The college idiom: Idioms in the COLL Corpus

David C. Minugh
University of Stockholm

Abstract

As with much of vocabulary, idioms in the stricter sense appear to be acquired continually throughout one's lifetime. Since most of the material in current large-scale corpora comes from writers well out of their teens, the 3.7 M word COLL corpus of college student online newspapers from Australia, the British Isles, New Zealand, North America and South Africa (Minugh 2002) provides one of the few already-compiled sources of writing by 20-year-olds, and thus is an interesting starting point for an investigation of which idioms are in use in the writing of university students in the English-speaking world when they address their peers. Using the idioms specified in the Collins COBUILD dictionary of idioms *as our starting point, the COLL corpus will be examined for use of idioms. Specific questions to investigate include which idioms occur, their geographic and subgenre distribution, their positions in the texts and their textual functions. Idiom-breaking, i.e. playful variation, may also be expected to occur in this particular genre, and the corpus can provide an indication of how prevalent this is, as well.*

1 Introduction

Since the advent of the telephone (not to mention the cell phone), young adults have had less and less incentive to write in non-educational environments. At the same time, they are the speakers who will be producing much of the written language of the future, so that how they write should be of major interest. However, most of the material in large-scale corpora such as the (written component of the) *BNC* comes from writers whose teenage years are a rapidly receding memory, and most of the relatively easily obtained material from younger writers takes the form of examinations or term papers, in a (secondary or tertiary) educational environment.

If we assume that, as for much of vocabulary, idioms are acquired continually throughout one's lifetime, the writings of young adults can form an interesting starting point for an investigation of idioms, as it may indicate which idioms are learned relatively early, or are coming into the language.[1] A range of different types of writing would of course be desirable, particularly since various informal idioms may not appear in the rather formal types of writing that examinations and term papers consist of.

It is also clear that, given the relative scarcity of individual idioms, unusually large samples are necessary: the *Collins* COBUILD *dictionary of idioms* (henceforth *CCDI*) notes that "[n]early one third of the idioms in [their] dictionary occur less often than once per 10 million words of the [*Bank of English*]" (2002: vi), figures confirmed elsewhere (Moon 1998: 45; Minugh 1999); cf. the discussion in Wray (2002: 25 ff.).

One sufficiently large source of material written by young adults for their peers is the COLL corpus of college student online newspapers (presented in Minugh 2002), the corpus used as the basis for the present investigation.

2 The corpus

The 3.7 M word COLL corpus of college student online newspapers from Australia, the British Isles, South Africa, Canada and the United States provides one of the few already-compiled sources of native-speaker writing by young people in the English-speaking world when they address their peers in writing.[2]

All of the COLL texts were available on the Internet in the spring of 1999, when they were collected, although some were posted somewhat earlier. The issues range from the minimal contribution of 12 words from *The Vermilion* (University of Southwestern Louisiana) to the largest, some 53,370 words from *The Colgate-Maroon* (Colgate College, NY). Some include arts sections with fiction and poems; a few include short police-blotter reports; there are also occasional messages from the dean or university president (see also the discussion in section 7, below). The material available on the Internet on the day the site was visited was normally all included (in 1999, these sites tended to be updated at long intervals or were still somewhat of a trial run, with relatively few being updated on a daily basis).[3] All available college newspaper sites from English-speaking countries were included. Nevertheless, North American newspapers were clearly overrepresented, particularly vis-à-vis the United Kingdom, due to the fact that North America had more enthusiastically embraced the Internet in 1999; the distribution is given as Table 1:

Table 1: Geographical distribution of online college student newspapers in COLL[4]

Country	Pop. (mil)	Target %	Actual %	Total words	No. papers	Avg. wds/paper
US	270	67 %	75 %	2,806,536	215	13,054
Can	25	6 %	12 %	436,456	35	12,470
UK	80	20 %	10 %	368,803	26	14,185
Austr	18	4 %	1 %	48,965	4	12,241
S Afr	2	0 %	1 %	25,963	3	8,654
Irel	4	1 %	1 %	28,552	3	9,517
NZ	3	1 %	1 %	23,671	2	11,836
Tot:	***402***	***100%***	***100%***	***3,738,946***	***288***	***12,982***

It is worth emphasizing that in a corpus of this kind, even if the individual newspapers vary widely in the size of their contribution, the overall result is a cross-section of student publications. Unlike typical newspaper collections, such as the articles in the 1998 *Independent* or *New York Times*, there are no gatekeepers at a higher level in COLL, i.e. no subeditor or house style sheets could have affected more than a tiny fraction of the corpus. Similarly, no individual writer contributed more than a tiny fraction of the material, whereas in corpus material based on major newspapers, certain reporters and editors will contribute articles on a daily basis.

The various newspapers in COLL vary widely in size, as Table 2 indicates:

Table 2: Words per newspaper included in COLL

Lowest	1st quart	Median	Average	3rd quart	Highest
12	6,749	10,860	12,982	16,185	53,370

3 *The idioms*

As used in the present paper, ‘idioms’ will be considered to be the idiomatic phrases listed in *CCDI*, the *Collins* COBUILD *dictionary of idioms* (2002).[5] Singular and plural headwords are not distinguished, so that while the *CCDI* headword **ace** lists *the ace in your hand, {come, be} within an ace of something, have an ace in the hole* and *play your ace*, and the *CCDI* headword **aces** lists *hold all the aces*, their results were combined in the present investigation: the headword **ace** thus contains a total of five idioms, one of which has two main variants. This method of counting yields a total of 1,390 headwords and 3,485 different idioms, a number of them with two or more main variants (which thus are not distinguished).

CCDI specifically claims to cover both British and American idioms, as well as "includ[ing] a few Australian English idioms which our evidence suggests are used more widely now. We have taken a similar approach with other varieties of English" (2002: vii). The coverage of *CCDI* can thus be expected to match that of the COLL corpus. Since COLL is from 1999, there should not be any major time disparity to skew the selection of idioms, either.

4 Finding the COLL idioms

Via WordSmith, the COLL corpus, a series of 284 .txt files, was queried for each headword, normally a noun (see e.g. *CCDI* 2002: viii). As much as possible, WordSmith's wild cards and alternatives were used to restrict the search and remove duplicates, so that e.g. {*land/lands/landed/landing on*, *fall/falls/fell/falling/fallen on*} within three words of {*feet/feet.*} would cover the expected range of variation for *fall on your feet*; *pull someone's leg* would also require a search that included forms of *leg-pull*; every duplicate instance of e.g. *an eye for an eye* had to be eliminated by hand.

The results were then inspected by hand for idiomatic expressions. Only metaphorical uses were counted, so that if someone e.g. *cashed in their chips* at a Reno casino, that instance would not be included unless the text indicated that they had died there.[6] There were occasional difficulties in classifying the exact idiom, as when *CCDI* lists two very similar metaphors, e.g. *get under your skin* (1) and (2), where (1) means "[something] annoys or worries you" and (2) means "[someone or something] begin[s] to affect you in a significant way, so that you become very interested in them or very fond of them" – the corpus context may not be sufficiently detailed to establish which idiom is involved. On the whole, however, even a relatively brief context is sufficient to establish the idiomatic nature of the example. And of course there are the usual difficulties in finding variants of idioms, such as "It's a game of inches, and sometimes I don't want to *go that extra <u>inch</u>*" (*Craccum*, New Zealand), rather than the canonical *go the extra <u>mile</u>*.

5 The idiom groups

A total of 5,439 idiom tokens were found in the COLL corpus. The vast majority were in an essentially canonical form and are therefore called **standard idioms**. They include simple permutations such as singular to plural, as in (1):[7]

(1) We have student agendas, not *hidden agendas*. (*The Daily Aztec*, San Diego State University)

It should be noted that *CCDI*'s information for individual items will occasionally indicate that an item is highly variable. Thus, *CCDI* specifies *sell/go like hot cakes* (2002: 204), but for *put/get your house in order* adds that "verbs such as 'keep' or 'set' can be used instead of 'put' or 'get'" (2002: 205); for *lend someone a hand*, *CCDI* adds that "**A hand** is used in many other structures with a similar meaning" (2002: 172). All variations recognized in *CCDI* (2002: x) are included here.

A much smaller group, 274 in all, are what may be called **variant idioms**, which are recognizably the same idiom, primarily varied via lexical substitution, leaving the meaning basically unchanged, as when *collect dust* or *shut up like a clam* become

(2) leave this one on the shelf with your Ace of Bass and Vanilla Ice CDs to *gather dust*. (*The Courier,* Western Illinois University)

(3) hearing loss caused him to withdraw from his hearing friends. "I just started *to clam up***.** I withdrew from neighborhood activities." (*The Indiana Daily Student*, Indiana University)

These variants are not found in *CCDI*, even if many, such as *gather (dust)* instead of *collect (dust)*, appear to be quite predictable.

The third group, **other idioms**, consists of idioms not found in *CCDI,* but whose headword is found there. They were not systematically searched for, but were observed while looking for standard idioms. Many appear to be traditional idioms that simply were not included in *CCDI*, such as *the long arm of the law*, which is listed in Longman's *Dictionary of contemporary English* (2005) and Oxford's *Advanced learner's dictionary* (2005). Others, such as *Johnny off the pickle boat*, appear to be rooted in more traditional forms, but transformed almost beyond recognition, in this case perhaps from *fresh off the boat* – or possibly merely a way of saying "a nobody":

(4) Mumia: "*The long arm of the law* will not wave its clubs at us." (*The Oberlin Review*, Oberlin College)

(5) "I've gone from Dean of the Mid American Conference coaches to feeling like *Johnny off the pickle boat,*" Walker said. "It's a little intimidating." (*The Digital Collegian*, Penn State University)

They are discussed further in section 7, below. Table 3 summarizes the raw data for idiom occurrence in the COLL corpus, broken down into these three categories, together with background information about the number of idiom types actually found in COLL:

Table 3: The COLL idioms: an overview

	Tot tokens of headwords[8]	Id types in *CCDI*	Id types in COLL	Id tokens in COLL
Totals	407,639	3,485	1,422	5,439
Standard idioms				4,701
Variant idioms				250
Other idioms				488

6 *Idiom density and geographical distribution in COLL*

Given the wide range in size of the various newspapers contributing to COLL, the average values are naturally only a rough approximation, but they indicate that an average newspaper, with 12,982 words, might be expected to contain (5,439 idioms)/(288 newspapers), or about 18.9 idioms, which normalizes to 14,547 idioms per 10 M words. Dividing that by the 3,485 different idiom types listed in *CCDI*, we obtain the average likelihood of an individual idiom occurring in the COLL texts. The result is some 4.20 idioms per 10 M words, which is well in line with the overall figures suggested in *CCDI*, 1st ed. (1995: xvii).[9]

As noted above, 488 'other' idioms were included in the COLL total of 5,439 idioms. Since they are not in *CCDI*, an alternate calculation is to remove them, leaving 4,951 idiom tokens. The average per 10M words then drops to 3.80, again well within the general guidelines suggested in *CCDI*, 1st ed.

If we instead look at the distribution of the newspapers with the highest and lowest densities (Tables 4 and 5, respectively), several points may be noted. First, the high-density newspapers do not exhibit a geographical clustering; i.e. all regions appear to have at least some paper that revels in using idioms, while the low-density newspapers appear to be a North American phenomenon. Secondly, there does not appear to be any correlation between newspaper size and use of idioms.

Table 4: The COLL idioms distributed by newspaper, highest densities (normed per 10,000 words)

Norm	Actual			Words
55.7	16	*Paradigm Press*	Palm Beach, Florida	2,871
40.7	1	*The Cowl*	Providence C, RI	246
36.4	49	*Trinity News*	Trinity U, Irel	13,465
31.9	3	*The Spike*	Bath U, UK	941
30.4	4	*Denver Advocate*	U of Colorado	1,316
29.8	33	*Shout*	U of Liverpool	11,075
29.1	50	*The Orange Source*	Syracuse, NY	17,185
28.8	12	*Jumbunna*	U of W Sydney	4,171
28.5	134	*Edinburgh Student*	Scotland	47,096
28.4	38	*Iowa State Daily*	Iowa	13,371
28.4	35	*The Ripple*	Leicester	12,316
27.0	49	*Warwick Boar*	Warwick	18,176
26.9	23	*The Gazette*	U of W Ontario	8,555
26.4	8	*Excalibur*	York, Ontario	3,027
26.2	14	*The Martelet*	U of Br Columbia	5,340
25.3	73	*Grip*	Manchester	28,847

Table 5: The COLL idioms distributed by newspaper, lowest densities (normed per 10,000 words)

Norm	Actual			Words
4.7	5	*The Linfield Review*	Linfield C, Oregon	10,707
4.7	3	*The DePauw Online*	DePauw U, Indiana	6,450
4.6	4	*UCF Knight Wire*	U of Central Florida	8,727

4.2	3	*E Carolina Tech Online*	North Carolina	7,091
4.2	1	*The Voice*	Langara C, Br Columbia	2,405
4.0	8	*Tulane Hullabaloo*	Toulane, La	20,027
3.8	1	*The Independent*	Clark C, Washington	2,616
3.8	2	*The Oracle*	U of S Florida	5,242
3.5	2	*The Online Lode*	Michigan Tech	5,665
3.3	1	*The Student Printz*	U of S Miss	2,986
3.3	1	*The Technicianonline*	N Carolina SU	3,059
3.2	2	*The Reveille*	Louisiana SU	6,157
2.8	4	*The Collegian*	Chico SU, Colorado	14,466
2.5	2	*Daily Lumberjack*	Humboldt SU, Ca	7,852
1.2	1	*On-Line Forty-Niner*	Cal State at Long Beach	8,269

Another aspect that can be investigated is the extent to which the *CCDI* geographical labels match the distribution in COLL. The first 745 idiom entries listed in *CCDI* (in alphabetical order) were therefore investigated for cases where a geographical label is provided. Since the numbers involved are so small, *CCDI*'s categories are here reduced to the British Isles (BrI), North America (NAm) and Australia/New Zealand (Austr), with *CCDI*'s 'mainly X' and 'X' categories merged.[10] The results are seen in Table 6, with the obvious caveat that the 'mainly' category in itself indicates a wider spread, so that 'incorrect' is a somewhat infelicitous label:

Table 6: Distribution of geographically marked idioms

	Neutral	BrI	Mainly BrI	NAm	Austr	Total
Idioms labels	565	109	32	36	3	745
Idioms found		21	10	49	1	
Label correct		16	6	47	1	
Label 'incorrect'		5	4	2	0	

One point that immediately becomes apparent is that the *CCDI* probably contains far more British Isles items than North American, although it is unclear whether this is a consequence of the structure of the Bank of English at that point (*CCDI* 2002: v) or – rather more improbably – genuinely reflects a lesser use of idioms in North American English. Since the number of 'BrI' idioms found in COLL's North American texts is disproportionately higher than 'NAm' idioms found in its BrI texts (five 'BrI' + four 'mainly BrI' of 31 versus two 'NAm' of 49 found in BrI texts), the probability is that the NAm component is underrepresented in the then-Bank of English, and that these idioms actually do have some currency in NAm English.

By comparison, it is striking how few of the idioms labeled NAm appear to be used in the British Isles newspapers. Here, the *CCDI* labels and the corpus agree, which suggests that young BrI writers match the patterns of their elders (as reflected in the Bank of English data). Given the generally accepted claim that the flow of new vocabulary items is currently primarily from North America, one might have assumed that young writers were more willing to accept trans-Atlantic loans: is it perhaps the case that idioms are not as easily transferred as less complex lexical structures? (One memorable case of actual transfer is the former Americanism *it isn't over until the fat lady sings*, which was bandied about during the 1992 Clinton-Bush election campaigns and thus began to appear in British political commentaries; by 1995, the first edition of *CCDI* was listing it without a geographical label.)

Since the frequency of idioms is relatively low, a comparison with other corpora should be of interest. Using three idioms where certain frequencies are specified in Moon (1998: 66) and matching them to items occurring in COLL, we obtain the comparison in Table 7. The range of variation does not suggest that a single factor, such as corpus size, register or geographical region, is decisive, although a much more extensive comparison would clearly be desirable.

Table 7: Frequency comparison to corpora cited in Moon (1998: 66)

Idiom	COLL	*OHPC* (Hector)	*BNC*	*Bank of English*	*Oxf Engl Corp*
(per m wds)	(4 m)	(18 m)	(100 m)	(211 m)	(1 b)
spill the beans	1.08	0.39	0.39	0.63	0.54
beg the question	2.17	2.5	1.16	1.4	1.81
call the shots	0.54	0.94	0.48	1.22	0.89

Another way of comparing COLL results to those from other corpora is to check them against results from searches of newspapers on CDs (such collections are usually annual, e.g. the 1996 *New York Times*). Drawing on a previous study of a small subset of the idioms in *CCDI* (Minugh 1999), we may note that of the top 25 idioms found in a series of such newspaper CDs, 18 were also found in COLL, with only the following missing: *fig leaf, fall into the wrong hands, the last gasp, shoot the breeze, go ape (crazy), never look back* and the BrI variant *take the biscuit* (the NAm variant *take the cake* did occur).[11]

7 Idiom frequency and variation in COLL

Table 8 lists the most common headwords, together with how many different senses (including those listed as ‘other idioms’) they occur in. As expected, parts of the body and other basic words are dominant, but there are a few notable exceptions, such as the adverb *there* (*to be there for* someone), the particle *up* (*up and running, up and coming*) and the logical linker *wake* (*in the wake of something*).

Table 8: The most common headwords, with total tokens and senses

Word	Tokens	Senses	Word	Tokens	Senses
eye	103	18	*home*	43	9
line	102	16	*name*	43	6
hand	101	21	*heart*	40	8
mind	98	9	*profile*	39	2
door	87	11	*there*	39	1
face	78	14	*day*	37	10
kick	71	7	*corner*	36	5
edge	59	7	*time*	35	3
track	56	7	*ball*	33	8
work	54	4	*foot*	33	20
finger	53	12	*business*	32	6
head	53	20	*hell*	32	12

ground	52	11	*air*	29	5
up	48	3	*blood*	28	12
mark	47	8	*dead*	28	5
wake	46	2	*sight*	28	4
boy	44	8	*way*	28	4
board	43	5	*fire*	26	6

The most frequent idioms are listed in Table 9, together with the number of tokens. Again, ‘other idioms’ (indicated by italics) are included in the calculations, in this case surfacing as the idiom *open/close doors*, which turns out to be the second-most frequent idiom in COLL, although not in *CCDI*. This may reflect the concept of higher education as a process that opens – or closes – (career) doors, something repeatedly reflected in the COLL articles about the approaching end of the school year and graduation, with the concomitant need to find gainful employment.

Table 9: The most common idioms, with total number of tokens

mind	61	keep/bear sth in mind
door	48	open/close doors
work	43	in the works
wake	42	in the wake of sth
face	39	face to face
kick	39	kick ass/butt
there	39	be there for sb
edge	34	the cutting edge (= advanced)
profile	33	a high profile, high-profile
hand	31	in your hands, in the hands of sb
finger	28	point the finger at, finger-pointing
shoe	25	step into/fill sb’s shoes
up	25	up and coming

eye	24	open your eyes
school	24	old school
hold	23	put on hold
line	23	the bottom line
corner	22	just around the corner
line	22	a fine line
scene	22	behind the scenes
time	22	(hit) the big time
name	21	call sb names, name-calling
up	21	up and running
belt	20	under your belt

In order to investigate the distribution more closely, the 1,036 idioms in the first 50 newspapers (in alphabetical order, including some Australian, Canadian and New Zealand material), comprising some 628,000 words, or roughly 16 percent of the corpus, were further examined for a number of possible positions, uses and forms in their individual articles. The first area investigated was their position. Since journalistic paragraphs tend to be brief even in the written form, let alone in electronic texts, it did not seem meaningful to consider their role in the paragraph. Instead, they were noted if they were part of the article's title, as 41 (4%) were (with half being echoed in the text proper), in the summary or opening sentence of the article, as 23 (2%) were, or at the end of the article, as 22 (2%) were. Thus, a rhetorical opening or summing-up is clearly a possible, but hardly a major function of these idioms. Nor did any individual idiom recur repeatedly in these positions, with the single exception of *thumbs up/down*, which one writer used as a structuring device for a list of items he approved/disapproved of.

Another dimension to consider is that of text types within the COLL newspapers. Eight types of text were distinguished, as specified in Table 10:

Table 10: COLL sample idiom distribution by genres (1,036 idiom tokens)

Papers[12] (of 50)	Size (1,000 wds)	Category all items	Idioms	I/10M
13	17.6	**Editorial**	50	8.14
21	43.3	**Profile**	95	6.29
37	117.8	**Arts**	241	5.87
35	87.3	**Sports**	138	4.54
49	288.8	**Feature**	435	4.32
19	21.4	**Letters**	28	3.76
29	43.1	**News**	45	3.00
9	9.0	**Ads**	4	1.28
Papers (of 50)	Size (1,000 wds)	Category no ‘other’	Idioms	I/10M
13	17.6	**Editorial**	45	7.33
21	43.3	**Profile**	84	5.56
37	117.8	**Arts**	227	5.53
35	87.3	**Sports**	126	4.14
49	288.8	**Feature**	405	4.02
19	21.4	**Letters**	23	3.09
29	43.1	**News**	44	2.93
9	9.0	**Ads**	3	0.96

Compared to standard newspapers, there are several primary differences:

- Most COLL papers did not come out on a daily basis in 1999; some were very infrequent, indeed. They carry relatively little “breaking” **News** of the standard local-national-international type. Instead, they primarily include presentations or discussions of university affairs and regulations, campus life or state legislation, all of which have been categorized as **Features**, except the obvious special areas of **Arts** and **Sports.**
- There are fewer “non-core” newspaper subject areas represented (no gardening, home-improvement or real-estate sections, no chess or bridge notes, market or financial analyses, science section or syndicated columns);

instead, extensive reflective (and sometimes parodic) **Feature** articles about college life frequently occurred.

- Very few of these papers had **Ads**, and not many had clearly demarcated **Editorials** or **Letters**. The fair number of relatively extensive feature articles that focused on an interview or in memoriam were instead designated as **Profiles**.
- The **Arts** articles also included the occasional story or poem.

The upper half of Table 10 includes all idioms tokens found, normalizing them to per 10 million words, and dividing by 3,485 (the number of idiom types in *CCDI*). However, there are 79 'other' idioms, i.e. idioms that are not in *CCDI*, included in these figures. The lower half of Table 10 removes them, and thus produces a comparison strictly within the *CCDI* frame. The rankings remain the same, although at a somewhat lower level. Nevertheless, as noted above, they all remain within the rough guidelines from *CCDI*, 1st ed.

The **Editorial** voice appears to be the most prone to use idioms, which fits in well with the concept of idioms as received wisdom (unlike **Letters**, where the writer would then appear not to be able to invoke such authority). A similar explanation may hold for the **Arts** and **Profiles** sections, where normative judgments are often expressed about the bands, music or movies they are reviewing, or the person being interviewed or remembered. Interestingly enough, **Sports** articles turn out to be nearly indistinguishable from **Feature** articles, which form the lion's share of the articles, both in number and in total number of words.

Another potential dimension is that of quotes versus "regular" text. Of the 1,036 tokens, 146, or 14.1 percent, were found in quotes, which is far above the percentage of quotes in the texts as a whole, even though the genres **Editorial** and **Letters** had no quotes at all. While this may indicate that people are more willing to talk in idioms than to write them, it is also entirely possible that this is an artifact of the reporter's deliberately choosing an idiom as part of the selection process, especially in cases where an interview is involved. This is in particular supported by the figures for **Profiles**, as seen in Table 11. As regards individual idioms, only one, *up in the air* ("not yet decided"), occurred 3 times (in 3 separate newspapers), and another 20 occurred twice, so again there is no clear tendency for an individual idiom to be strikingly frequent in this respect.

Table 11: COLL sample idiom distribution of quotes by genres

Size (1000s)	Category all items	Idioms found	Percent found	Idioms expected	Percent expected
117.8	**Arts**	18	12.3%	30	20.2%
288.8	**Feature**	59	40.4%	72	49.6%
43.1	**News**	16	11.0%	11	7.7%
43.3	**Profile**	25	17.1%	11	7.4%
87.3	**Sports**	28	19.2%	22	15.0%
582.2	***Total***	*146*	*100%*	*146*	*100%*

Table 12 looks instead at the distribution of the 79 ‘other’ idioms (those not in *CCDI*). Again, even if the numbers are quite small, **Editorial** and **Profile** (and Letters) are the most frequent innovators, while **News** is the least common. The others cluster around a middle value. (Note that this is merely the likelihood of any new idiom at all appearing, not that of a specific new idiom.)

Table 12: COLL sample idiom distribution by genres (1,036 idiom tokens)

Size	Category	Idioms	Idioms
(1000s)	all items	number	normed/10M
17.6	Editorial	5	28.4
43.3	Profile	11	25.4
21.4	Letters	5	23.4
87.3	Sports	12	13.8
117.8	Arts	14	11.9
9.0	Ads	1	11.6
288.8	Feature	30	10.4
43.1	News	1	2.3

Table 13: COLL sample geographical distribution of ‘other’ idioms

Area	Idioms	Percent
New Zealand	5	32.5%
Great Britain	21	21.0%
US	47	18.8%
Australia	2	14.1%
Canada	4	13.6%

Table 13 indicates the geographical region for these innovations. Again, with such small numbers, it is difficult to place too much faith in the high percentage for New Zealand, particularly since these ‘other’ idioms, upon examination, often turn out to be old friends. For the entire COLL corpus (as opposed to the subcorpus just examined), there are a total of 182 types and 488 tokens with already-identified headwords, but where the idiom in context clearly did not fit the given *CCDI* meaning(s). The great majority appear to be well-established, with only some 40 types unknown to the author. Consider the most common of these idioms, as shown in Table 14:

Table 14: The 18 most frequent ‘other’ idioms (i.e., not in *CCDI*)

word	token	idiom
door*	48	*open/close ~s*
boy	15	*good old ~s*
bang	13	*finish with a ~*
bridge*	13	*~ the gap*
business*	13	*none of sb’s ~*
hand	13	*a helping ~*
kick*	13	*~ back* (= relax)
fun*	10	*~ and games*
boy	9	*bad ~*

boy	9	*wonder ~*
ground*	9	*stand your ~*
board	8	*get on the ~* (=first goal)
board	8	*get/come on ~* (= join)
rein	8	*take the ~s*
door*	7	*foot in the ~*
melting*	7	*the ~ pot* (=assimilation)
back*	5	*~ to basics*
bomb	5	*drop a ~*

Of these 18, exactly half (those with an asterisk) are to be found in both Longman's *Dictionary of contemporary English* (2005) and Oxford's *Advanced learner's dictionary* (2005). These idioms (Oxford also lists *rein*, *bang*) would all appear to be well-established. Other, less frequent items are also plausibly already established, such as (6), where *the plank in your own eye* is clearly a modernization of *the beam in your own eye*:

(6) As it says in Matthew 7:4-5, "How can you say to your brother, 'Let me take the speck out of your eye,' when all the time there is *a plank in your own eye*? You hypocrite, first take the plank out of your own eye." (*The Tech Talk Online*, Louisiana Tech University)

Perhaps the only genuinely new idiom noted (3 tokens) is seen in (7):

(7) "We're really, really far *ahead of the curve* in terms of our technology here at the moment," he said. (*Montana Kaimin*, University of Montana)[13]

The tentative conclusion is that the idiom list in *CCDI* could probably easily be expanded to include many of these 'other' idioms, but that it will not materially alter the frequency patterns observed.

A much more important point is that a number of idioms are highly frequent because they have been used as proper names: the COLL corpus contains some 157 different name types, with 260 tokens (many of which are multiple references to the same book, film or band), as detailed in Table 15. It is possible to question whether these should be considered as idioms at all (Wray 2002: 3-4, Coates 2000: 1166); yet it seems clear that most names that are based on idioms also involve a playful evocation of the *meaning* of the idiom, and occasionally of its composite parts, as well. Thus, the play entitled *The Other Side of the Closet* deals with issues that arise after a homosexual “comes out of the closet” and “on the other side” of a passage (– or is it a Narnian wardrobe?), while the coffee shop called *The Common Ground* is not merely evoking “finding common ground [for discussions]”, but also *coffee grounds.*

Table 15: Major type distributions of the names in the COLL corpus (tokens)

Type	Nr	%	Examples
Films	79	31%	*Wonder Boys, The Full Monty, Drop Dead Gorgeous*
Albums	32	12%	*Salad Days, Bury the Hatchet, Heads Are Gonna Roll*
Coinages	30	12%	*Out of the Cold (housing), Glass Ceiling (report)*
Bands	26	10%	*Ten Foot Pole, Eagle-Eye Cherry, Top Dog*
Songs	17	7%	*My Name is Mud, Another Brick in the Wall*
Organization	15	6%	*Hotter Than Hell (Harley Days), the Old Boys Union*
Bar/Restaurant	13	5%	*The Fat Cat, The Dark Horse Tavern, Mick E Fynn's*
TV shows	9	3%	*Short and Curlies, Get a Life, Seventh Heaven*
Sports	8	3%	*Dream Team, The Trail Blazers, Wild Card playoffs*
Stores	7	3%	*The Paper Trail (stationary), The Common Ground*
Plays	5	2%	*The Other Side of the Closet*
Books	4	2%	*Be My Guest (Hilton), Daniel in the Lion's Den*
Magazines	4	2%	*Raw Deal*
Theaters	4	2%	*The Black Box, Off Limits (exhibition)*
Commerce	2	1%	*Helping Hands (nanny agency), Spic ‘n Span*
Computers	2	1%	*Cloak and Dagger* (game)
History	2	1%	*the Reign of Terror*

As might be expected in a college corpus, about 70 percent of these names refer to various aspects of the entertainment industry, which means that they are idioms chosen by others, and thus not actually “used” by COLL writers. However, this would be true under any circumstances: one can hardly keep referring to *The Full Monty* as “the-film-which-must-not-be-named”. At the very least, such contemporaneous names help keep the phrases alive and in circulation.

Any corpus-based investigation of idioms also invariably stumbles on meta-level awareness of idioms, often triggered by sheer curiosity about an idiom’s origins or meaning. There were, however, remarkably few such instances in the COLL corpus, with a total of only 13 tokens, including four that mark this status by the use of ‘proverbial’, as in (8):

(8) Fisher learned Monday afternoon that her team, which *was on the proverbial bubble* for the NCAA Championships, would not be making their first trip to a regional site as the bids were announced via conference call Monday afternoon. (*The Arkansas Traveler*, University of Arkansas)

Looking at other features, one point that may be noted is that roughly 10 percent of the COLL idioms occur in articles that indulge in what we might call ‘verbal fireworks’, where word play, puns, slang, parody or outrageous statements compete for attention, as in the quote attributed to Edith Hubbard that introduces an article in (9):

(9) “Editor: a person employed by a newspaper, whose business it is *to separate the wheat from the chaff*, and to see that the chaff is printed.” (*The Gauntlet*, The University of Calgary)

More extreme examples are to be found in *Scrapie* (University of Bradford), where the author used *in my crystal ball* no less than ten times in an article introducing his fellow editors, or the inspired slanger in *Red and Blackmail* (Jesus College, Cambridge) who ended ten consecutive short statements with a summarizing *pants* (i.e. “terrible”) and the following 14 with the equally pithy *wicked* (i.e. “great”); cf. the discussion of *thumbs up/down* at the beginning of this section.

8 Idiom variation and context

Recently, several works have examined the question of variation in idioms as regards form and context (Gustawsson 2006, Langlotz 2006). The full-scale model presented by Langlotz, particularly in Chapters 6 and 7, covers the full range of variations and indicates how they may be categorized, not least with the help of cognitive schemas. Gustawsson takes up a number of the same issues, and both provide examples that are similar to those seen in Table 16, which gives examples of the variation to be found in COLL idioms.[14] All in all, 270 of COLL's 5,439 tokens fall into this group.

Table 16: Types of variation in the COLL idioms

Type of change	Tokens	Example
substitution	136	a *juggling* act, alarm bells *go off,* your cup of *juice,* easy as *torte,* a blot on our *history*
grammar/roles	31	*come>bring* full circle, *draw* the curtain > *is drawn,* in *pole position* >a *pole-sitter*
elaboration or specification	30	much ado about *the choreography,* over the *hump* > *Hump Day,* all dressed up and nowhere to *sing,* a *purchasing* wedge
shift	18	that's all *he* wrote, *in* > *find* the groove, to jump *off* the bandwagon
reversal of usual meaning	15	*miss* the bandwagon, the buck *starts* here, all that glitters *is* gold, time *off* your hands
puns	12	crash and *learn,* the "*Cookie*" crumbles (person's nickname), to *Kerry* the torch (Am politician)
blend of 2 idioms	9	a whole other ball *of wax*, the other side of the *closet*, blaze a trail *of glory*
twists	9	wake up and smell the *birdseed*, at the crack of *noon,* come hell or high *budget overruns*
inverted WO	5	to add *injury* to *insult, silver* clouds have a *dark* lining, more *dollar* for the *bang*
errors	5	score the *brunt* of their runs, put on the (~~*back*~~) burner, hand/tongue *and* cheek

In a recent paper, the author explored how idioms can incorporate an extraneous element, particularly into their primary NP (Minugh, 2007). In this context, Langlotz (2006: 194-215, 257-73) draws the distinction between **usual** and **occasional** variation, which roughly corresponds to Gustawsson's **compatible** and **incompatible** modifiers (2006: 92-104), both based on whether the element can be considered a modifier of the idiom itself (and thus can be institutionalized/cognitively entrenched (Langlotz 2006: 199)) or is derived from the context within which the idiom is used (Gustawsson speaks of elements of the idiom as 'domain delimiters' (2006: 94)). Most of these occasional/incompatible elements thus correspond to **anchorings** in the text (Minugh 2007), i.e. refer to the surrounding context, although they are integrated into the idiom. Langlotz additionally notes that puns and the like cannot be dismissed as mere word play, as they often function in a similar, anchoring manner.[15]

The 1,036 idioms in the subcorpus contain only 9 NP anchorings, which would extrapolate to roughly 50 in the entire COLL corpus, or about 1 percent. Some examples:

(10) I actually got to stand on the field at Shea Stadium once, after the Mets had reached their peak and had begun their descent to *the bottom of the baseball barrel*. (*The Argus*, Wesleyan University)

(11) Global Connections will also hold our annual Bon Voyage Party on Saturday, April 25 from 1 pm to 4 pm at Eagle Beach. It will be *an international potluck* held in coordination with the American Indian Science and Engineering Society and a number of other clubs here on campus. (*The Whalesong*, University of Alaska)

(12) The defensive secondaries are strong, and the linebackers and defensive line *have some experience under their immense belts*. (*The Gauntlet*, University of Calgary)[16]

This low number is not unexpected (cf. Minugh, 2007), but anchoring per se is much more frequent, either by the use of postmodification or by embedding the entire idiom in an NP: there are a total of 341 instances in the subcorpus, which would extrapolate to roughly 1,785 in the entire COLL corpus. Langlotz states that "*[o]f*-constructions are the most frequent postmodifying PPs to be found in idiom variants by far. Interestingly, the *of*-complements in the variation-data only contain target-related NPs and thus perform the primary function of topic indication" (2006: 263). However, in the COLL material, only 40 of 341 were of this type, and 21 of these idioms require an obligatory NP, as in (13):

(13) … all of which rest uneasily *in the hands of* a director more accustomed to understated narrative, carefully crafted dialogue and visual self-indulgence. (*The Edinburgh Student*, Edinburgh University)

Instead, it appears that in addition to adnominals (e.g. *a high-profile project*), there are a range of postmodifiers (various PPs, *that*-clauses, and even colon + main clause) available for this process of textual anchoring.

9 Final remarks

Although we have had the luxury of “very large” corpora such as the *BNC* long enough to begin to forget how revolutionary multimillion word corpora were, when it comes to idioms, with their unusually low frequencies, we have not yet exhausted the need for more primary data. Consider John Sinclair’s advice: “So if we need, say, fifty occurrences of a sense of a word in order to describe it thoroughly, then the corpus has to be large enough to yield fifty instances of the least common sense” (1991: 102). We have not yet really reached that stage for most idioms, and there is presumably a great deal still to learn about these fascinating expressions.

Notes

1. The term *idiom* will be defined and discussed in section 3, below.
2. The advent of large-scale blogging is beginning to change this situation, not merely as regards collecting writing by young people, but more fundamentally by the very fact that blogging encourages public writing by young people. Thus, blogs comprise about 15 percent of the 1 billion word online collection in the Oxford English Corpus (Oxford University Press 2006) – although we of course do not know the actual age of most of these bloggers.
3. Sites that published in PDF format were not included, however, as conversion into text format was at that point beyond the author’s resources. Syndicated and wire-press material was deliberately excluded, although a few such articles appear to have slipped in.
4. The figures in Table 1 are slightly different than those cited in Minugh (2002:76) (with percentages rounded to the nearest percent). In the present paper, the aggregate has been adjusted downward, due to the subsequent removal of duplicate articles. The change does not materially affect the present discussion.
5. For a more general discussion of idioms and related terms involving formulaic sequences, see e.g. Wray (2002), whose Figure 1.2 lists more than fifty different terms and variants “used to describe aspects of formulaicity” (2002: 9).
6. In hindsight, if we consider the question of idioms and metaphors that are “alive” or “dead” (cf. e.g. Gibbs 1994: 273–278, and implicit in Gustawsson 2006: 34–35), this failure to keep track of purely “literal” uses was ill-advised. There were, however, quite few such rejected instances.
7. For an example of detailed examination of such permutations, see Gustawsson (2006).

8. Since the headwords are extracted by WordSmith from an untagged corpus, the total number includes many cases where the headword is the wrong part of speech or a homograph.
9. This information is not found in their second edition. For its relevance to standard newspaper corpora, see the discussion in Minugh (1999). For a more general discussion, see Moon, Ch. 4 (1998) and Gustawsson (2006: 72–73).
10. *CCID* has a vanishingly small proportion of idioms labeled [SCOTTISH] or [IRISH]. For example, *set the heather on fire* is labeled [mainly SCOTTISH].
11. Interestingly enough, a less polite variant of *shoot the breeze* (and one not found in *CCID*), *shoot the shit*, was not found in COLL, either. As a whole *shit/shite* was astonishingly infrequent (only 93+10 tokens).
12. This column merely indicates the number of newspapers (out of the total of 50) that had text in the respective categories. This indicates above all that a certain caution is necessary for the figures for **Editorial** and **Ads**.
13. For example, there were no examples of *ahead of/behind the curve* in the *BNC*.
14. Both Langlotz and Gustawsson provide systematically defined categories for variation types; the labels in Table 15 are not intended to be as carefully worked out.
15. He approvingly quotes Burger: “Die Erweiterung hat in diesen Fällen nicht den Effekt eines Sprachspiels, sondern sie gibt eine Art Anweisung, wie man die phraseologische Ausdrucksweise in den wörtlichen Gedankengang zu ‘übersetzen’ habe[…].” (Langlotz 2006: 200)
16. While (12) would probably not qualify under Gustawsson’s system (after all, a belt *can* be ‘immense’), the adjective really only makes sense in this context if you know that a defensive line in (North American) football usually consists of men that weigh well over 200 pounds.

References

Advanced learner’s dictionary, 7th ed. 2005. Oxford: Oxford University Press.

Coates, R. 2000. Singular definite expressions with a unique denotatum and the limits of properhood [electronic version]. *Linguistics* 38(6): 1161–1174.

Collins *COBUILD dictionary of idioms*, 2nd ed. 2002. Glasgow: HarperCollins.

Collins *COBUILD dictionary of idioms*, 1st ed. 1995. Glasgow: HarperCollins.

Dictionary of contemporary English, 4th ed. 2005. Harlow: Longman.

Gibbs, Raymond.W., Jr. 1994. *The poetics of mind*. Cambridge: Cambridge University Press.

Gustawsson, Elisabeth. 2006. Idioms unlimited. A study of non-canonical forms of English verbal idioms in the British National Corpus. Gothenburg: Department of English, Gothenburg University.

Langlotz, Andreas. 2006. *Idiomatic creativity: A cognitive-linguistic model of idiom-representation and idiom-variation in English*. Amsterdam: John Benjamins.

Minugh, David. 2007. The filling in the sandwich: Internal modification of idioms. In R. Facchinetti (ed.). *Corpus linguistics 25 years on*, 207–224. Amsterdam: Rodopi.

Minugh, David. 2002. The COLL Corpus: Towards a corpus of web-based college student newspapers. In P. Peters, P. Collins and A. Smith (eds.). *New frontiers of corpus research*, 71–90. Amsterdam: Rodopi.

Minugh, David. 1999. You people use such weird expressions: The frequency of idioms in newspaper CDs as corpora. In J. M. Kirk (ed.). *Corpora galore: Analyses and techniques in describing English*, 57–71. Amsterdam: Rodopi.

Moon, Rosamund. 1998. *Fixed expressions and idioms in English: A corpus-based approach.* Oxford: Clarendon.

Oxford English Corpus. 2006. Oxford University Press. Database. Retrieved on Nov. 27, 2006 from http://www.askoxford.com/oec/mainpage/oec01/?view=uk.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.