Through children's eyes?

Corpus evidence of the features of children's literature

Paul Thompson and Alison Sealey
University of Reading / University of Birmingham

This article reports on an analysis of a small corpus of fiction written for children, extracted from the BNC. Quantitative analyses of most frequent words and sequences of words, and of parts-of-speech, were conducted, and compared with their equivalents in two other sub-corpora of the BNC, of adult fiction and of newspaper texts. The main findings point to some characteristics of both the fiction corpora which are very similar, and which contrast markedly with the news texts. However, more nuanced comparison of concordance lines in which the frequent items occur reveal subtle but telling differences between their use in context in adult fiction and in fiction written for children.

Keywords: discourse analysis, children's literature, POSgrams, frequency information, semantic tagging

1. Introduction

This article reports on an investigation into the language used in imaginative fiction written for a child audience. The analyses discussed here were part of an exploratory study of the potential for using corpus evidence with primary school children (8–10 year olds) in their learning about language in an L1 context.[1] For the purposes of this project, a small corpus was created of texts written for a child audience, taking texts from the British National Corpus.[2] This corpus was given the acronym 'CLLIP', which stands for 'Corpus-based Learning about Language In the Primary-school'. It contains 40 texts, and includes a range of types of writing, predominantly imaginative fiction, but also others such as Brownie annuals and a book on Christianity.[3]

International Journal of Corpus Linguistics 12:1 (2007), **1–23**.
ISSN 1384–6655 / E-ISSN 1569–9811

Although one strand of the research was to discover how younger learners respond directly to corpus-based approaches to learning about language — and some of our data were therefore collected in classrooms — a parallel objective was to explore features of the corpus itself. Pilot work with learners of the age-group we worked with (8–10-year-olds) had established that it was necessary for examples of language they encountered in concordance output to be drawn from texts that they might be familiar with or be likely to read. Thus we had, for pedagogic application, a 'specialized' corpus, prepared for this group of learners, but this raised some questions about the corpus itself. To what extent does such a corpus represent a 'scaled-down' version of 'language in general', simplified to be made accessible to these younger readers, and to what extent is it distinctive in its linguistic properties? If corpus-based teaching were to be extended to this population on a larger scale, would a corpus such as this one be adequate for pedagogical purposes, or would the range of investigations that teachers and learners could conduct with it be restricted because of the properties of the language used in the texts it contains?

We therefore decided to conduct linguistic analyses of the corpus, in order to establish what the distinctive features of writing aimed at a child audience might be, in comparison with writing aimed at a 'general' audience — which effectively means adult readers. We reduced our original CLLIP corpus from 40 texts to the 30 that were classified as imaginative fiction, and we then created a comparison corpus of imaginative fiction written for an adult audience, composed of 317 texts from the BNC. Ideally, the CLLIP corpus would contain many more texts, but it was essential to have a set of texts that was already POS-tagged, and the BNC contains very few texts written for a child audience. Comparisons between the two corpora would go some way towards answering the question: 'Does writing for children demonstrate different linguistic properties from writing for adults?' However, although any differences found should be largely attributable to the adult/child variable, the restriction to fiction excludes other kinds of differences, so, as a further point of comparison,

Table 1. The three corpora used in this study. 'COMP' is the abbreviation used throughout for the corpus of adult fiction

CLLIP corpus	imaginative fiction written for child audience, from the BNC	30 texts	698,286 tokens
COMP corpus	imaginative fiction written for an adult audience, from the BNC	317 texts	12,869,883 tokens
Newspaper corpus	newspaper texts from the BNC	114 texts	1,270,798 tokens

we created a third corpus, of newspaper texts, extracted from the BNC, to allow us to contrast the features of imaginative fiction writing in general with those of newspaper writing. This should help to shed light on the kinds of patterns in language which younger readers may need to learn about as they progress through their schooling.

As can be seen, the three corpora are of different sizes, an issue which is discussed in Section 3 below.

2. Research questions

The general question underpinning the investigation reported here is: 'What is distinctive about the discourse of the CLLIP corpus?' The first more specific question to be derived from this one examines relative frequencies within the three corpora, with an interest not only in the frequencies of words, but also in sequences of words.

1. What similarities and differences are there in the overall frequencies of words, parts of speech, and word and POS sequences in the three corpora?

The investigation was focused further in order to explore the issue of whether language deployed in writing for children can be seen to represent the world and human experience differently from the ways in which they are represented in writing for adults. Researchers into fiction written for children have noted the role it plays in their socialisation, and how these texts are inevitably suffused with ideology (Hunt 1992; Lesnik-Oberstein 1994; Sealey 2000; Stephens 1992; Wall 1991). Few, however, have taken a corpus linguistic approach to analysis, although there are some exceptions. Stubbs (1996) used corpus techniques to analyse gender-related differences in two specific texts addressed to boys and girls, and he cites Baker and Freebody's (1989) analysis of the different distributions of the lemmas GIRL, BOY and CHILD and their collocates in initial reading books (p. 94). Knowles and Malmkjaer's study (1996) is in the 'critical linguistics' tradition, concerned with how 'an awareness of patterns of textual structure and of language choices may provide information about how the author wants his/her readers to view society' (p. 263), and they use a Hallidayan framework for their clause level analysis and 'a neo-Firthian framework for the analysis of collocation' (p. 69). Concordancing was used in some parts of this study to analyse collocational patterns, with a particular interest in how the selection of linguistic expression functions ideologically in this genre of discourse, which is:

> culturally formative, and of massive importance educationally, intellectually, and socially. Perhaps more than any other texts, they [novels for children] reflect society as it wishes to be, as it wishes to be seen, and as it unconsciously reveals itself to be.
>
> (Hunt 1990:2, cited in Knowles & Malmkjaer 1996)

Our own approach is rather different, and concerned particularly with the ways in which the world is represented to the child reader — itself an issue which has occupied many researchers into children's literature. Wall (1991), for example, suggests that "… the narrator-narratee relationship … is the distinctive marker of a children's book" (p. 9), and she provides an extensive overview of the changing ways during the last two centuries in which adults have met the challenge of writing **for** children, and **about** children's concerns, while standing in relation neither to the world nor to their subject matter as the implied child reader does. Writers of fiction for children have a range of options about their authorial stance, as Wall and others have demonstrated, including: first-person narrators who are ostensibly children themselves; the narrator-as-adult who presents events as though remembered from childhood; anthropomorphic animal narrators with whom child readers can identify as both in the world (perhaps with an 'adult' experience and perception) and yet on its margins (since they are non-human); narrators who, though detached themselves, "allow the narratee to stand in [a child character]'s shoes and to move as she moves" (Wall 1991:197). We decided to explore further the linguistic means by which the world — including the fictional, even fantasy, world — is represented as though from a child's perspective.

We wondered, firstly, whether specific lexical items are used in particular ways when representing the world to child readers; and, secondly, what evidence there is for depictions of the relationship between protagonist(s) and world being handled in a distinctive way compared with the depiction of that relationship in fiction aimed at adult readers.

The following two questions were thus posed of the two imaginative fiction corpora:

2. Are there differences between the child and adult fiction corpora in the uses of particular lexical items?
3. Is the discourse of the CLLIP corpus distinctive in its representation of the self in the world and of the world by the self?

3. Methods and methodological issues

We chose to use texts from the BNC for a number of reasons: firstly, as the texts are in the BNC, they are readily available both for us as researchers, but also for others who might want to do similar work; secondly, the texts are tagged for part-of-speech (POS), which offers the possibility for users of the corpus to make searches by POS and — importantly for our classroom research purposes with young learners — the tagging made it possible to create a colour output for concordance lines in which words were coloured according to POS. However, BNC texts are not parsed, which restricts the range of possible questions that can be asked of the data, without advanced processing of the files. For this study, it was not feasible to parse the data, and so the questions posed are focused on frequency and collocational information rather than complex syntactic patterning.

We decided to compile our basic word lists according to types (strings of characters) rather than lemmas, on the basis that information regarding relative frequency of different morphological forms can be revealing (see, for example, Sinclair's (1991) discussion of *yield*, *yielding* and *yielded*). Frequency information was also compiled for parts of speech, to show the overall relative usage of different parts of speech, and the most frequent tokens in each part of speech category. In order to find out about word sequences, we then compiled frequency lists for 4-grams (sequences of four strings), and also for 6 part sequences of POS tags, and compared the frequency lists.

The method used for extracting POS information from the three corpora (given that the BNC files are already tagged for part of speech), was to compile separate files for each corpus for each part of speech. All occurrences of a particular POS tag and the string of letters following it were extracted (using purpose-built filters in the data manipulation program TextPipe Pro), then the tags were stripped and wordlists constructed. The wordlists and other statistical information were derived using Oxford Wordsmith Tools, Version 4. Identification of n-grams and of POS-grams was performed using two applications written by William Fletcher, called kfNgram and POSonly. For concordance work on particular expressions (single- and multi-word units), MonoConc Pro 2.2 was used, as this (unlike Oxford WordSmith Tools) allows the formation of regular expression searches.

As the size of the three corpora, in numbers of tokens, was unequal, all frequency counts were normalized to show the frequencies as percentages, so that the figures would be comparable.

4. Results & Discussion

4.1 Frequency of types

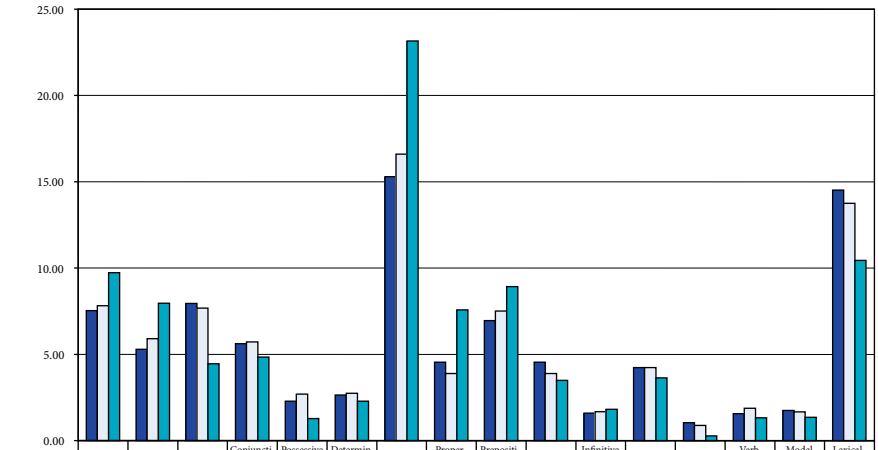
Table 2 shows the ten most frequent types in the three corpora, and it can be seen that there is a relatively similar profile between the two imaginative fiction corpora, with mostly the same types appearing in the top ten, and, perhaps most notably, *the* occurring at just over 5%. The newspaper corpus, on the other hand, has *the* at 7%, which suggests that there is a higher degree of nominalisation, a conjecture which is supported by the appearance of *of* in second place, with a percentage rating of 3.23% (compared to 1.67% and 2.11% in the other two corpora) — the high frequency of *of* is likely to be due to its use in the pattern: *the*+N+*of*+N. The newspaper corpus also has no personal pronouns in the top ten, unlike the fiction corpora.

Table 2. The ten most frequently occurring types in each of the three corpora

	CLLIP			COMP			News		
N	Word	Freq	%	Word	Freq	%	Word	Freq	%
1	*the*	35,868	5.14	*the*	651,685	5.06	*the*	89,143	7.00
2	*and*	19,566	2.80	*to*	335,442	2.61	*of*	41,156	3.23
3	*to*	17,440	2.50	*and*	335,074	2.60	*to*	36,552	2.87
4	*a*	14,397	2.06	*a*	284,278	2.21	*a*	28,599	2.25
5	*he*	12,234	1.75	*of*	272,255	2.11	*and*	28,437	2.23
6	*of*	11,691	1.67	*I*	221,448	1.72	*in*	27,120	2.13
7	*it*	11,481	1.64	*he*	208,592	1.62	*that*	13,005	1.02
8	*was*	11,467	1.64	*was*	202,970	1.58	*for*	12,863	1.01
9	*you*	9,927	1.42	*she*	190,574	1.48	*is*	11,717	0.92
10	*I*	9,677	1.39	*in*	180,394	1.40	*was*	10,910	0.86

4.2 Frequency of parts of speech

Figure 1 shows the relative occurrence of different parts of speech (following the classification of parts of speech used in the CLAWS tagset) in the three corpora, expressed in percentages. What is most striking about the figures is, once again, the similarity in profile between the first two columns for each POS, which represent the percentage figures for the two imaginative prose corpora. There are slightly higher proportions of nouns in the COMP corpus and of proper nouns and pronouns in the CLLIP corpus, suggesting a higher degree of reference to people in CLLIP, but overall the profiles are similar, especially



	Article	Adjective	Adverb	Conjunction	Possessive	Determiner	Noun	Proper noun	Preposition	Pronoun	Infinitive *to*	Verb *be*	Verb *do*	Verb *have*	Modal verb	Lexical verb
CLLIP	7.53	5.29	7.95	5.62	2.29	2.63	15.29	4.53	6.95	4.53	1.59	4.23	1.03	1.56	1.74	14.51
Comparison	7.82	5.91	7.68	5.72	2.69	2.74	16.60	3.89	7.51	3.89	1.68	4.23	0.88	1.87	1.65	13.74
Newspaper	9.73	7.96	4.45	4.83	1.28	2.28	23.15	7.57	8.92	3.49	1.81	3.64	0.27	1.32	1.35	10.44

Figure 1. Relative frequency of different parts of speech in each of the three corpora shown in percentages. In each set of three columns, the CLLIP corpus is the first column (with dark shading), the COMP corpus is in the middle and the News corpus is on the right.

when compared with the profile for the News corpus. In the latter, we see a high proportion of nouns, articles, adjectives and proper nouns, and a far lower use of lexical verbs and adverbs (proportionately), which confirms the conjecture made in the previous section that a distinctive feature of the newspaper corpus is the high degree of nominalisation.

The higher proportions of lexical verbs and adverbs in the fiction corpora, on the other hand, indicate a more central role for human agency and actions in the worlds of fiction.

4.3 Frequency data — lexical verbs

Table 3 shows the top ten lexical verb forms in the three corpora. The pattern that we see in this, and the next two sections, is that the CLLIP and COMP corpora continue to show a very similar profile, with the newspaper corpus providing a useful contrast. In the case of the lexical verbs, ten of the words (types) in the CLLIP list also occur in the COMP list, with the only differences being *going* and *made*. The high frequency of *got* is explicable by its delexicalised (or rather 'desemanticised' (Stubbs 2001:32)) status, and it is arguable that the other verbs in the lists represent the 'core' processes with which narrative

Table 3. The 10 most frequent lexical verb forms in the three corpora. The figures in the 'Freq' columns show the raw frequency of the lexical verb form, and the figures in the '%' column show what percentage of all the lexical verbs in the corpus that particular type accounts for.

CLLIP	Freq	%	COMP	Freq	%	News	Freq	%
said	6,711	6.30	*said*	64,513	3.71	*said*	5,715	4.27
see	1,420	1.33	*know*	26,015	1.50	*made*	1,140	0.85
know	1,355	1.27	*see*	20,345	1.17	*told*	828	0.62
go	1,319	1.24	*think*	18,048	1.04	*take*	805	0.60
get	1,309	1.23	*get*	17,530	1.01	*make*	798	0.60
looked	1,251	1.17	*go*	17,455	1.01	*say*	698	0.52
got	1,237	1.16	*looked*	16,611	0.96	*says*	688	0.51
come	1,071	1.01	*thought*	15,696	0.90	*go*	660	0.49
going	1,039	0.98	*come*	15,126	0.87	*put*	631	0.47
think	1,005	0.94	*got*	14,599	0.84	*get*	583	0.44

is primarily concerned, so that the marked similarity between the adults' and children's fiction lists is unsurprising. Stories are about protagonists with whom readers are invited to identify. These protagonists make their way through their fictional world (*go, come*), and readers are told about these characters' perceptions of it and their experience in it (*see, think / thought, know, looked*).

One distinctive feature of the CLLIP corpus is that *said* is proportionately more common (6.3%) than it is in the COMP corpus (3.7%) which is indicative of the prevalence of direct speech in the CLLIP corpus, a point which is confirmed by the high frequency (relatively) of the entities *&bquo;* and *&equo;* (beginning and end quote marks) in that corpus.

4.4 Frequency — adjectives

Table 4 shows the ten most frequent adjectives. In this respect again, as can be seen by the high number of shaded items in the CLLIP and COMP columns, the two fiction lists are much more similar to each other than either of them is to the News corpus. In the latter, adjectives denote social attributes, indicating the groups to which people belong; these, of course, are salient features of the kinds of relations among collective forces with which news reporting is concerned. It is worth noting, too, that the most frequent adjectives in the CLLIP corpus do more work, as it were, than those in the COMP corpus; when the percentage figures are added together, the CLLIP adjectives account for 14.63% of the total adjective use in the CLLIP corpus, while the figure for the COMP

Table 4. The 10 most frequent adjectives in the three corpora. The figures in the 'Freq' columns show the raw frequency of the adjective, and the figures in the '%' shows what percentage of all the adjectives in the corpus that particular adjective accounts for. Adjectives which are shaded occur in more than one of the three lists.

CLLIP	Freq	%	COMP	Freq	%	News	Freq	%
old	788	2.13	*good*	12,071	1.62	*new*	2,057	2.02
good	704	1.91	*other*	10,702	1.43	*labour*	1,538	1.51
little	656	1.78	*old*	10,246	1.37	*other*	1,385	1.36
other	570	1.54	*little*	8,494	1.14	*political*	1,236	1.22
long	485	1.31	*small*	6,783	0.91	*British*	1,220	1.20
small	408	1.1	*sure*	6,773	0.91	*national*	1,042	1.02
big	400	1.08	*long*	6,502	0.87	*European*	810	0.80
great	391	1.06	*young*	5,965	0.8	*prime*	731	0.72
sure	355	0.96	*new*	5,693	0.76	*public*	708	0.70
right	344	0.93	*right*	5,515	0.74	*Soviet*	690	0.68

corpus is 11.28%. Despite the similarities between CLLIP and COMP, these lists provide some small evidence relevant to our conjecture that there might be contrasting depictions of 'being in the world' in fiction for adult and child readers respectively. Relative to people of a child's size and age, more people and things are likely to be perceived as *big*, and fewer as *young*; the former item is more frequent in CLLIP, and the latter more frequent in COMP.

4.5 Frequency — noun*s*

With nouns, once again, we find considerable overlap between CLLIP and COMP, and a very different list in News. Consistent with the claim made above, that from these analyses we can almost begin to discern a 'core' vocabulary for narrative texts, the most frequent items denote aspects of the experience of fictional protagonists: embodied people (*man, people, head, eyes, face, hand*); settings where events take place (*house, room, door*); items relating to chronology (*time, day*). However, frequency data such as these, of course, should be used with caution and it would be unwise to draw any firm conclusion about the meanings of these items without examining the contexts in which they occur.

We note that the two fiction lists differ in the inclusion in CLLIP but not COMP of *thing*, and in COMP but not CLLIP of *hand* and *room*. *Man* is in both lists, but much more frequent in COMP (0.86% compared with 0.57%); *eyes* too appears in both. We consider in more detail below words such as *hand*, which denote parts of the body, as an example of the more subtle differences

Table 5. The 10 most frequent nouns in the three corpora. The figures in the 'Freq' columns show the raw frequency of the noun, and the figures in the '%' show what percentage of all the nouns in the corpus that particular noun accounts for. Nouns which are shaded occur in more than one of the three lists.

CLLIP	Freq	%	COMP	Freq	%	News	Freq	%
time	1,265	1.18	*time*	23,446	1.12	*party*	2,865	0.96
way	917	0.86	*man*	18,014	0.86	*government*	2,863	0.96
thing	766	0.72	*way*	17,005	0.81	*people*	2,380	0.80
head	760	0.71	*eyes*	16,670	0.79	*years*	1,796	0.60
eyes	756	0.71	*face*	13,058	0.62	*year*	1,443	0.49
face	673	0.63	*head*	12,349	0.59	*time*	1,427	0.48
door	672	0.63	*door*	11,297	0.54	*cent*	1,398	0.47
people	666	0.62	*hand*	11,135	0.53	*per*	1,398	0.47
day	664	0.62	*room*	10,444	0.5	*minister*	1,275	0.43
man	606	0.57	*people*	10,144	0.48	*police*	1,148	0.39
house	572	0.53	*day*	9,742	0.46	*election*	1,085	0.36

we found when looking at the semantic distributions of items in the two fiction corpora (see Section 4.8 below).

4.6 N-gram data

The pattern observed so far is that the two imaginative prose corpora are remarkably similar in profile, especially in comparison to the newspaper corpus. To test this similarity further, we decided to look at the most common N-grams (otherwise known as lexical bundles, or clusters) in the three corpora. We identify 4-grams as the most revealing (as do Biber et al. (1999) and Partington & Morley (2004)), on the basis that there are usually, in a sizeable corpus, too many 3-grams to make them tractable to careful analysis and too few 5-grams. Following this preference, we calculated the most frequent 4-grams, using the kfNGram programme, and the top 22 4-grams for each corpus are shown in Table 6. The figure of 22 was chosen as there were only 22 4-grams that were used at least, on average, once per text.

The third column in Table 6 shows, firstly, the position of the 4-gram given in the first column in the CLLIP corpus, and, secondly, the ranking of the same 4-gram in the COMP corpus. Where the letter X appears, this indicates that the given 4-gram does not appear in the first 22 of the 4-grams in the COMP corpus. As can be seen, there are only 4 X's, which means that 18 of the 4-grams appear in the top 22 of both lists, which is a remarkably high degree of

Table 6. The top 22 4-grams in each of the three corpora. The third column indicates the relative frequency positions of 4-grams in the CLLIP and COMP corpora — the first figure is the position in the CLLIP corpus and the second is the position in the COMP corpus. 'X' indicates that the 4-gram does not appear in the top 22; only 4 of the 4-grams in the CLLIP corpus do not feature in the top 22 of the COMP corpus.

CLLIP (30)	Freq	CLLIP–COMP	COMP (317)	Freq	Newspaper (114)	Freq
in the middle of	70	**1–5**	*the end of the*	1041	*per cent of the*	172
the top of the	68	**2–10**	*for the first time*	963	*the end of the*	145
the end of the	65	**3–1**	*the rest of the*	905	*for the first time*	118
the edge of the	63	**4–9**	*at the end of*	836	*at the end of*	108
in front of the	60	**5–11**	*in the middle of*	733	*secretary of state for*	93
the back of the	51	**6–8**	*at the same time*	673	*one of the most*	81
the middle of the	51	**7–12**	*I don't want to*	669	*at the same time*	78
at the end of	50	**8–4**	*the back of the*	654	*a member of the*	76
I don't want to	49	**9–7**	*the edge of the*	645	*as a result of*	69
the rest of the	48	**10–3**	*the top of the*	556	*the secretary of state*	69
for the first time	45	**11–2**	*in front of the*	515	*the labour party conference*	68
out of the window	45	**12–X**	*the middle of the*	511	*the second world war*	63
the side of the	45	**13–13**	*the side of the*	501	*by the end of*	61
on the other side	42	**14–20**	*he was going to*	490	*is likely to be*	56
the other side of	40	**15–15**	*the other side of*	488	*letter to the editor*	56
what are you doing	40	**16–X**	*what do you mean*	481	*was one of the*	55
but there was no	36	**17–X**	*I don't know what*	475	*will be able to*	55
other side of the	34	**18–X**	*he shook his head*	453	*the house of commons*	54
what do you mean	34	**19–16**	*she shook her head*	431	*the house of lords*	53
at the same time	33	**20–6**	*on the other side*	427	*the rest of the*	53
he was going to	31	**21–14**	*on the edge of*	415	*in an attempt to*	47
I don't know what	30	**22–17**	*what do you think*	412	*is one of the*	47

similarity. Only four of the N-grams in the Newspaper list also appear in the other two: *the end of the*, *for the first time*, *at the end of* and *at the same time*. (Note, incidentally, that some 4-grams are duplications, as they are parts of 5-grams: for example, *at the end of* and *the end of the* are two portions of the larger 5-gram *at the end of the*.)

Continuing our conjecture about the linguistic characteristics of fiction, we would expect the two fiction corpora to contain N-grams denoting times and places where narrative events unfold, and this indeed seems to be the case: *for the first time*, *out of the window*. Also prevalent in this set of data are expressions denoting interpersonal interaction and communication: *what do you mean, she shook her head.* News discourse, by contrast, indicates a concern

with identifying social relations, through naming social positions (*secretary of state for*), groupings and group membership (*the house of lords*, *the labour party conference*), and historical events (*the second world war*).

4.7 POS-gram data

A further line of enquiry[4] was to ascertain the most frequent part-of-speech sequences in each corpus. This was accomplished by stripping the corpora of all words and all non-POS tags, and then running a small programme developed by William Fletcher, called 'POS only', which identifies the most frequent sequences. There is a small weakness in this method, as it does not take into account sentence boundaries, and some of the sequences are in fact intersentential sequences, but the figures are robust enough to show the most frequent patterns. After experimenting with different lengths of sequence, we finally decided on a setting of 6 POS tags as the most useful length, as this overcame the problem of an excessive number of sequences inherent in any choice of 5 or fewer POS, and also provided a richer source of insight than the use of 7 or more, where the variety of sequences was extremely limited, and tended more towards highly conventionalised expressions. The top 6-POS-gram sequences for each corpus are shown in Table 7.

Table 7. The top four 6-POSgram sequences for each of the three corpora.

Corpus	1	2	3	4	5	6	Raw	No of files	No of tokens	Rounded	
CLLIP	prep	art	NN1	*of*	art	NN1	523	30	698,286	7.5	A
	prep	art	NN1	prep	art	NN1	385			5.5	B
	art	adj	NN1	prep	art	NN1	274			3.9	C
	art	NN1	prep	art	adj	NN1	186			2.7	D
COMP	prep	art	NN1	*of*	art	NN1	9167	317	12,869,883	7.1	A
	prep	art	NN1	prep	art	NN1	5259			4.1	B
	art	adj	NN1	prep	art	NN1	4907			3.8	C
	art	NN1	prep	art	adj	NN1	3555			2.8	D
News	prep	art	NN1	*of*	art	NN1	848	114	1,270,798	6.7	A
	art	NN1	of	art	adj	NN1	776			6.1	
	art	adj	NN1	prep	art	NN1	655			5.2	C
	adj	NN1	prep	art	adj	NN1	462			3.6	D

The columns numbered 1–6 show the POS in the sequences and these are followed by the number of actual occurrences of each sequence, the number of files in the corpus, the number of tokens and then a rounded factoring of each 6-POS-gram sequence, which is arrived at by taking the number of actual occurrences, dividing by the number of tokens in the corpus and then multiplying by 10,000. While this measure does not represent a real indication of the POS-grams' occurrence relative to other POS-grams in the corpus (as each POS-gram can contain parts of other POS-grams within it), it allows a comparison of relative frequency to be made between corpora. The final column identifies a particular 6-POS-gram sequence, and it can be seen that the sequence 'preposition-article-singular noun-*of*-article-singular noun' is the most frequent sequence in all three corpora. Furthermore, the repeated set of A-B-C-D (each sequence is identified with a single letter identifier) shows that the top four sequences in both the CLLIP and the COMP corpora are the same, which is remarkable. However, three of the four sequences appear in the top four for each corpus, with only the B sequence not appearing in the News corpus.

The recurrence of the A sequence prompted us to investigate what these patterns were and why they might be so prevalent in all three corpora. Examples of these sequences are: *at the top of the hill*; *in the bottom of the bag*; *to the end of*

Table 8. The most common nouns that appear as the first noun in the type A sequence: Prep+art+noun+*of*+art+noun

Word	Freq	%	Word	Freq	%	Word	Freq	%
CLLIP			**COMP**			**News**		
end	49	**9.44**	*end*	757	**8.30**	*end*	84	**9.98**
middle	47	**9.06**	*back*	546	**5.98**	*back*	26	**3.09**
edge	45	**8.67**	*edge*	478	**5.24**	*heart*	21	**2.49**
back	44	**8.48**	*middle*	464	**5.08**	*middle*	20	**2.38**
top	41	**7.90**	*side*	384	**4.21**	*result*	18	**2.14**
side	38	**7.32**	*top*	323	**3.54**	*rest*	16	**1.90**
bottom	19	**3.66**	*centre*	297	**3.25**	*time*	13	**1.54**
rest	18	**3.47**	*rest*	288	**3.16**	*centre*	13	**1.54**
centre	13	**2.50**	*bottom*	232	**2.54**	*wake*	11	**1.31**
corner	12	**2.31**	*corner*	179	**1.96**	*start*	11	**1.31**
front	10	**1.93**	*front*	150	**1.64**	*side*	11	**1.31**
foot	7	**1.35**	*direction*	116	**1.27**	*future*	11	**1.31**
floor	6	**1.16**	*foot*	115	**1.26**	*top*	10	**1.19**
head	5	**0.96**	*door*	85	**0.93**	*head*	9	**1.07**
direction	5	**0.96**	*rear*	76	**0.83**	*Department*	9	**1.07**

Table 9. The six most frequent nouns as shown in Table 8, with the aggregate of percentage of nouns in the gap accounted for by these types.

CLLIP	%	COMP	%	News	%
end	**9.44**	*end*	**8.30**	*end*	**9.98**
middle	**9.06**	*back*	**5.98**	*back*	**3.09**
edge	**8.67**	*edge*	**5.24**	*heart*	**2.49**
back	**8.48**	*middle*	**5.08**	*middle*	**2.38**
top	**7.90**	*side*	**4.21**	*result*	**2.14**
side	**7.32**	*top*	**3.54**	*rest*	**1.90**
Total (rounded)	**51**		**32**		**22**

the lane. We therefore extracted all the occurrences of each sequence, by means of regular expression searches in the original files, using Monoconc Pro, and created spreadsheets of the sequence for each 6-POS-gram. From this database, we then extracted all the nouns that occur in third position of the A type sequence; for example, in the case of *at the top of the hill*, we extracted the noun *top*. These nouns were counted and the results tabulated to produce Table 8 above, in which the 15 most frequent nouns to appear in third position in the A sequence are shown, for each of the three corpora. The two columns after the noun show the raw frequency and then the percentage of all occurrences of this 6-POS-gram in which the noun appears.

It can be seen that once again the profiles of the CLLIP corpus and the COMP corpus are remarkably similar. Of the 15 most frequent nouns in the CLLIP corpus, only two (*head* and *floor*) do not appear in the top 15 of the COMP corpus list. By contrast, only eight appear in the Newspaper corpus list. The majority of the nouns in the two fiction corpora listed in Table 9 are used in expressions of location, as in the examples above (*at the top of the hill*; *in the bottom of the bag*), of direction (*to the end of the lane, in the direction of the house*) and of temporality (*at the end of the day, for the rest of the afternoon*). The news corpus list contains items used in expressions of causality (*in the wake of, as a result of*) and also expressions that work at a higher level of metaphoricity, such as *at the heart of the controversy*, where an abstract notion (*controversy*) is conceived of physically, as having a heart, which is used both metaphorically and also with the added evaluative load of implicit importance.

A further point to be made is that the top six nouns in the CLLIP data in Table 8 occur relatively frequently in the type A sequence. This is demonstrated in Table 9, in which it can be seen that the top six nouns are used in 51% of the occurrences of the A sequence, a far higher figure than those for the other two corpora.

4.8 Semantic analysis

The analysis so far has focused primarily on frequency data, addressing in the main the first of our three research questions (i.e. What similarities and differences are there in the overall frequencies of words, parts of speech, and word and POS sequences in the three corpora?). Our findings indicate a remarkable similarity between the two fiction corpora, pointing to the existence of expressions consistent with the key characteristics of narrative, and a contrast with the news corpus on all the measures we used.

However, these analyses have suggested some subtle differences between fiction for adults and for children, and in the following sections, we investigate the uses of particular words and phrases by writers in the CLLIP and the COMP corpora in more detail, with a view to addressing the question of whether the discourse of the CLLIP corpus is distinctive in its representation of the self in the world and of the world by the self.

From a purely numerical point of view the two corpora are broadly similar, as we have seen, in the proportion of uses of POS, in the most frequently used lexical items and in the 6-POS-grams used. We are interested to identify the distinctive features of the world and perspectives on the world that are presented in literature written for children. To investigate this, we have used a web interface developed by Paul Rayson for the semantic analysis of corpus data. The WMatrix tagger (www.comp.lancs.ac.uk/ucrel/wmatrix/) automatically tags input data using the USAS tagset, a set of semantic categories loosely based on McArthur (1981). (Details of the tagging system can be found at www.comp.lancs.ac.uk/ucrel/usas/.)

A comparison of the overall weighting of different semantic categories, based on log-likelihood scores, is shown in Table 10. The world as construed in literature written for a child audience appears to be highly distinctive in the importance of animals generally (living creatures), food and plants. Communication stands out too, with, as observed in 4.3 above, a high degree of direct speech and also of speech acts. It is a world of bravery and fear, where movement and speed are important, a world of objects, and a world in which sight and size are emphasised. The world of adult fiction, on the other hand, is distinguished by intimacy and sexuality, and is a world in which beliefs and broad questions about life predominate, and is a world of social laws. While there is much more that could be said about the semantic analysis, space is limited here and we simply observe that the representations of world and self in the two corpora are clearly different.

Table 10. A summary table showing the ranking of semantic categories that are distinctive in either corpus, based on log-likelihood scores.

Semantic categories more highly represented in the CLLIP corpus	Log-likelihood score	Semantic categories more highly represented in the COMP sub-corpus	Log-likelihood score
Living creatures generally	429.1	Relationship: intimate/sexual	173.95
Personal names	263.9	Drinks	145.29
Food	254.2	Life and living things	135.35
Plants	215.2	Law and order	124.75
Objects generally	201.8	Anatomy and physiology	90.35
Communication in general	163.8	Medicines	79.59
Time: general, future	113.3	Strong or weak	69.98
Measurement: size	105.3	Thought, belief	66.94
Sensory: sight	84.0		
Speech acts: communicative	83.3		
Moving, coming and going	77.9		
Fear, bravery, shock	64.1		
Measurement: speed	63.0		
Location and direction	58.4		

4.9 Words denoting parts of the body

As noted above, lexical items denoting the embodied self are common in both fiction corpora, so we next take the names for different parts of the body, and investigate how they are used in each corpus.[5] Various observations we have made as we investigate these data have pointed to the possibility that fiction written for children may make greater use of the more literal meanings of words and less use of their more figurative or metaphorical meanings than the discourse found in the comparison corpora. We therefore set out to determine firstly, the relative use of figurative expressions involving names for parts of the body, and secondly, the variation in experience of the world through the body represented in each fiction corpus.

The first word investigated was NECK, and the most frequent left collocates are shown in Table 11. The lemma NECK was chosen in preference to any of the most frequent lemmas such as HEAD or EYE as these would have thrown up too many instances to examine in detail. NECK occurs 90 times in the CLLIP corpus and 1897 times in the COMP corpus.

In many of the instances of *the neck*, NECK is used of entities other than living bodies, as in *the neck of the bottle* or *the neck of her dress*, and the occurrences of *the* with NECK can be taken to give an indication of the relative

Table 11. The most frequent left collocates of NECK in the CLLIP and COMP corpora.

	CLLIP	COMP
Her	15 (16.7%)	516 (27.2%)
His	31 (34.4%)	533 (28.1%)
The	14 (15.6%)	285 (15.0%)
*(a)round * neck*	18 (20%)	409 (21.6%)

instances of the use of NECK in a figurative sense. As can be seen from the percentage figures, *the* is a left collocate of *neck* in approximately 15% of the instances of NECK in both corpora, which provides no evidence to suggest that there are more figurative senses of the word in the CLLIP corpus than in the COMP corpus. There are, of course, other uses of NECK that do not involve *the* as a left collocate and which are used in a figurative sense, such as *stick your neck out* (from the CLLIP corpus) but analysis of the concordance lines found only a slightly higher number of these in the COMP corpus.

From the data, it appears that there is greater mention of male than female characters' necks in the CLLIP corpus, relatively speaking, but a reading of the concordance lines shows that several of the instances of *his neck* actually refer to male animals rather than male human characters in the stories. In the CLLIP corpus, there is little description of physical contact between characters; what is *round the neck* is usually an item of clothing, some form of pendant or a pair of violent hands, and where intimacy is expressed, it is between a child and an animal. In the adult stories, by contrast, the neck is often the site of desire — whereas in the CLLIP corpus, it tends to be a site of pain — as well as a place for ornamentation, often an ornament that symbolises the relationship of the wearer to other people in the story. Furthermore, what is placed around the neck is often the arms of a loved one.

The second body part explored was FINGER. The concordance lines were examined and categorised as figurative or literal, according to the context in which FINGER appeared. Comparison of the number of uses of FINGER in a figurative sense again did not show a clear predominance in the COMP corpus, although the evidence suggests slightly more differentiated use: 13% in the CLLIP corpus and 19% in the COMP corpus. To investigate this question more rigorously, however, a larger children's corpus is needed, as there is a high chance that the differences in proportions are due to the small number of texts in the present CLLIP corpus.

Verbs that are used with FINGER in the CLLIP corpus include JAB, PROD, LAY, RUN and PUT, while adjectives that premodify it include *accusing* and *admonishing*. The actions that fingers are used to perform include drawing,

indicating the need for silence and pulling triggers. In the COMP corpus, the verbs that precede FINGER are JAB, RAISE, WAG, POINT, RUN and PUT, while premodifying adjectives are *furtive*, *tentative*, and *negligent*. The purposes fingers are put to are those of communicating, feeling (contours and textures) and wearing rings. This comparison indicates how representations of the idea of the human finger, while overlapping to a great extent in the two corpora, also differ in subtle ways. One such is the metonymic expression which transfers the quality of an action to the body part itself.

Although it was clearly not practical to explore all aspects of adjectival expressions, we did notice other examples of the contrast between children's and adult fiction consisting in differential uses of the same core vocabulary. *Angry*, for example, was found in the CLLIP corpus mainly to modify nouns denoting people (and sometimes their *faces* or inanimate objects, such as *fires*), whereas in the COMP corpus the entities described were often found to be less concrete: *an angry movement / gaze / silence*. Common sense suggests that progress in literacy from childhood to adulthood should be marked by a developing command of an increasing range of vocabulary. Less obvious, but suggested by our research, is the possibility that progress towards maturity in literacy would also be indicated by an ability to comprehend — and deploy — common adjectives in more abstract contexts (see Sealey & Thompson 2006).

4.10 Time and space

As discussed above, protagonists in children's stories may be represented as experiencing both time and space in slightly different ways from those in adult fiction. Just as the quality of size is relative, so time may be represented as having different qualities when perceived from a child's perspective. These areas of experience have both physical — or literal — constituents and sociocultural ones, as the following examples illustrate. Creatures or objects which pose no threat to adults may seem to tower over a child protagonist, and the very size of adult characters can contribute to their superordinate status. Time may pass more slowly in the child's perception, and it may also be more subject to regulation by others. We looked at the expression *in time*, generating the following set of concordance lines from the CLLIP corpus.

```
1  the direct route by road and arrived in time. She didn't intend to miss her
2  ight sometimes," but stopped herself in time. She handed him her ferns
3  very, very good thing you stopped us in time!" she said to Brenda.
4     lot if you think you can get back in time." She was beaten, and knew
```

5 the men from it to the others just **in time**. Six ships they had had
6 fficult for Gloria to get out of bed **in time**, so they arrived late.
7 He stopped just **in time**, teetering on the edge.
8 But she remembered **in time** that Gloria had warned her to
9 They were only just **in time**. The sky gave one almighty
10 whoever it was could not get there **in time**. The chair was slipping
11 He never arrived **in time** to find out who, or what it
12 Gloria was busy tapping her feet **in time** to the music.
13 got back to the garden shed **in time** to flake out until morning.
14 Nutty, last to disappear, was just **in time** to hear the visitor say

Below is a comparable set of selected concordance lines from the COMP corpus:

1 "no doubt it will all come out **in time**. Is there anybody who can vouch for
2 "Well, not at this moment **in time**, perhaps. But you never know,
3 er was imputing to me, but saw **in time** that to do so would be to rise
4 be pleaded, though if caught **in time**, the house could have been converted
5 t would occur in the soul and, **in time**, the fripperies of science would
6 r two. Everything would change **in time**. Things changed faster as time
7 ng his head from side to side, **in time** to the beat. He lay face down beside
8 talion were already moving off **in time** to take their appropriate places
9 nd Tummel Catchment Area. Just **in time** to preserve the high quality of
10 But if all goes well, **in time** we should be able to sell our cast
11 "They will see **in time**, when she is over this, she will
12 that particles move backwards **in time**. Without the emergence of Hitler

In the majority of the CLLIP examples, *in time* refers to the meeting of a deadline, the accomplishment of an action that must be completed before a penalty or some other unwanted outcome should occur. In the COMP corpus, *in time* can be used to refer to the same local sense of time (as in line 3, *saw in time*) but in other lines time is invoked on a larger scale, either in the sense of the centuries passing (line 6, *Everything would change in time*) or in terms of a gradual unfolding of events (line 10, *in time we should be able to …*).

In order to explore representations of space, we decided to choose a common 3- or 4-gram which includes at least two of the most frequent lexical items from other lists. Although we had established that many of the most frequent adjectives denote qualities of size, it was difficult to analyse them in their own right, because they throw up too much data (as they are, of course, so frequent). Adjectival *long*, for example, occurs 485 times in the CLLIP corpus, and 6,502 times in the COMP corpus. A look at the collocation patterns, however, shows that the most common words appearing to left and right of long are *a* and *the*, and *time* and *way* respectively. Highly frequent words are increasingly delexicalised

and can take on a wide range of meanings. From this statistical information, we chose to look at the range of meanings that a common 3-gram can convey, taking here as an example for discussion the 3-gram *a long way*, which occurs 44 times in the CLLIP corpus, and 276 times in the COMP corpus.

Only 3 out of 44 instances of *a long way* in the CLLIP corpus do not relate to physical distance directly. These instances are as follows:

1 paused, then added, 'It goes back **a long way** .' 'It's not my family's
2 in the dark and he'd have to go **a long way** round if he wasn't to upset
3 but cold it's inedible and there's **a long way** to go till lunchtime.' 'I

In the first line, *a long way* forms part of a longer expression *go(es) back a long way*, where *way* is used in a temporal sense, and this invokes the common metaphor of 'time is a journey' (Lakoff & Johnson 1980). This metaphor also underlies the third line, in which the speaker asserts that lunchtime is, as it were, still a long way away (a long time away). The second line is somewhat ambiguous, but one can read it as saying that the character will have to go a long route to get to where he is going, and that the *way* in this case is synonymous with 'path' or 'route' rather than 'distance'.

In the CLLIP corpus, then, there are only these three exceptions to the general rule that *a long way* refers to physical distance. In the COMP corpus, by contrast, there are 108 instances of *a long way* (out of 276) that do not carry the sense of physical distance, and some examples of the various meanings that this expression can carry are given in the following lines:

1 could not imagine it. 'We've come **a long way**, you and I,' Michael went on.
2 of the sort that make a little go **a long way** . I reckon he was careful.'
3 sixty-five, and that's a hell of **a long way** away. What would they do in
4 der. 'I think that we are getting **a long way** from the subject,' he said.
5 he used to say. 'That girl will go **a long way** -' 'The further the better,' I
6 eer on his breath. 'And they'd go **a long way** to 'elp 'im if ever 'e 'ad
7 he said dryly. " The grave seems **a long way** off." " I keep saying the
8 lly a deer; a very dead deer, and **a long way** from fresh. Its eyes and part
9 ooked up at him sadly. He seemed **a long way** off. But now the truth was
10 ast of all because she still felt **a long way** from figuring him out! That

Lines 1, 3 and 7 are based around the metaphor 'time is a journey', or, alternatively, 'life is a journey', which is a frequent sense of the expression in the COMP corpus. As has been seen in the discussion of 'time' above, the perception of temporality in the adult fiction corpus is markedly different from that in the CLLIP corpus. Line 9 involves an element of abstraction with the distance being a mental, rather than a physical one, while line 4 is similar in that it

draws on the metaphor of 'topic of conversation as location'. Another common idea conveyed in the expression *a long way* is that of social improvement, as instanced in line 5, where going a long way suggests that the person will be successful in career or self advancement. This again exhibits a sense of time on a broad scale, and also of the future as a subject for speculation.

In the case of *a long way*, then, as with *in time* and the uses of NECK and FINGER, the senses in which the expression is used indicate clearly different representations of the fictional world as experienced by its characters and narrators.

5. Conclusions

The conclusions we draw from these analyses are tempered with caution, including the recognition that the number of texts from which the evidence is drawn, particularly in the case of the CLLIP corpus, is relatively small. Several lines of enquiry could be pursued much further if there were available a larger and more heterogeneous corpus of writing for a child audience, while even a more extensive corpus of children's fiction would provide material for a more thorough analysis of the questions discussed here. A parsed corpus, also, would facilitate analyses complementary to those we have reported on, and we suggest that these are all areas for corpus development, with the potential for a range of applications, including pedagogical ones.

Nevertheless, we believe that the analysis has generated some interesting findings, not least by deploying the approach using POS-grams which makes possible a novel perception of the frequency of the complex prepositional phrase. There is scope for further investigation of the roles that such phrases play within different forms of discourse, in their expression of a variety of relations (temporal, causal, spatial, etc) and of their role in grammar.

In respect of our overarching question, about the distinctive linguistic properties of fiction written for children, we conclude the following on the basis of this investigation. Linguistic analysis of the CLLIP corpus, carried out in comparison with the two other corpora, has revealed a close similarity between the two fiction corpora, in terms of the overall frequency lists, and the proportions of different parts of speech, prompting us to suggest that narrative fiction, regardless of the age of its intended audience, may be characterised by linguistic characteristics such as those we have identified. However, close analysis of concordance output for some particular words and phrases has identified some of the marked ways in which the world, and the human relation to that world, is represented in the language of writing for children.

Notes

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'. The authors are grateful for comments on an earlier version of this article from Ramesh Krishnamurthy, the editor of the journal and the anonymous referees.

2. This corpus was used for devising a set of worksheets and a number of hands-on computing activities with small groups of primary-school children; a number of initial findings from that fieldwork are reported in Sealey and Thompson (2004) and Sealey and Thompson (Forthcoming).

3. Brownies are a junior branch of the Girl Guides organisation, and 'annuals' are books published once a year, often at Christmas, containing collections of short stories, jokes, quizzes etc.

4. This approach was stimulated by discussion with Mike Stubbs and Bill Fletcher, for whose suggestions we are very grateful.

5. A similar line of investigation is reported in Sveen's (2006) study of gender differences in Victorian and contemporary children's fiction, which also makes use of the BNC.

References

Baker, C. & Freebody, P. (1989). *Children's First School Books.* Oxford: Blackwell.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* Harlow: Pearson Education Limited.

Hunt, P. (Ed.) (1992). *Literature for Children: contemporary criticism*. London: Routledge.

Hunt, P. (Ed.) (1990). *Children's Literature: the development of criticism*. London: Routledge.

Knowles, M. & Malmkjaer, K. (1996). *Language and Control in Children's Literature*. London: Routledge.

Lakoff, G. & Johnson, M. (1980). *Metaphors We Live By*. Chicago: University of Chicago Press.

Lesnik-Oberstein, K. (1994). *Children's Literature: criticism and the fictional child*. London: Clarendon Press.

McArthur, T. (1981). *Longman Lexicon of Contemporary English.* London: Longman.

Partington, A. & Morley, J. (2004). At the heart of ideology: Word and cluster/bundle frequency in political debate. In B. Lewandowska-Tomaszczyk (Ed.), *Practical Applications in Language and Computers* (pp. 179–192). Bern: Peter Lang.

Sealey, A. (2000). *Childly language: children, language, and the social world*. Harlow: Longman.

Sealey, A. & Thompson, P. (2004). "What do you call the dull words?" Primary school children using corpus-based approaches to learn about language. *English in Education,* 38 (1), 80–91.
Sealey, A. & Thompson, P. (2006). "Nice things get said": corpus evidence and the National Literacy Strategy. *Literacy,* 40 (1), 22–28.
Sealey, A. & Thompson, P. (Forthcoming). Corpus, concordance, classification: Young learners in the L1 classroom. *Language Awareness*.
Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.
Stephens, J. (1992). *Language and Ideology in Children's Fiction*. Harlow: Addison Wesley Longman.
Stubbs, M. (2001). *Words and Phrases: corpus studies of lexical semantics*. Oxford: Blackwell.
Stubbs, M. (1996). *Text and Corpus Analysis*. Oxford: Blackwell.
Sveen, H. A. (2006). *"Honourable" or "Highly-Sexed": Adjectival Descriptions of Male and Female Characters in Victorian and Contemporary Children's Fiction* (Acta Anglistica Upsaliensia 129). Uppsala: Acta Universitatis Upsaliensis.
Wall, B. (1991). *The Narrator's Voice: the dilemma of children's fiction*. London: Macmillan.

Corpus Tools

KfNgram (2002). William Fletcher, http://www.kwicfinder.com/kfNgram/kfNgramHelp.html
MonoConc Pro 2.2 (1996, 2002). Mike Barlow, http://athel.com/index.php
Oxford WordSmith Tools (2005). Oxford University Press, http://www.oup.co.uk/episbn/0-19-459400-9
POSonly (2004). William Fletcher, [No URL available]
Textpipe Pro (1996, 2005). DataMystic Corporation, http://www.datamystic.com/
WMatrix (2001). Paul Rayson, www.comp.lancs.ac.uk/ucrel/wmatrix/

Authors' addresses

Paul Thompson
Department of Applied Linguistics
The University of Reading
P.O. Box 241
Reading RG6 6AA
U.K.

p.a.thompson@reading.ac.uk

Alison Sealey
English Department
University of Birmingham
Edgbaston
Birmingham B15 2TT
U.K.

a.j.sealey@bham.ac.uk