worked on the computer rather than on paper, they used a version of *WordSmith Tools* to which a number of new features had been added, including a colour chooser. This operates on the POS-tagged files in the BNC and makes it possible to view concordance lines in which individual words appear in colours to indicate part of speech classification. There is also a floating, resizeable colour palette which allows the user to turn off any or all POS colours. This facility contributed to the feature of corpus-based teaching noted above, of making an information tool available for learners to manipulate in ways which grammar textbooks and exercises simply cannot provide. The corpus output (although mediated in Phase 1 by the researchers in the form of worksheets and so on) is not a set of instructions or claims, merely evidence, and the idea that the machine itself generates data to be responded to emerges during many of the classroom interactions.

For example, during one Phase 1 session, the children noticed that some words on the printed concordance lines were coloured purple (although the most accurate name for the colour in question was also discussed: 'dark blue' / 'purple'). BA1 suggests that the words for which this colour has been used are proper nouns: 'There's names of people in like that bluey sort of thingymabob' [B4/317]. However, BA2 notices a discrepancy:

~~Data~~ Extract 4a: School A, Session 4

BA2 [374]: er there's
are they er
I think [BA1-Name] was talking about this kind of colour
R1 [375]: yes yes
BA3 [376]: oh
BA2 [377]: but there's a name there that isn't
R1 [378]: right
there's a name there that isn't in that colour er
BA2 [379]: *mama*

It is not quite clear what BA2 is challenging here – BA1's accuracy in linking this colour with proper names, or the computer's accuracy in failing to code 'mama' as a proper noun ('name'). Evidence that it is the program which is being challenged, rather than a fellow pupil's judgement, emerges after a digression (common with this group), when BA2 takes up this theme again, wondering whether 'Mummy' is a name or not. His intonation and the emphasis on 'wasn't' suggests that he has revised his earlier judgement about 'mama' needing to be shown in purple. The referent of 'that' in his first turn below is unclear, but we believe he means the idea that there is a name that has not been coloured purple:

~~Data~~ Extract 4b: School A, Session 4

BA2 [453]: that's actually wrong because er that
it **wasn**'t supposed to be purple because I think that means 'Mummy, are you poorly?' or something

R1 [454]: u-huh I think it does
yes

BA2 [455]: so and *Mummy*'s not a name

R1 [456]: *Mummy*'s not a name
I think sometimes we use it as a name
and sometimes we use it as just an ordinary noun

BA1 [457]: yeah they call people Mummy

R1 [458]: so I think I think [BA2-Name]-,
it's a good point [BA2-Name]
it's
it could be one or the other couldn't it?

BA4 [459]: *Mama*
Mama is a person
a thing
a person is a thing

This final comment, from BA4, is interesting. Falling back, as usual, on semantic criteria to establish what a noun is, he nevertheless observes that, for the purposes of this classification system, 'a person is a thing'.

Another salient feature of this discourse, in the context of the pedagogic role of the corpus, is the speculative interpretations which are articulated as the children try to decide if there is a mistake in the classification, and if so what it is. BA2 is reading from the concordance line when he says 'I think that means *Mummy are you poorly* or something'. No one, not the teacher/researcher, nor the children, knows for sure very much about the texts which have generated the concordance lines, so there is no arbiter to rule on some of these issues. The teacher/researcher can observe 'I think sometimes we use it as a name, and sometimes we use it as just an ordinary noun', but this is an appeal to knowledge outside the corpus.

Another example comes from much later in the project, when the children had become used to manipulating the data and had begun to suggest sometimes unexpected lines of inquiry. The main activity in this session had been designed in response to a discussion about adverbs earlier in Phase 2. The children had been fairly sure that adverbs modify verbs, so the idea of adverbs of degree, used to modify adjectives, had been new to them. In particular, the idea that *very* might be classified as an adverb had caused some problems, which we referred to in this follow-up session.

Extract 5a: School B, Session 10

R1 [21]: how could you find out using the er the computer what sort of a word *very* is?
GB4 [22]: you use the tag and colours
R1 [23]: c-, do you think you can do that?
can you remember how to do it?
GB4 [24]: well the boxes are already up so
we go into er
BB4 [25]: go to ...
GB4 [26]: what word are we typing in then
'cause it's **...**
BB4 [27]: type in the word *very*

In the following sequence, the children experiment with the tags and colours, somewhat baffled at first by the array of colours which appear when the full set of tags is activated. GB4 comments 'It doesn't say like it's an adjective. It's just got all the different colour+' [B10/48]. She then remembers that they can 'uncheck' any of the tags in the palette to see words in that class in monochrome, recalling, ''cause last time we found out it was a determiner [i.e. the word *some*]. Maybe we can click on that and it'll get rid of all the determiners' [B10/63]. They gradually progress to trying to deduce from words whose class they are sure of which colour has been assigned to which class. They are fairly secure about the notion of *angry* being an adjective, and decide to look for words immediately preceding it in the corpus:

Extract 5b: School B, Session 10

GB4 [194]: could you like type in *angry* and then see how many times it comes out?
R1 [195]: yeah
GB4 [196]: see if adverbs ...

Then, to make it easier to see if these are adverbs, the same girl suggests, 'er if we took everything off apart from the adverbs, then we'd find out if there was any adverbs before+'. So, manipulating the program themselves, they set up a query to see both which words precede *angry* (having predicted some from intuition – including *quite*) and, using the colour coding function appropriately, how many of these words are in fact adverbs. As the results are generated, BB4 reports 'got three *really*s, … a *very* and a *recklessly*' [B10/230-1]. As this session continues, the children seem to be grasping the idea of adverbs modifying adjectives, with BB4 correcting GB5 when she notices that *was* precedes *angry* in one of the lines, saying 'but that's verb be so+' [B10/280]. As often happens with young learners, BB4 has assimilated the label 'verb be', used in the list of tags, very quickly, and uses it here to correctly classify *was*, supported by the colour in which it is displayed on screen.

2.5 Semantic classification

The final example to report here, of how working with the corpus encouraged the children to embrace semantics and grammar as they explored linguistic categories, and to move easily between them, comes from later in this same session. As explained above, the phrase ‘very angry’ had been used as a starting point to illustrate the patterning of adverbs of degree modifying adjectives. The children then investigated patterns involving these adverbs, on the one hand, searching the corpus for *really* and *quite*, and then speculated whether *extremely* would precede similar adjectives or different ones. They made some predictions, citing *big*, *small* and *high*, as words likely to follow *extremely*. They carry out their search, and BB4 reports that, ‘every one’s come before … well, the ones that we could see, came before er adjectives’ [B10/487-93]. Interspersed – and sometimes overlapping – with this description of what he can see on screen are examples of this pattern: ‘they’ve got *extremely annoying*’ [GB4], ‘*extremely large*’ [GB5].

At this point, the children began to make use of a facility in *WordSmith Tools* which we had not anticipated using in this project. As well as flexible sorting within the concordance lines, there is a ‘Set’ feature, which allows users to assign a code to specific lines so that they can be grouped by criteria chosen by them. Asked by the researcher, ‘Can you remember how to sort them?’ GB5 replies, ‘You go to the side and write down the letters’ [B10/512] – that is, the user types in a code for concordance lines of a particular type. The researcher asks which adjectives are related to size, and suggests coding these as ‘S’. They then look for other potential semantic categories, and BB4 suggests ‘behaviour.’ Collaboration in this idea is apparent in the next two turns – note the use of ‘‘cause’ at the start of both:

Extract 5c: School B, Session 10

BB4 [529]: c-, ‘cause they’ve got *rude*, *stupid*
GB4 [530]: ‘cause there’s *annoying*
GB5 [531]: *useful*
R1 [532]: okay
BB4 [533]: *unpleasant*

BB4 next identifies *brave*, in response to which GB4 wonders, ‘Is that to do with behaviour?’ [B10/544]. The classification continues:

Extract 5d: School B, Session 10

BB4 [545]: ‘cause it could be brave
GB5 [546]: yeah
R1 [547]: is it behaviour do you think
GB5 [548]: yeah
R1 [549]: go on then
GB4 [550]: er *cold* no

After this, coding becomes a little more problematic, with discussion, and more co-text, required:

Extract 5e: School B, Session 10

GB4 [561]: *hungry*
GB5 [562]: *hungry*
GB4 [563]: yeah ‘cause you’re hungry
is hungry to do with behaviour
I di’n’t think it was
‘cause I thought behaviour was like ‘good’ ‘bad’

BB4 [564]: shall we leave that one?
R1 [565]: yeah leave it out
if you don’t think it belongs
GB4 [566]: *interested*
I think might be
‘cause it’s to do with your behaviour be interested ... like not bothered
BB4 [567]: yeah
GB4 [568]: okay
GB4 [572]: er *old*
it’s not to do with your behaviour
BB4 [573]: *rude* is the next one
GB5 [574]: er
BB4 [575]: *rude*
GB4 [576]: behaviour

The children then re-sorted the concordance lines, causing the ‘size’ and ‘behaviour’ related adjectives to appear together in sets, which left a number of lines still unclassified. Prompted to consider whether the adjectives at the centre of these might have anything in common, the following discussion occurs:

Extract 5f: School B, Session 10
GB4 [631]: I think there might be only one of them
but they could be feelings ‘cause it if you’re feeling hungry
GB5 [632]: feeling strong
BB4 [633]: feeling dangerous
GB5 [634]: yeah
BB4 [635]: it could be condition as well

They decide to tag them with ‘F’ for ‘feelings’, and further discussion is provoked by *cold*, and whether this describes a feeling or not. Here again, with no prompting from the researchers, the children switch from the vertical reading of adjectives appearing at the centre of concordance lines produced by using *extremely* as the search word to horizontal reading of the phrase in context so as to determine the sense in which the item is being used.

Conclusions

It seems inevitable that corpus-based, data-driven learning about language will become increasingly common in an increasingly wide range of learning contexts, and evidence is needed of how learners make use of corpus resources. This small-scale study goes some way towards exploring how one particular group of learners – of primary-school age, working with a syllabus that requires a particular kind of metalinguistic knowledge about their first language – responded to this experience.

In reporting some of the findings from this exploratory project, much has necessarily been omitted. On the one hand, we have focused in this article on one particular aspect of metalinguistic awareness, namely the capacity to classify words by features they share, particularly their grammatical class. This means that we have not discussed here the full range of activities: in various sessions, topics arose that touched on synonymy, register and collocations, for example, while a further development was the construction of small corpora of the children’s own writing, with extensive interest in word frequencies expressed as their statistics were compared with those of the original corpus. On the other hand, the themes and extracts we have presented here have been selected because of what they reveal about these children's insights and perceptions about the metalinguistic topic on which we have chosen to focus, and we have not included the portions of talk which were 'off-task', or about procedural issues and problems – including the 'colour-blindness' of two of the children.

We recognise that we are not in a position to make claims about the general receptivity of children of this age-group to the approach we presented to these 18 individuals, in small group teaching sessions away from the main classroom and its many demands. We also acknowledge, as explained in the introduction, that this kind of study cannot demonstrate a simple cause-and-effect set of outcomes attributable to particular teaching experiences – if indeed any study can; (for discussion of the epistemological and methodological issues raised by this question, see Sealey and Carter 2004).

However, a number of conclusions can be drawn from our analysis of this particular source of data. Firstly, it has been claimed (e.g. McEnery et al 1997; Tsui 2004) that corpus-based teaching facilitates explorations of language patterns which are less evident in other kinds of materials. This is borne out by these children's spontaneous movement between 'horizontal' and 'vertical' readings of concordance lines, and their references to linguistic categories as they respond to the visual evidence presented to them as corpus output. Secondly, the terms in which the corpus is referred to suggest a conception of it (and the associated software) as a resource which the children can deploy to find out various kinds of things about language. Associated with this are (a) the flexibility of this resource, as the children confidently went about setting up queries, to answer both questions set by the researchers and also questions they devised for themselves and (b) the externality of the knowledge retrievable from the corpus, so that intuitions about, for example, the likely combinations of certain groups of words can be overtly tested. Finally, our study suggests that this kind of approach has the potential to mediate between those teaching goals which prioritise responses to texts and authentic discourse, on the one hand, and the need for learners to understand the systemic properties of language, on the other. The empirical evidence which a corpus comprises can demonstrate for learners both patterns and instances, keeping teaching about language itself firmly anchored in genuine discourse.

References

Benelli, B., Belacchi, C., Gini, G. and Lucangeli, D. (2006) ‘To define means to say what you know about things’: the development of definitional skills as metalinguistic acquisition. *Journal of Child Language* 33 (1), 71-97.

British Association for Applied Linguistics (1994) *Recommendations on Good Practice in Applied Linguistics*. http://www.baal.org.uk/about_goodpractice_full.pdf

Carter, R. (1998) Orders of reality: CANCODE, communication, and culture. *ELT Journal* 52 (1), 43 - 56.

Cook, G. (1998) The uses of reality: a reply to Ronald Carter. *ELT Journal* 52 (1), 57 - 63.

Cook, G. (2001) ‘The philosopher pulled the lower jaw of the hen.’ Ludicrous invented sentences in language teaching. *Applied Linguistics* 22 (3), 366 - 387.

Department for Education and Employment (DfEE) and Qualifications and Curriculum Authority (QCA) (1999) *English: the National Curriculum for England*. London: QCA.

Department for Education and Skills (DfES) (2004) *The National Literacy Strategy*. The Standards Site: Department for Education and Skills

Edwards, H. and Kirkpatrick, A.G. (1999) Metalinguistic awareness in children: a developmental progression. *Journal of Psycholinguistic Research* 28 (4), 313 - 29.

Francis, N. (2002) Literacy, second language learning, and the development of metalinguistic awareness: a study of bilingual children’s perceptions of focus on form. *Linguistics and Education* 13 (3), 373-404.

Galambos, S.J. and Goldin-Meadow, S. (1990) The effects of learning two languages on levels of metalinguistic awareness. *Cognition* 34 (1), 1-56.

Gombert, J.E. (1992) *Metalinguistic Development*. Hemel Hempstead: Harvester Wheatsheaf.

Hunston, S. (1995) Grammar in teacher education: the role of a corpus. *Language Awareness* 4 (1), 15 - 31.

Hunston, S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S. and Francis, G. (2000) *Pattern Grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

Johns, T. (1990) http://www.eisu.bham.ac.uk/johnstf/timconc.htm. Accessed October 2006

Karmiloff-Smith, A., Grant, J., Sims, K., Jones, M.C. and Cuckle, P. (1996) Rethinking metalinguistic awareness: representing and accessing knowledge about what counts as a word. *Cognition* 58 (2), 197-219.

McEnery, T., Wilson, A. and Barker, P. (1997) Teaching grammar again after twenty years: corpus-based help for teaching grammar. *ReCALL* 9 (2), 8 - 16.

Scott, M. (1999) *WordSmith Tools*. Software. Oxford: Oxford University Press.

Sealey, A. and Carter, B. (2004) *Applied Linguistics as Social Science*, (Advances in Applied Linguistics series). London: Continuum.

Sealey, A. and Thompson, P. (2004) ‘What do you call the dull words?’ Primary school children using corpus-based approaches to learn about language. *English in Education*. Vol 38 No 1: 80 – 91

Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J.M. (ed.) (2004a) *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins

Sinclair, J.M. (2004b) *Trust the Text: language, corpus and discourse*. London: Routledge.

Stubbs, M. (2001) *Words and Phrases: corpus studies of lexical semantics*. Oxford: Blackwell.

Stubbs, M. (2002) On text and corpus analysis: a reply to Borsley and Ingham. *Lingua* 112 (1), 7 - 11.

Tsui, A. B. M. (2004). What teachers have always wanted to know - and how corpora can help. *How to Use Corpora in Language Teaching*. J. M. Sinclair (ed.). Amsterdam, John Benjamins**:** 39 - 61.

Tunmer, W.E., Pratt, C. and Herriman, M.L. (eds) (1984) *Metalinguistic Awareness in Children*. Berlin: Springer-Verlag.

Wichmann, A., Fligelstone, S., McEnery, T. and Knowles, G. (eds) (1997) *Teaching and Language Corpora*. Harlow: Addison Wesley Longman.

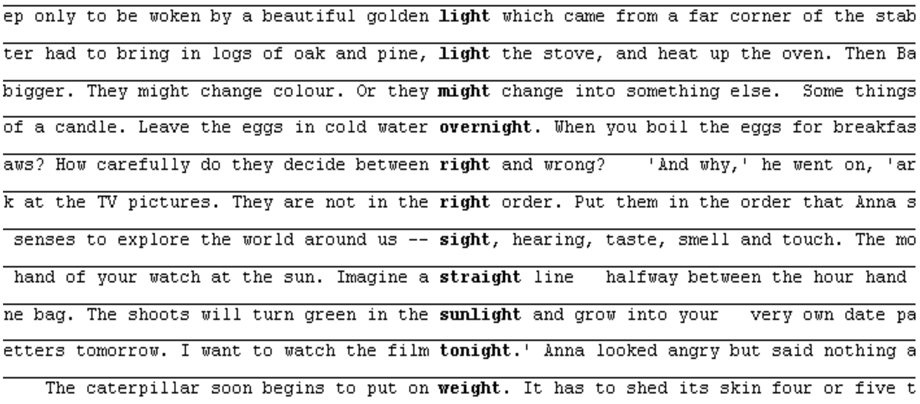
Widdowson, H. (2000) On the limitations of linguistics applied. *Applied Linguistics* 21 (1), 3 - 25.

Appendix 1: Transcription conventions

line breaks	Short pause or silent beat in the rhythm; marks speaker ‘parcellings’ of talk and makes long utterances readable
italics	‘Citations,’ including reading aloud from text on screen or paper
bold	Stressed words or syllables
?	Utterances interpreted as questions
+	Incomplete word or utterance
-	False start

Figures

Figure 1: Printout of concordance lines for words ending in –*ight.*



```
ep only to be woken by a beautiful golden light which came from a far corner of the stab
ter had to bring in logs of oak and pine, light the stove, and heat up the oven. Then Ba
bigger. They might change colour. Or they might change into something else.  Some things
of a candle. Leave the eggs in cold water overnight. When you boil the eggs for breakfas
aws? How carefully do they decide between right and wrong?    'And why,' he went on, 'ar
k at the TV pictures. They are not in the right order. Put them in the order that Anna s
 senses to explore the world around us -- sight, hearing, taste, smell and touch. The mo
 hand of your watch at the sun. Imagine a straight line   halfway between the hour hand 
ne bag. The shoots will turn green in the sunlight and grow into your   very own date pa
etters tomorrow. I want to watch the film tonight.' Anna looked angry but said nothing a
    The caterpillar soon begins to put on weight. It has to shed its skin four or five t
```

[1] The study is part of a research project funded by the Economic and Social Research Council of England (R000223900), entitled ‘An investigation into corpus-based learning about language in the primary school’. We are grateful to the editors of the journal and two anonymous referees for comments on an earlier version of this article.