CATEGORY AMBIGUOUS WORDS COMPARED ACROSS THREE TYPES OF CORPORA

Elaine W. Vine: *School of Linguistics and Applied Language Studies, Victoria University of Wellington <elaine.vine@vuw.ac.nz>*

Abstract

This study draws on three types of corpora: general English corpora, learner English corpora and a corpus of English language teaching coursebooks, in an investigation of high frequency words which are category ambiguous, that is, one word-form has different grammatical uses. Four words are analysed and discussed: *about*, *over*, *end* and *place*. Variation in patterns of use of these words is shown within and across the corpora, which gives rise to discussion of pedagogical implications. Patterns of use in coursebooks often differ from patterns of use in general English, and it is argued that while corpus frequency data should not necessarily be pedagogically prescriptive, they should inform pedagogy except where there is good reason for them not to do so.

1. Introduction

Over the past decade, there has been significant progress in the development of 1) general English corpora (e.g. O'Keeffe et al. 2007) and 2) learner corpora (e.g. Granger 2009) and their application to English language teaching. There has also been study of English language use in 3) coursebook corpora (e.g. Meunier & Gouverneur 2009). This work has built on the insights of, for

 © Linguistic Society of New Zealand (Inc.)

example, Kennedy (1987) who argued in his study of how quantification is expressed in written English that corpus-based information about patterns of use 'surely is an improvement on using impressionistic means alone as a basis for English language teaching' (p. 283).

Many words in English, particularly words that are used frequently, are multifunctional. As Leung (1991) noted in his study of the meanings of *over*, this poses challenges to learners of English, and to their teachers, who have to decide when and how to teach the words and their various uses. However, Leung's study was not corpus-based, so he could not comment on how frequently each of the meanings of *over* that he identified is used in English.

If, in English language teaching contexts, we are to apply Kennedy's suggestion about corpus-based information and Leung's insight about multifunctional words, it is relevant to first explore patterns of use across different types of corpora. In this paper, I draw on the three types of corpora mentioned above to focus on category ambiguity in high frequency words such as *over*, that is, where one word-form has different grammatical uses, for example, *over* as adverb (i) and *over* as preposition (ii):

i. i'll come **over** and stay a few days (dpc192)

ii. there is a row in the government **over** the report (msn062)
(examples from *Wellington Corpus of Spoken New Zealand English*).

Other main types of lexical ambiguity do not involve grammatical differences: for so-called 'within-category homonymy', unrelated words share the same word-form, for example, *bat* as a flying mammal and as a piece of sports equipment, and for polysemy a single word has related senses, such as *point* as the tip of a blade or as a promontory.

My research questions are:

1. What are the patterns and frequencies of use of high frequency category-ambiguous words in general English corpora, learner corpora and English language coursebooks?

2. To what extent are the patterns and frequencies of use of these words in general and learner English corpora reflected in pedagogical applications in coursebooks?

In this paper, I have chosen to focus on *over* first, to provide a corpus-based perspective on that word which was missing from Leung's (1991) study. I have

also chosen three other words to compare with *over*: *about*, because, like *over*, it has adverb (iii) and preposition (iv) uses:

iii. i stayed there for **about** an hour afterwards (dpc036)

iv. quentin was thinking **about** becoming a d j (dpf024)

The other two words are *end* and *place*. These were chosen to contrast with *over* and *about* in that, while teachers regard *over* and *about* as 'function words', they regard *end* and *place* as 'content words'. *End* and *place* have both noun (v and vii) and lexical verb (vi and viii) uses:

v. you did all walk out at the **end** just as I was trying to get one question resolved (dpc308)

vi. i suppose it's quite hard to **end** a movie like that though (dpc273)

vii. it's a nice **place** for a couple (dpc008)

viii. that's the only er view i could **place** on it (dgi148)

All four words are high frequency words in English, as indicated by their frequency rankings in an unlemmatised list of word-forms prepared from the BNC for use by English learners (Audience Dialogue 2006): *about* (55th), *over* (74th), *end* (169th) and *place* (173rd).

I report here on analyses and comparisons of patterns of use of these four category ambiguous words in general English corpora, learner corpora, and a coursebook corpus.

2. Comparison corpora

The comparison corpora for this study are:

General English
British English (BE)
Spoken BE – spoken section of the British National Corpus
Written BE – written section of the British National Corpus

The British National Corpus (BNC) was completed in 1994. The spoken section contains *c.* 10 million words and the written section *c.* 90 million words. For this study the BNC was searched using the on-line interface *BYU-*

BNC: The British National Corpus (Davies 2004). All other corpora were searched using *WordSmith Tools* (Scott 2004).

New Zealand English (NZE)
Spoken NZE – Wellington Corpus of Spoken New Zealand English
Written NZE – Wellington Corpus of Written New Zealand English

The two Wellington corpora contain 1 million words each. The spoken corpus is made up of texts collected between 1988 and 1994 (Holmes *et al.* 1998) and the written corpus is made up of writings published between 1986 and 1990 (Bauer 1993).

L1 Learner English (L1 LE)
Spoken L1 LE – Louvain Corpus of Native English Conversation (LOCNEC) – learner turns only

The learner turns section of LOCNEC contains 124,352 words. It is composed of interviews with British university undergraduate students. It is a comparison corpus for LINDSEI (see below) and has the same format and tasks.

Written L1 LE – Louvain Corpus of Native English Essays (LOCNESS)

The LOCNESS corpus contains 288,177 words: 149,574 from argumentative essays by American university students, 18,826 from literary-mixed essays by American university students, 59,568 from argumentative and literary essays by British university students, and 60,209 from British A-level argumentative essays. It is a comparison corpus for ICLEv2 (see below).

L2 Learner English (L2 LE)
Spoken L2 LE – Louvain International Database of Spoken English Interlanguage – learner turns only (LINDSEI: Gilquin *et al.* 2010)

LINDSEI is a 1 million word corpus of 554 interviews with undergraduate university students who are higher intermediate to advanced EFL learners, most from countries in Europe, with some from China and Japan. The learner turns within the interviews total just under 800,000 words.

Written L2 LE – International Corpus of Learner English (ICLEv2: Granger *et al.* 2009)

ICLEv2 is a 3.75 million word corpus of undergraduate university writing (mostly argumentative essays) by higher intermediate to advanced EFL

learners, most from countries in Europe, with some from China, Japan and South Africa.

English language teaching coursebook series

Cutting Edge (CE)

Starter (1) (Cunningham, Redston & Moor 2007), Elementary (2) (Cunningham, Moor & Eales 2005), Pre-Intermediate (3) (Cunningham, Moor & Carr 2005), Intermediate (4) (Cunningham & Moor 2005), Upper Intermediate (5) (Cunningham & Moor 2007), Advanced (6) (Cunningham, Moor & Carr 2007)

New Headway (NH)

Beginner (1) (Soars & Soars 2002), Elementary (2) (Soars *et al.* 2000), Pre-Intermediate (3) (Soars *et al.* 2007), Intermediate (4) (Soars & Soars 2003b), Upper-Intermediate (5) (Soars & Soars 2005), Advanced (6) (Soars & Soars 2003a)

The coursebook corpus comprises the full texts of all student books in the two series of coursebooks, including contents pages; activities and exercises; written texts; transcripts of spoken texts, words on pictures, diagrams, charts; answer keys. The two series are widely used with adult learners of English, ranging from beginners through to advanced learners.

Table 1 shows the number of words in each book of each of the two coursebook series. Some differences can be seen between books at each level in the two series, but the total number of words for each series is similar.

Table 1: Coursebook corpus: CE and NH series

CUTTING EDGE (CE)	NO. OF WORDS	NEW HEADWAY (NH)	NO. OF WORDS
CE 1	35,209	NH 1	33,773
CE 2	70,102	NH 2	57,443
CE 3	81,231	NH 3	79,318
CE 4	90,306	NH 4	78,537
CE 5	96,435	NH 5	99,458
CE 6	88,097	NH 6	96,510
Total	461,380		445,039

While the comparison corpora vary considerably in size, even the smaller ones are adequate for the purposes of this study given that the four target words are all high frequency words.

3. Comparison findings

All of the analyses presented in this paper relate to exact forms of the relevant words (e.g. *end*), not inflected forms (e.g. *ends*, *ended*, *ending*) because some of the inflected forms (e.g. *ended*) are not category ambiguous. In the following tables and figures, frequency data are presented in standardised form per 10,000 words. However, the log-likelihood (LL) statistics were calculated from the raw data, using the online log-likelihood calculator at http://ucrel.lancs.ac.uk/llwizard.html which was developed by the University Centre for Computer Corpus Research on Language at Lancaster University.

3.1 General English corpora

As shown in Table 2, there are statistically significant differences between the two spoken corpora for all four words, but *about*, *over* and *place* occur more frequently in NZE than in BE, while *end* occurs more frequently in BE than NZE.

Table 2: Spoken BE and NZE: occurrences per 10k words

	about	*over*	*end*	*place*
BE – Spoken	36	10	5.4	3.9
NZE – Spoken	40	13	4.6	5.6
	LL 46.96 p<0.0001	LL 64.86 p<0.0001	LL 10.14 p<0.01	LL 61.90 p<0.0001

As shown in Table 3, in the written corpora, three words, *about*, *over* and *place*, occur more frequently in NZE than in BE, and one word, *end*, occurs more frequently in BE than in NZE. but the differences are only statistically significant for *about*, *over* and *end*.

Table 3: Written BE and NZE: occurrences per 10k words

	about	*over*	*end*	*place*
BE – Written	18	14	4.8	4.9
NZE – Written	20	15	4.3	5.2
	LL 38.17 p<0.0001	LL 5.20 p<0.05	LL 8.10 p<0.01	LL 0.54 ns

3.2 Learner English corpora

3.2.1 L1 learner English corpora

The data for L1 learner English (L1 LE) (see Table 4) show that *about*, *end* and *place* occur more frequently in spoken L1 LE than in written L1 LE. By contrast, *over* occurs at the same frequency in both corpora.

Table 4: Spoken and written L1 LE: occurrences per 10k words

	about	*over*	*end*	*place*
L1 LE – Spoken	40	10	9.4	6.8
L1 LE – Written	18	10	6.3	5.0
	LL 153.96 p<0.0001	LL 0.00 ns	LL 10.55 p<0.01	LL 4.66 p<0.05

3.2.2 L2 learner English corpora

The data for L2 learner English (L2 LE) (see Table 5) show that *about* occurs more frequently in spoken L2 LE than in written L2 LE. By contrast, the other three words occur more frequently in written L2 LE than in spoken L2 LE, though the difference for *end* is not significant.

Table 5: Spoken and written L2 LE: occurrences per 10k words

	about	*over*	*end*	*place*
L2 LE – Spoken	37	3.6	5.1	6.4
L2 LE – Written	27	6.3	5.2	8.0
	LL 210.91 p<0.0001	LL 89.70 p<0.0001	LL 0.01 ns	LL 23.92 p<0.0001

3.3 L1 and L2 learner English corpora compared

3.3.1 Spoken L1 and L2 learner English corpora

In the parallel learner spoken corpora, L1 learners use all four words more frequently than L2 learners, but the differences are statistically significant only for *over* and *end* (see Table 6).

Table 6: Spoken L1 LE and L2 LE: occurrences per 10k words

	about	*over*	*end*	*place*
Spoken – L1 LE	40	10	9.4	6.8
Spoken – L2 LE	37	3.6	5.1	6.4
	LL 1.97 ns	LL 78.76 p<0.0001	LL 28.16 p<0.0001	LL 0.20 ns

3.3.2 Written L1 and L2 learner English corpora

The pattern is different for the parallel learner written corpora. Here, L1 learners use *over* and *end* more frequently than L2 learners, but the reverse is the case for *about* and *place*. In all four cases the differences are statistically significant (see Table 7).

Table 7: Written L1 LE and L2 LE: occurrences per 10k words

	about	*over*	*end*	*place*
Written – L1 LE	18	10	6.3	5.0
Written – L2 LE	27	6.3	5.2	8.0
	LL 102.30 p<0.0001	LL 63.21 p<0.0001	LL 7.39 p<0.01	LL 39.94 p<0.0001

3.4 Coursebook corpora

One of the target words, *about*, occurs significantly more frequently in the CE series of coursebooks than in the NH series (see Table 8). The differences for the other three words are not statistically significant.

There are also differences between the two series on a book-by-book basis (see below).

Table 8: Target words in CE and NH series per 10k words

	about	*over*	*end*	*place*
CE	52	8.2	4.4	6.3
NH	44	8.5	5.0	5.8
	LL 32.30 p<0.0001	LL 0.21 ns	LL 1.59 ns	LL 1.05 ns

3.5 General English, L2 and L1 learner English and coursebook corpora compared

This section includes calculations of frequencies of use of the target words according to the grammatical category of use of each occurrence of the word. For the BNC, the standardised data were calculated by BYU-BNC (Davies 2004) on the basis of the automatic part-of-speech tagging in the corpus. For all other corpora, the standardised data for grammatical categories were calculated on the basis of frequencies extrapolated from a manual analysis of a random sample, generated by *WordSmith Tools*, of 100 occurrences of the target word from throughout each corpus. A sample size of 100 occurrences from each corpus was chosen to be large enough to provide sufficient evidence of different grammatical uses of the target word, while also being manageable for manual analysis. Significance of findings has been calculated in SPSS using Chi square for comparing categorical data, assuming expected equal frequencies.

3.5.1 about

In the case of *about* (see Figure 1), the overall pattern noted earlier that *about* occurs more frequently in CE than in NH (see Table 8) is replicated in each of the six books in the series. There is also a trend in both series for the frequency of *about* to decrease across the series, though that trend is not evenly spread across the series.

However, Figure 1 also shows that *about* is used more frequently in coursebooks than in spoken English from all corpora, and much more frequently than in written English from all corpora. Also, written L2 learner English stands out in that the frequency of *about* does not sit with the other written Englishes. It approaches but does not reach the spoken levels of the other corpora.

When we consider uses of *about* in terms of the grammatical category of

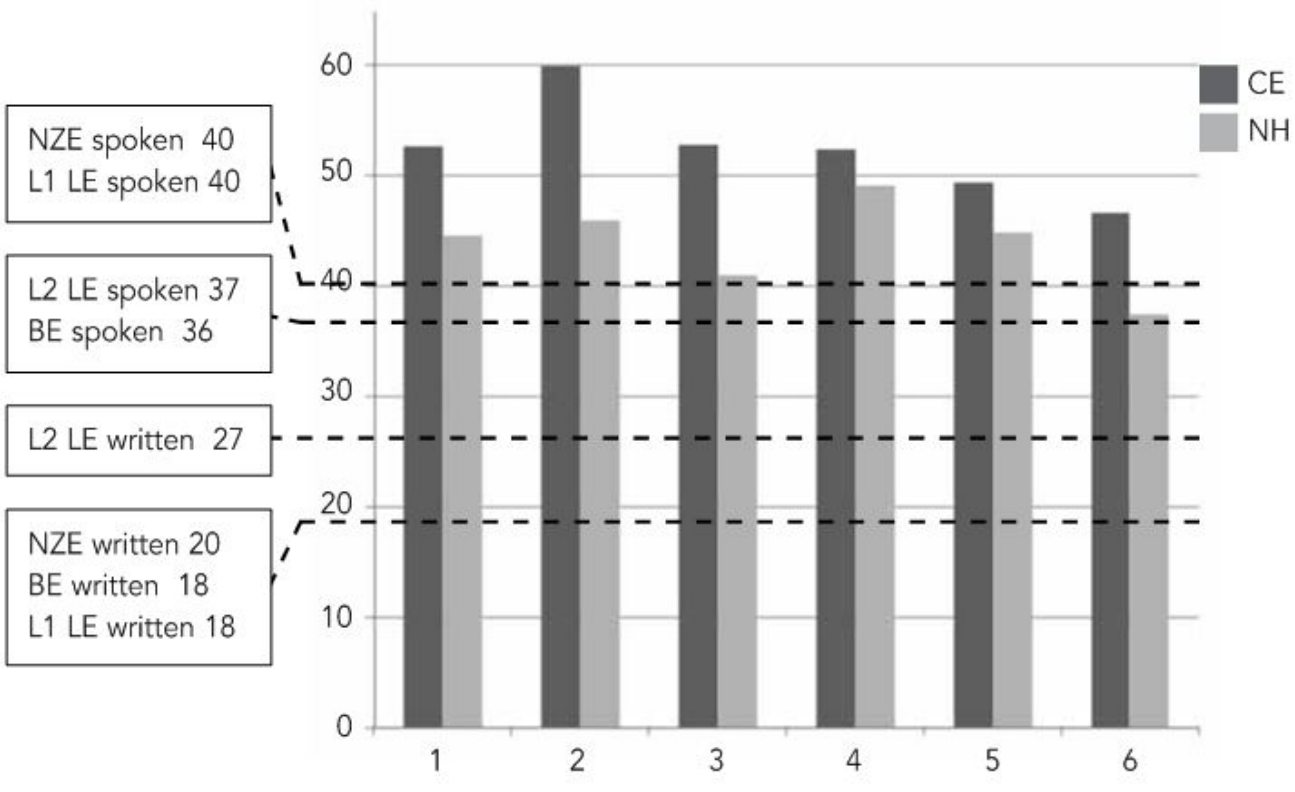


Figure 1: *About* in CE and NH coursebooks and other corpora per 10k words

each use (see Table 9), we see that *about* occurs mostly as a preposition in all the corpora, and much less frequently as an adverb. The differences between preposition and adverb in each corpus are all statistically significant at the level of $p<0.0001$.

The following are examples of use in each grammatical category:

(1) Preposition: Write questions ***about*** the story using these question words. (NH series)

(2) Adverb: i think the minimum wage is ***about*** six dollars an hour or something that he can get (NZE spoken).

The two coursebook series use *about* as a preposition more frequently than L2 learners do, and both coursebooks and L2 learners use *about* more frequently as a preposition than BE, NZE and L1 learner English speakers and writers. There is no such clear pattern for uses of *about* as an adverb. Notice, for example, that the two series have rather different rates of use of *about* as an adverb.

Table 9: Part-of-speech in corpora per 10k words – *about*

	PREPOSITION	PREP *vs* ADV	ADVERB
BE spoken	27	χ^2 = 9466.328, df = 1, p<0.0001	8.7
BE written	14	χ^2 = 44546.643, df = 1, p<0.0001	4.1
NZE spoken	29	χ^2 = 767.206, df = 1, p<0.0001	12
NZE written	15	χ^2 = 457.508, df = 1, p<0.0001	5.4
L1 LE spoken	24	χ^2 = 25.138, df = 1, p<0.0001	15
L1 LE written	14	χ^2 = 185.470, df = 1, p<0.0001	3.8
L2 LE spoken	32	χ^2 = 1688.538, df = 1, p<0.0001	4.4
L2 LE written	24	χ^2 = 6182.201, df = 1, p<0.0001	2.9
NH series	39	χ^2 = 1108.864, df = 1, p<0.0001	4.8
CE series	41	χ^2 = 828.840, df = 1, p<0.0001	10

3.5.2 over

In the case of *over* (see Figure 2), the overall pattern noted earlier that there is no significant difference between the frequency of *over* in the two series, CE and NH (see Table 8) is evident also in the six books in the series. There is a trend in both series for the frequency of *over* to increase across the series, though that trend is not evenly spread across the series.

However, Figure 2 also shows that *over* is used less frequently in coursebooks than in all corpora except the L2 learner English written and spoken corpora. In other words, general English users and L1 learners use *over* more frequently than L2 learners, and coursebook English patterns are closer to the patterns for L2 learners than to those for L1 users of English both general and learner. This is unexpected in that we might expect coursebooks to use high frequency words even more frequently, if anything, than L1 users of the language, in order to make their texts more accessible to L2 learners.

When we consider uses of *over* in terms of the grammatical category of each use (see Table 10), we see that *over* occurs more frequently as a preposition than as an adverb in all the corpora except spoken NZE and spoken L2 learner English. The differences between preposition and adverb are all statistically significant at the level of p<0.0001, except for the CE and

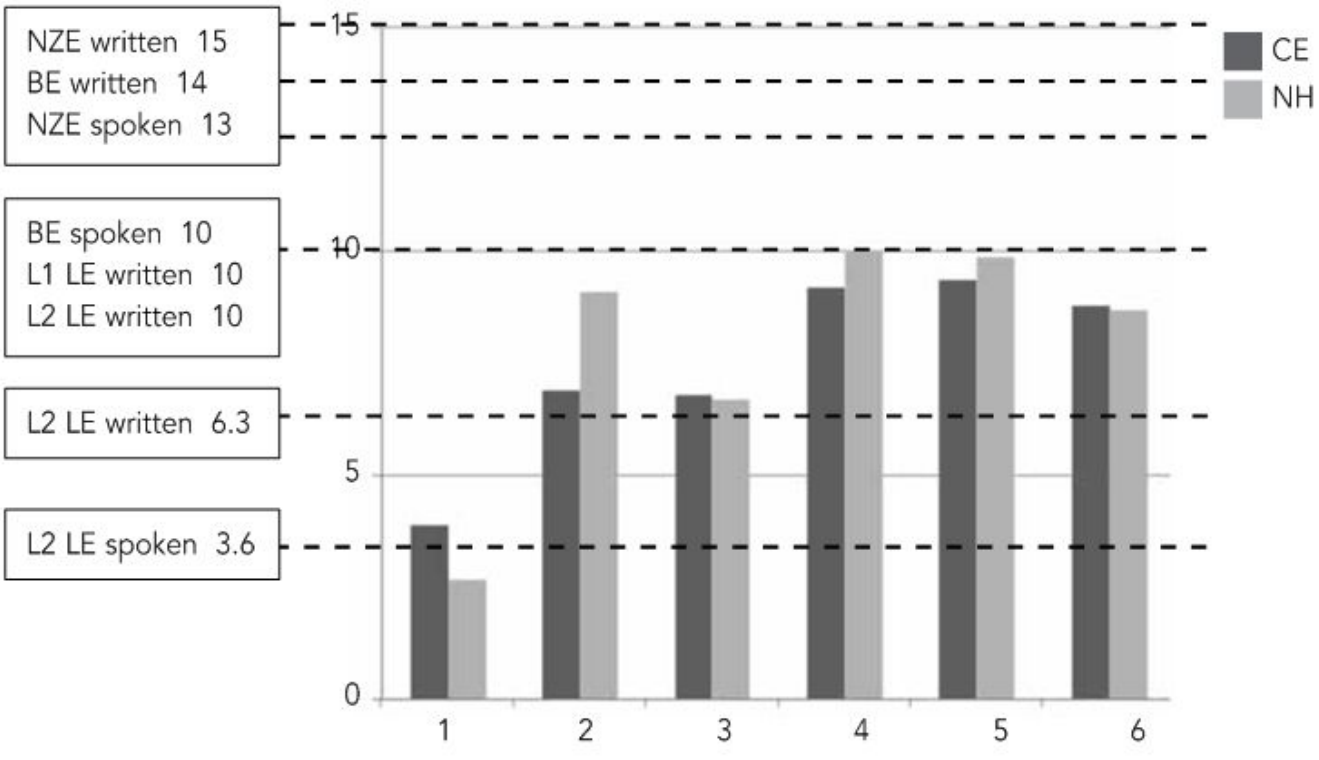


Figure 2: *Over* in CE and NH coursebooks and other corpora per 10k words

NH coursebook series, where the differences are still significant but at lower levels of confidence. Uses of *over* as adjective and noun also occur in some of the corpora, but only at very low frequencies. The following are examples of use in each grammatical category:

(1) Preposition: You definitely shouldn't reveal the details of your bank account ***over*** the telephone. (CE series)

(2) Adverb: it was just when they all came ***over*** on the horses and it was just like oh wow this is a good film (L1 LE spoken)

(3) Adjective: oh God my university life is ***over*** (L2 LE spoken)

(4) Noun: At the start of the final ***over***, England wanted five to win. (BE written)

The difference between preposition and adverb uses is much less marked for *over* than was the case for *about*; preposition uses of *over* are only slightly more frequent than adverb uses across the corpora. However, it is noteworthy that while preposition uses of *over* are consistently higher in the written corpora than the spoken, the reverse was the case for preposition uses of *about*.

Table 10: Part-of-speech in corpora per 10k words – *over*

	PREPOSITION	PREP *vs* ADV	ADVERB	ADJECTIVE	NOUN
BE spoken	5.6	χ^2 = 103.789, df=1, p<0.0001	4.6	0.03	<0.01
BE written	7.5	χ^2 = 1609.719, df=1, p<0.0001	5.9	0.02	0.02
NZE spoken	4.8	χ^2 = 76.849, df=1, p<0.0001	7.8	0.13	0.26
NZE written	9.3	χ^2 = 105.947, df=1, p<0.0001	5.3	0.61	0
L1 LE spoken	7.4	χ^2 = 23.902, df=1, p<0.0001	2.9	0	0
L1 LE written	8.7	χ^2 = 154.742, df=1, p<0.0001	1.7	0	0
L2 LE spoken	1.2	χ^2 = 25.610, df=1, p<0.0001	2.2	0.2	0
L2 LE written	3.8	χ^2 = 160.251, df=1, p<0.0001	2.2	0.03	0.02
NH series	4.6	χ^2 = 5.391, df=1, p= 0.0202	3.6	0.3	0
CE series	4.3	χ^2 = 6.260, df=1, p= 0.0123	3.3	0.2	0

3.5.3 end

In the case of *end* (see Figure 3), the overall pattern noted earlier that there is no significant difference between the frequency of *end* in the two series, CE and NH (see Table 8) is evident also in the six books in the series. There is a trend in both series for the frequency of *end* to increase across the series.

However, Figure 3 also shows that *end* is used less frequently in the first three coursebooks (but not NH Book 2) in both series and CE Book 4 than in all other corpora. In NH Book 2 occurrences of *end* reach the level of NZE written and spoken. In CE Book 5, occurrences of *end* reach the level of BE written, and in NH Book 4 they reach the level of L2 LE spoken and written. NH Books 5 and 6 and CE Book 6 reach the level of BE spoken and L1 LE written, but they are still well below the level of L1 LE spoken. These patterns suggest that coursebook usage builds towards L1 English usage.

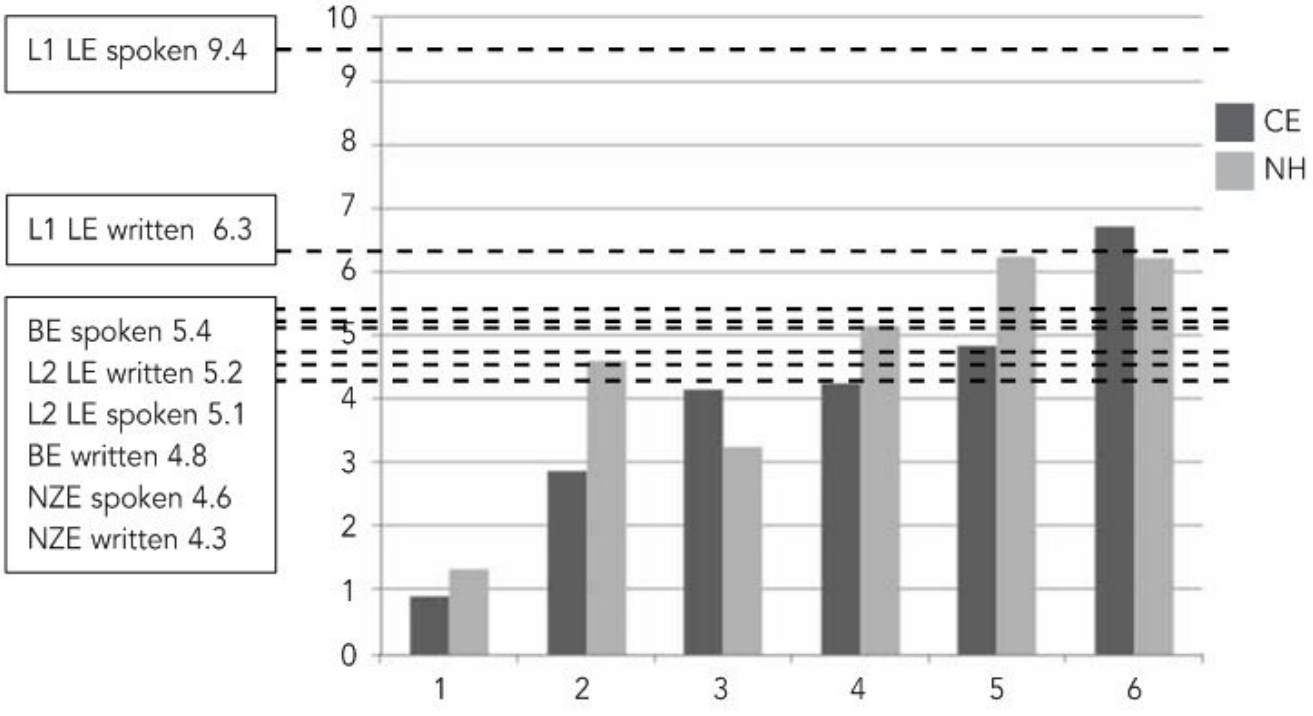


Figure 3: *End* in CE and NH coursebooks and other corpora per 10k word

Table 11: Part-of-speech in corpora per 10k words – *end*

	NOUN	NOUN *vs* VERB	VERB
BE spoken	4.7	χ^2 = 2985.196,df = 1, p<0.0001	0.7
BE written	4.2	χ^2 = 24729.031,df = 1, p<0.0001	0.5
NZE spoken	3.9	χ^2 = 217.815,df = 1, p<0.0001	0.8
NZE written	3.9	χ^2 = 298.906,df = 1, p<0.0001	0.3
L1 LE spoken	8.6	χ^2 = 77.919,df = 1, p<0.0001	0.8
L1 LE written	4.8	χ^2 = 53.020,df = 1, p<0.0001	1.6
L2 LE spoken	4.7	χ^2 = 271.488,df = 1, p<0.0001	0.5
L2 LE written	3.7	χ^2 = 342.250,df = 1, p<0.0001	1.5
NH series	3.2	χ^2 = 28.922,df = 1, p<0.0001	1.4
CE series	3.3	χ^2 = 52.083,df = 1, p<0.0001	1.0

When we consider uses of *end* in terms of the grammatical category of each use (see Table 11), we see that *end* occurs more frequently as a noun than as a verb in all the corpora. The differences between noun and verb are all statistically significant at the level of $p<0.0001$. The following are examples of use in each grammatical category:

(1) Noun: Just got to tie a bit of wood on the ***end*** of that bit where it's rotted. (BE spoken)

(2) Verb: The wood is split, and it isn't hard to ***end*** up with a splinter in your hand (NZE written)

Spoken L1 learner English stands out as having much more frequent occurrences of *end* as a noun than any of the other corpora, though this is not the case for verb uses. It may be the case that in this smaller corpus, for this word, topic has had an influence, in that the noun uses include many that relate to story plots, and an optional task in the interviews with learners was to talk about a film or play, and a required task was to describe four pictures making up a story (Gilquin et al. 2010: 8).

3.5.4 place

In the case of *place* (see Figure 4), the overall pattern noted earlier that there is no significant difference between the frequency of *place* in the two series, CE and NH (see Table 8), obscures some significant differences between books at particular levels. CE Book 1 has significantly ($LL = 12.62$, $p<0.001$) more occurrences of *place* than NH Book 1, as does CE Book 4 than NH Book 4 ($LL = 5.26$, $p<0.05$), while NH Book 6 has significantly ($LL = 4.37$, $p<0.05$) more occurrences of *place* than CE Book 6. There is a trend in the NH series for the frequency of *place* to increase across the series, while in the CE series the frequencies go up and down from book to book across the series.

However, Figure 4 also shows that, apart from NH Book 1, coursebook uses of *place* fall within much the same range as the L1 corpora, from BE spoken at 3.9 occurrences per 10,000 up to L1 learner English spoken at 6.8 occurrences. L2 learner English, both spoken at 6.4 and written at 8.0, fall towards or above the coursebook range. This is the only one of the four words which L2 learners have used more than L1 speakers of all sorts, and more than the coursebooks. It is possible that the learner corpus tasks/topics may have contributed to its higher use in the L2 corpora, but if that is the case, then it is odd that the same tasks in the learner written corpora have given different outcomes for L2 and

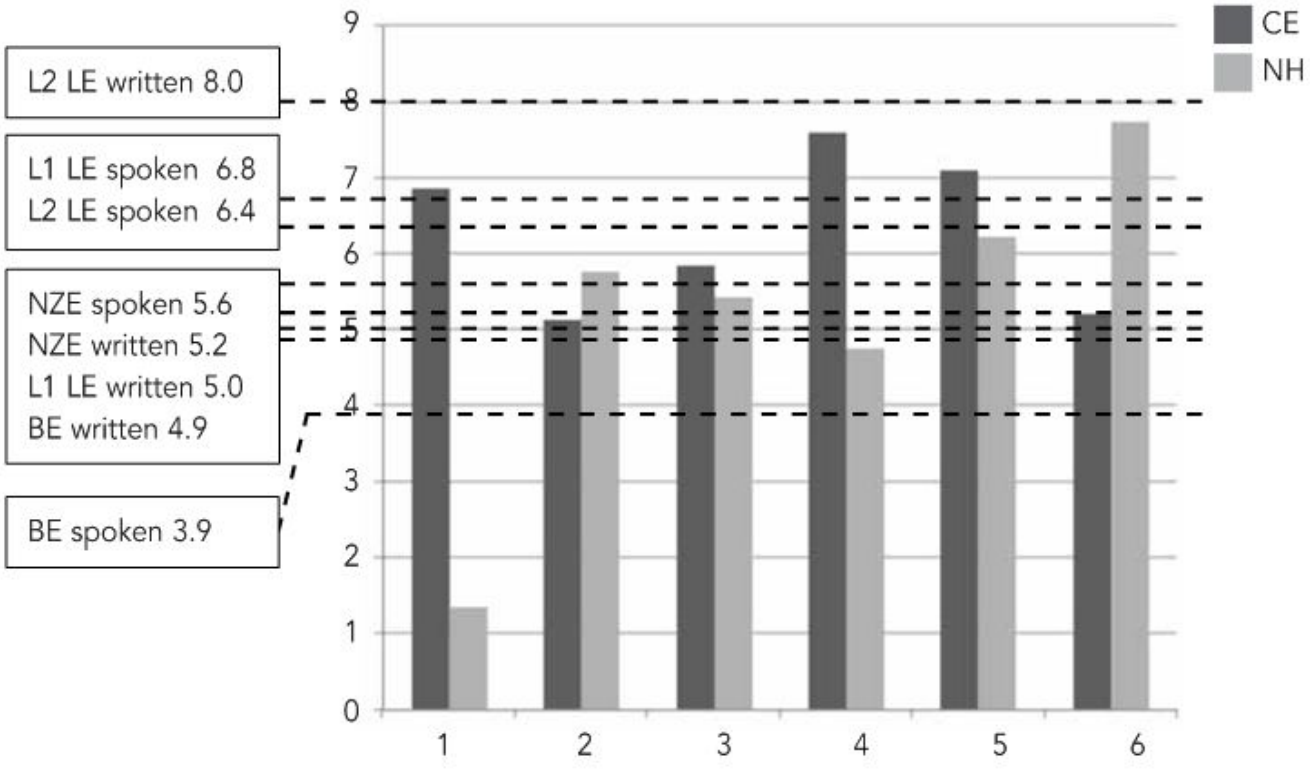


Figure 4: *Place* in CE and NH coursebooks and other corpora per 10k word

Table 12: Part-of-speech in corpora per 10k words – *place*

	NOUN	NOUN *vs* VERB	VERB
BE spoken	3.6	χ^2 = 2795.315, df = 1, p<0.0001	0.3
BE written	4.5	χ^2 = 27979.803, df = 1, p<0.0001	0.5
NZE spoken	5.4	χ^2 = 553.958, df = 1, p<0.0001	0.1
NZE written	4.9	χ^2 = 417.240, df = 1, p<0.0001	0.3
L1 LE spoken	6.8	χ^2 = 80.000, df = 1, p<0.0001	0
L1 LE written	4.7	χ^2 = 123.484, df = 1, p<0.0001	0.3
L2 LE spoken	6.3	χ^2 = 487.197, df = 1, p<0.0001	0.06
L2 LE written	8.0	χ^2 = 3019.000, df = 1, p<0.0001	0
NH series	5.5	χ^2 = 200.674, df = 1, p<0.0001	0.2
CE series	6.1	χ^2 = 247.918, df = 1, p<0.0001	0.2

L1 learners. The unexpectedly high comparative frequency of *place* may also have to do with its usefulness as a 'general noun' (Mahlberg 2005).

When we consider uses of *place* in terms of the grammatical category of each use (see Table 12), we see that *place* occurs more frequently as a noun than as a verb in all the corpora. The differences between noun and verb are all statistically significant at the level of $p<0.0001$. The following are examples of use in each grammatical category:

(1) Noun: It's also an extremely dangerous ***place***, and every year we have several cases where people, usually tourists have serious problems in the desert and have to be rescued. (CE series)

(2) Verb: Please ***place*** your tray tables in their fully upright and locked positions. (NH series)

There is no consistent pattern to the use of *place* as a noun in spoken and written corpora. BE and L2 LE show higher frequency in written than spoken use, while NZE and L1 LE show higher frequency in spoken than written use.

4. Pedagogical implications

We have seen that there is variation in the extent to which the frequencies of occurrence of different high frequency category-ambiguous words in general and learner English corpora are consistent with each other and with pedagogical applications in coursebooks. Such variation raises issues about common pedagogical practices.

Corpus frequency data should not necessarily be pedagogically prescriptive, but they should inform pedagogy. If coursebook writers and language teachers are well informed about issues raised by corpus comparison findings such as those presented here, they may choose not to be guided by them, but in that case they will do so in a principled way. For example, the findings of this study show that *about* is consistently used more in spoken than in written contexts, which suggests that pedagogical attention should follow that pattern. However, there may be a particular genre or topic in a written context that requires more attention to *about*. An example could be written fiction contexts where written dialogue resembles spoken usage.

Teachers need to consider what is going on where, for example, coursebook frequencies are lower than general English frequencies (e.g. *end*, particularly

as a noun, and *over*, particularly as a preposition) or higher (e.g. *about*, particularly as a preposition). Another example is where, for both 'content' (*end, place*) and 'function' (*about, over*) words, L1 learners are using these high frequency words more frequently than L2 learners in parallel corpora.

On the whole, it appears that the category ambiguity of these high frequency words is paid little explicit attention in the coursebook series. For example, in NH Book 2, *end* is included in a word list at the end of the book, and it is labeled as both noun and verb. However, there is no explicit attention paid to its category ambiguity in the relevant unit of work. It is simply the case that it occurs in both uses, noun and verb:

On 2 July she was nearly at the **end** of her journey, …

When did the Second World War begin / **end**?

Corpus comparison data suggest paying more pedagogical attention to the category ambiguity of high frequency words. The data also give language educators grounds for making reasoned choices about which category-ambiguous words and which uses of those words might warrant more, or less, pedagogical time and emphasis.

References

Audience Dialogue. 2006. Sustainable English - A vocabulary (Dennis List – 6 December 2006). http://www.audiencedialogue.net/susteng.html (last accessed on 20 June 2012).

Bauer, Laurie. 1993. *Manual of Information to Accompany The Wellington Corpus of Written New Zealand English*. Wellington, NZ: Department of Linguistics, Victoria University of Wellington.

Cunningham, Sarah and Peter Moor. 2005. *New Cutting Edge Intermediate Students' Book.* Harlow: Pearson Longman.

Cunningham, Sarah and Peter Moor. 2007. New *Cutting Edge Upper Intermediate Students' Book.* Harlow: Pearson Longman.

Cunningham, Sarah, Peter Moor and Jane Comyns-Carr. 2005. New *Cutting Edge Pre-Intermediate Students' Book.* Harlow: Pearson Longman.

Cunningham, Sarah, Peter Moor and Jane Comyns-Carr. 2007. *Cutting Edge Advanced Students' Book.* Harlow: Pearson Longman.

Cunningham, Sarah, Peter Moor and Frances Eales. 2005. *New Cutting Edge Elementary Students' Book.* Harlow: Pearson Longman.

Cunningham, Sarah, Chris Redston and Peter Moor. 2007. *Cutting Edge Starter Students' Book.* Harlow: Pearson Longman.

Davies, Mark. 2004. *BYU-BNC: The British National Corpus.* Available online at http://corpus.byu.edu/bnc

Gilquin, Gaëtanelle, Sylvie De Cock and & Sylvian Granger (eds). 2010. *Louvain International Database of Spoken English Interlanguage (LINDSEI)*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, Sylvian. 2009. 'The contribution of learner corpora to Second Language Acquisition and foreign language teaching: A critical evaluation'. In Karin Aijmer (ed.) *Corpora and Language Teaching*. Amsterdam: John Benjamins. 13–33.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier and Magali Paquot (eds). 2009. *International Corpus of Learner English, Version 2*. Louvain-la-Neuve: Presses universitaires de Louvain.

Holmes, Janet, Bernadette Vine and Gary Johnson. 1998. *Guide to the Wellington Corpus of Spoken New Zealand English*. Wellington, NZ: School of Linguistics and Applied Language Studies, Victoria University of Wellington.

Kennedy, Graeme, D. 1987. Quantification and the use of English: A case study of one aspect of the learner's task. *Applied Linguistics* 8, 264–286.

Leung, Matthew Wing Kwong. 1991. Prototype theory and teaching English prepositions. *Perspectives (City University of Hong Kong)* 3, 89–97.

Mahlberg, Michaela. 2005. *English General Nouns. A Corpus Theoretical Approach*. Amsterdam: John Benjamins.

Meunier, Fanny and Céline Gouverneur. 2009. New types of corpora for new educational challenges: Collecting, annotating and exploiting a corpus of textbook material. In K. Aijmer (ed.) *Corpora and Language Teaching*. Amsterdam: John Benjamins, 179–201.

O'Keeffe, Anne, Michael McCarthy and Ronald Carter. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Scott, Mike. 2004. *WordSmith Tools version 4*. Oxford: Oxford University Press.

Soars, Liz and John Soars. 2002. *New Headway Beginner Student's Book*. Oxford: Oxford University Press.

Soars, Liz and John Soars. 2003a. *New Headway Advanced Student's Book*. Oxford: Oxford University Press.

Soars, Liz and John Soars. 2003b. *New Headway Intermediate Student's Book (3rd ed.)*. Oxford: Oxford University Press.

Soars, Liz and John Soars. 2005. *New Headway Upper-Intermediate Student's Book (3rd ed.)*. Oxford: Oxford University Press.

Soars, Liz, John Soars and Sylvia Wheeldon. 2000. *New Headway Elementary Student's Book*. Oxford: Oxford University Press.

Soars, Liz, John Soars and Sylvia Wheeldon. 2007. *New Headway Pre-Intermediate Student's Book (3rd ed.)*. Oxford: Oxford University Press.