Studies in Corpora and Idioms: Getting the cat out of the bag
David Minugh
Abstract

“Idiomatic” expressions, usually called “idioms”, such as *a dime a dozen*, *a busman’s holiday*, or *to have bats in your belfry* are a curious part of any language: they usually have a fixed lexical (why a *busman*) and structural composition (only *dime* and *dozen* in direct conjunction mean ‘common, ordinary’), can be semantically obscure (why *bats*?), yet are widely recognized in the speech community, in spite of being so rare that only large corpora can provide us with access to sufficient empirical data on their use.

In this compilation thesis, four published studies focusing on idioms in corpora are presented. Study 1 details the creation of and data in the author’s medium-sized corpus from 1999, the 3.7 million word *Coll* corpus of online university student newspapers, with comparisons to data from standard corpora of the time. Study 2 examines the extent to which recognized idioms are to be found in the *Coll* corpus and how they can be varied. Study 3 draws upon the *British National Corpus* and a series of British and American newspaper corpora to see how idioms may be “anchored” in their contexts, primarily by the device of premodification via an adjective appropriate to the context, not to the idiom. Study 4 examines idiom-usage patterns in the *Time Magazine* corpus, focusing on possible aspects of diachronic change over the near-century *Time* represents.

The introductory compilation chapter places and discusses these studies in their contexts of contemporary idiom and corpus research; building on these studies, it provides two specific examples of potential ways forward in idiom research: an examination of the idioms used in a specific subgenre of newspapers (editorials), and a detailed suggestion for teachers about how to examine multiple facets of a specific modern idiom (*the glass ceiling*) in the classroom. Finally, a summing-up includes suggestions for further research, particularly at the level of the patterning of individual idioms, rather than treating them as a homogeneous phenomenon.

**Keywords:** *Coll corpus*, corpora, corpus creation, idioms, idiom variation, idiom-breaking, online newspapers, student newspapers, college newspapers
Studies in Corpora and Idioms:

Getting the cat out of the bag

David Minugh

Therefore I wol seye a proverbe,
That he that fully knoweth th'erbe
May savely ley hvt to his epe—
Withoute drede, this ys no lye.

– Geoffrey Chaucer, *House of Fame*
To my wife and children, who like to remind me what fun it is to have a Fountain of Useless Knowledge in the family, yet who patiently listen to me nevertheless.
Publications Included in This Thesis


Three previous studies of relevance to this thesis should be briefly mentioned: Minugh 1997 introduced the use of mainstream quality British and American newspaper CDs as corpora, which during the 1990s had become available (and searchable) via annually issued CDs. ¹ Minugh 1999 returned to these newspaper CDs in order to examine their use of idioms, and finally, Minugh 2001 investigated the occurrence of idioms in three entirely different genres: well-known popular songs from the 20s and 30s (interesting as they are both written, performed and heard), transcripts from modern TV soap operas (largely tightly scripted), and transcripts for TV talk shows like

---

¹ The most important were *The Independent* (from 1987 until the late 1990s) and *The New York Times* (from Oct. 1990 until 2000, after which lawsuits about control of the texts cited brought the enterprise to a stand-still); several other English papers, such as *The Times*, were available then, as well as e.g. the *Los Angeles Times, USA Today* and *the Sydney Morning Herald*. Internet-based search engines at the individual newspapers have of course replaced them.
*Oprah* and *Geraldo* (oral, and largely unscripted, including audience and caller participation). All of these studies confirmed the general frequency patterns mentioned in *Collins COBUILD Dictionary of Idioms* (henceforth *CCDI*), and failed to turn up a strikingly large use (or lack) in any specific genre. They may be seen as background and preparation for the subsequent studies actually included in the thesis.

Since this is a compilation thesis, the four studies are re-publications; they have not been reworked, but are published here as they were published originally. All comments on their potential merits and regrettable deficiencies have instead been worked into the introductory chapter. This procedure will occasionally feel somewhat cumbersome, but it ensures that the reader will encounter the original studies as they were published.
Abstract ...........................................................................................................................................3
Publications Included in This Thesis ..........................................................................................9
Acknowledgements ...................................................................................................................14
Preface ............................................................................................................................................15
1 Introduction ..................................................................................................................................18
  1.1 Idioms and Corpora Conjoined ...........................................................................................18
     1.1.1 Idioms: A familiar puzzle ............................................................................................20
     1.1.2 Corpus linguistics and idiom studies: An expanding universe ...................................37
  1.2 Research Aim and Questions ..............................................................................................39
  1.3 Method ....................................................................................................................................42
     1.3.1 Overview ......................................................................................................................42
     1.3.2 Creating a corpus: The Coll Corpus .............................................................................46
     1.3.3 Examining an entire corpus: The College Idiom .........................................................48
     1.3.4 Examining an idiom structure in corpora: The Filling in the Sandwich ....................51
     1.3.5 A diachronic investigation of idioms in a corpus: Is Time A-changin’? ......................54
  1.4 Summary of the main findings .............................................................................................57
     1.4.1 The Coll Corpus ...........................................................................................................57
     1.4.2 The College Idiom ......................................................................................................60
     1.4.3 The Filling in the Sandwich: Internal modification of idioms ....................................62
     1.4.4 Is Time A-changin’?: A diachronic investigation of the idioms used in Time ..........65
  1.5 Discussion of the Findings ......................................................................................................68
     1.5.1 Corpus creation .............................................................................................................68
     1.5.2 Comparing corpora .......................................................................................................70
     1.5.3 Choosing your idioms ....................................................................................................73
     1.5.4 Idiom distribution in the corpora ..................................................................................75
     1.5.5 Idiom variation in corpora ............................................................................................80
  1.6 A Qualitative Substudy: Idiom use in Coll editorials .........................................................84
  1.7 A Pedagogical Application: Exploring the glass ceiling ....................................................96
1.8 Conclusions and Future Research.................................................................103

References ...........................................................................................................110

2 Papers .............................................................................................................123
  2.1 Study 1: The Coll Corpus: Towards a corpus of web-based college student newspapers........................................................................................................125
  2.2 Study 2: The College Idiom: Idioms in the COLL Corpus .................147
  2.3 Study 3: The Filling in the Sandwich: Internal modification of idioms .........................................................................................................................169
  2.4 Study 4: Is Time A’Changin’?: A diachronic investigation of the idioms used in Time.................................................................................................189

Sammanfattning (Summary in Swedish)..............................................................209
Abbreviations

ACE Australian Corpus of English
ALD [Cambridge] Advanced Learner’s Dictionary
AmE American English
ANC The American National Corpus
Archer The Archer Corpus
BE06 The BE06 Corpus of British English
BLOB-1931 The BLOB-1931 Corpus
BNC British National Corpus
BoE Bank of English
BrE British English
BYU-BNC The BYU interface for the BNC
CCDI Collins COBUILD Dictionary of Idioms
CID Cambridge Idioms Dictionary
COCA Corpus of Contemporary American English
COHA Corpus of Historical American English
Coll The Coll Corpus
DAI A Dictionary of American Idioms
DOI Dictionary of Idioms and Their Origins
FEI Fixed idioms and expressions
GloWbE Corpus of Global Web-based English
Google-UK Google Books (BYU/Advanced): British English
Google-US Google Books (BYU/Advanced): America English
ICE International Corpus of English
LID Longman Idioms Dictionary
LDOCE Longman Dictionary of Contemporary English
MED Macmillan English Dictionary for Advanced Learners
NODE New Oxford Dictionary of English
ODEI Oxford Dictionary of English Idioms
OEC Oxford English Corpus
OED (Online) Oxford English Dictionary (Online)
OHPC Oxford Hector Pilot Corpus (subsumed in the BNC)
Time TIME Corpus
WWC Wellington Corpus of Written New Zealand English
XARA The Oxford interface for the BNC

A number after the abbreviation (e.g. ALD7) indicates the edition.
Acknowledgements

In the creation of a work such as a thesis, debts, both personal and intellectual, are invariably accrued, so that it is a great pleasure to in some small measure be able to acknowledge, and in so doing, repay them.

During my undergraduate years at Wesleyan University, I was fortunate enough to encounter gifted scholars with an abiding love of language and literature, and a commitment to passing on this joyous passion. The results of their efforts have stood me in good stead throughout my teaching career.

Numerous friends at the English Department of Uppsala University must not be forgotten, but above all thanks to Professor Sven-Johan Spånberg, who from the very first offered me unfailing support and friendship. Many thanks to him, and to many other colleagues, for friendship and fruitful years of discussion about English literature!

Over the years at Stockholm University, it has been my privilege to work with many scholars, both in our Department and in the Humanities. Spurred on by the curiosity and interest of my students, I drew heavily on the knowledge of my colleagues as I made the transition to becoming a linguist. It is with great pleasure that I have noted how our linguistic seminars have become ever more vibrant and challenging over the years, while retaining a supportive sense of collegial friendship.

To Professor Magnus Ljung I owe the first impulse to do more with corpora than just check on data, and his pioneering grasp of that field was helpful indeed as I began to do research. At the various ICAME conferences on corpus linguistics, I have had the privilege to meet many of the scholars whose names occur in the References section. Their extraordinary openness and willingness to discuss these matters even with a part-time scholar such as this then-Director of Studies was of inestimable value.

To Professor Philip Shaw and Professor Maria Kuteeva, great thanks for encouraging me into bringing this work to fruition, and to Docent Erik Smitt-terberg for meticulous and insightful consideration of the draft version. And to Professor Nils-Lennart Johannesson, grateful thanks for having had the privilege of enjoying your many years of friendship, learning, cooperation and support.
Preface

... like some watcher of the skies
When a new planet swims into his ken.
— John Keats

The origins of this work lie firmly embedded in the joy I have experienced in more than a third of a century of working with, and two thirds of a century of using, that most complicated and living of human abilities, language. It has been my privilege, not merely to exercise the birthright of every man, woman and child—that of acquiring a native language—, nor merely to have had the opportunity to seek the elusive mastery of a number of other languages, but also to have experienced the forced bilingualism that comes with emigration to another culture. Even as I continually strive to mold my native English to a yet more malleable instrument that may do my bidding, it changes me; and while I persist in hopes that Swedish and I may yet form ‘a more perfect union’, it, too, daily reveals depths and subtleties that surprise and delight. Living as I do in Europe, I need travel only a short distance to reach Finland, a country whose dominant language quickly reduces me to a poor scholar whose every word comes with difficulty, and whose eastern borders mark the point where there live peoples whose languages I do not speak at all.

It is thus from an ambiguous position, indeed, that I for many years have worked as a so-called language expert, whether as translator, dictionary-maker, teacher, or researcher. For where was certainty to be found? Numerous discussions over the years with British colleagues and their reference books convinced me that English contained many terms, collocations, stress patterns, constructions and idioms that were definitely not mine, and yet were English—and beyond them lay Australian and still other varieties. When I wrote on the blackboard (we used blackboards in those days), He snuck out the door without anyone noticing, my Swedish students were not slow to object, as their grammar books did not record such a verb form. Learned Swedish colleagues have patiently contributed to my education by calling to my attention still other patterns and changes.

2 Elfstrand-Gabrielsson (1960), Svartvik-Sager (1978), Quirk et al. (1985), Svartvik-Sager (1996) are silent on this; Estling Vannestål (2007) is the first to mention it. Modern learner
And to crown it all, American colleagues have also had their comments and evidence that shook my native-speaker certainty. When I showed up at work with a bag of chicken corn, they calmly informed me that this classic Halloween candy from my childhood was called candy corn. Since the label on the bag (packed in Chicago) agreed with them, I was forced to consult Volume I of *The Dictionary of American Regional English* (1985), which indeed confirmed that it is only the New York City region that calls (called?) it chicken corn. But why believe a book?

On first looking into a CD-ROM of *The New York Times* in 1992 I stood, like Keats, silent by a desk in the Wesleyan University library, looking indeed ‘with a wild surmise’. Coming from the English-speaking New World, I now lived in a corner of the Old World where my vocation required me on a daily basis to judge the correctness of (non-native, often would-be British) English. At that point I realized that the ability to instantly access huge amounts of corpus data to check such language matters need no longer be the preserve of the high priests of punch-cards at a few isolated research centers—with the help of such CDs, it could be done anywhere, even in far-off Ultima Thule, a land of other tongues (*Ohthere sæde his hlaforde, Ælfrede cyninge, þæt he ealra Norðmonna norðmest bude...He cwæð þæt nan man ne bude be norðan him*).³

I have thus spent much of the intervening years working with these CDs, and subsequently with the unfolding world of large corpora, both to discover and document details about many aspects of English, not least my growing interest in idioms, those most recalcitrant of language phenomena, and to seek out the limitations of these corpora—what do they reveal about language, and what do they distort? How typical are they of English in general, how consistent are they when compared to one another, are there areas they miss entirely? Such are the matters that form the core of this thesis.

But beyond the quest for knowledge and the desire to speak of language norms and patterns with as well-informed a voice as possible, there lies yet another powerful motive: the desire to empower my students, to provide them with tools that will enable them to undertake the same quest, to make their own well-informed judgments about language:

---

³ Although Ottar was a Scandinavian, his story has come down to us in Old English: “Ohthere said unto his lord, King Alfred, that he lived the furthest north of all Northmen. … He said that no one lived to the north of him.” (Bately 1980: 13/29–30, 16/1–2). Even if he probably lived not far from modern Tromsø (which is far north, indeed), his isolation is metaphorically familiar to many who throughout the world live in linguistic exile.
Meek young men grow up in libraries, believing it their duty to accept the views which Cicero, which Locke, which Bacon, have given; forgetful that Cicero, Locke, and Bacon were only young men in libraries when they wrote these books. (Emerson, The American Scholar)

Replace Cicero with Quirk, Locke with Chomsky, Bacon with Webster, and the majority of the meek young men with their sisters: as Emerson so urgently recommended, today’s students must encounter the living language, seek out its patterns and idiomatic leaps, not merely hear tell of them. When my students, working in an EFL environment, but with tools such as these CDs and other data bases, come back to me with proof positive that I have yet more to learn about English, then I feel that there may yet be hope for me as an educator, as one who, rather than regaling them with yet another threadbare tale of Rocs and diamonds on his seventh voyage, prepares the way for his students, that they in turn may stand upon a peak in Darien, lost in wonder at the world of language.
1 Introduction

1.1 Idioms and Corpora Conjoined

The present compilation thesis concerns the intersection of two areas of linguistics, idioms (in the narrow sense) and corpus linguistics. While idioms (and proverbs) have long been recognized as curious linguistic phenomena, corpus linguistics is a child of the late twentieth century, a tool that arguably did not become fully accessible to scholars and laymen alike until the arrival of the Internet. To some extent, the individual papers collected here are a history of the expanding opportunities for today’s linguists to explore and come to terms with idiomatic expressions in an explicitly quantitative way, and subsequently to use this wealth of data to produce more detailed qualitative studies of idioms.

Implicit in this description is that this thesis was not originally conceived of as a single project, but rather to some extent, like Topsy, just grew—in the sense that as new corpora came into being, new questions could be asked: in particular, studies of idioms based on real data derived from large corpora covering numerous genres and geographical areas. The awareness of and subsequent focus on idioms is, however, a theme that runs throughout the thesis and indeed, its projection into the future.

In an important book-length study of what she designated as FEI (‘fixed expressions and idioms’), Moon (1998: 59–74) presents extensive corpus evidence that idiomatic expressions are quite rare. The first edition of the Collins COBUILD Dictionary of Idioms (CCDI) specifically states that they primarily occur in the range of 1–5 instances per million (1995:xvii). Given

---

4 The classic layman’s and linguist’s examples of idioms are It’s raining cats and dogs (‘It’s raining very hard’) and He kicked the bucket (‘He died’), respectively. The first example has the distinction of being more taught than used, while the second is the source of a memorable Jimmy Durante sight gag in the comedian-loaded Hollywood film It’s a Mad, Mad, Mad, Mad World (1963).

5 “I spect I grow’d. Don’t think nobody never made me.” Uncle Tom’s Cabin, Ch. 20.

6 This was more cautiously expressed in the second edition (2002): “Nearly one third of the idioms in [the Collins COBUILD Dictionary of Idioms] occur less than once per 10 million words of the corpus.” For corroborative evidence for idiom frequencies in newspapers, see Minugh 1999. In his corpus-based examination of the language used in universities, Biber similarly notes: “In fact, most longer idioms are far too rare to be considered lexical bundles. Stereotypical idioms such as kick the bucket (meaning ‘die’) and a slap in the face (meaning ‘an affront’) are rarely attested in natural speech or writing” (2006: 134).
their relative infrequency, idioms are not easily collected from actual text samples or even moderate-sized corpora.

In the early days of English corpus linguistics, when the Brown and LOB corpora, at 1 million words each, were the standard, it was obviously difficult to use corpus data to examine such phenomena. During the last two decades, the basis for collecting and analyzing idioms has changed radically—for the better. At present, publicly accessible corpora such as the Corpus of Historical American English (COHA) and the Corpus of Contemporary American English (COCA) are around 400 and 450 million words, respectively; various corpora with more limited access, such as the Oxford English Corpus (OEC), with web-based data collection, contain roughly 2 billion words, and the Google-US corpus of American English from Brigham Young University, at 155 billion words, is yet again two orders of magnitude larger.\footnote{For the Brown and LOB corpora, see Kucéra & Francis 1967 and Garside, Leech & Sampson 1987, respectively. For further information on the BYU corpora produced by Mark Davies, see http://corpus.byu.edu/faq.asp. In addition, the massive web-based corpora such as ukWaK produced by the SketchEngine team under Adam Kilgarriff for a steadily expanding number of languages (http://www.sketchengine.co.uk/) offer immense opportunities, for a relatively modest annual subscription fee.}

Although strong claims have been advanced for considering corpus linguistics as theoretically important in its own right (particularly Sinclair 1991, Tognini-Bonelli 2001; see Gilquin & Gries 2009 for a more recent general view of its methodology and goals), the present focus is on the opportunities that expanding resources and understanding within corpus linguistics open up for the study of idioms. In this area, it is still true that “bigger is better”, if we consider the claim about words made by COBUILD founder John Sinclair:

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)

Clearly we need very large corpora indeed to find fifty examples of most idioms, let alone of their possibly multiple senses (for example, put your foot down, meaning either ‘exert your authority’ or ‘drive very fast’). Although this is now becoming possible, at least diachronically, it was clearly not the case when the studies presented here were conceived and carried out.
1.1.1  Idioms: A familiar puzzle

[Idiomatic expressions] are of special interest to linguists, psycholinguists, and lexicographers, mainly because of their syntactic and semantic idiosyncrasies as well as their unclear lexical status.  
(Fazly, Cook & Stevenson 2009: 61)

Idioms, as used in this thesis, have a hyponymous relationship to the much larger and rather amorphous linguistic area Moon (1998: 2–5) designates as phraseology or fixed expressions and idioms (FEI). This latter ‘field’ primarily consists of various multi-word units (MWU) that, as a first approximation, may be seen as some form of semantic and psychological whole. The following is intended as a brief review of the main issues involved in producing a satisfactory definition of what such a multi-word unit or fixed expression actually is.

1.1.1.1  Defining idioms

Terminology in this field has always been problematic... (Moon 1998: 2)

There appear to be two major areas of uncertainty among linguists about the definition of phraseology as a whole and its subfields: the labels themselves (and what they imply), and which subfields belong where in this taxonomy. To take just one extreme example, in an extensive discussion of Hockett’s A Course in Modern Linguistics (1958), Makkai (1972: 27–38) notes that Hockett’s discussion of the concept word (1958: 166–76) leads him to include among idioms such distinctly surprising candidates as personal pronouns, numbers, proper names, private codes and unfinished allusions. Hockett also explicitly includes morphemes as a type of idioms, and as Moon (1998: 10) notes, his “reductionist model of the lexicon” and “extension of idiom to morphemes and ad hoc formulations [mean] that idiom becomes too broad to be a practical category.”

We will begin with an overview of various phraseological labels and what they imply, then move on to our primary focus, the subgroup of (“classical”) idioms and how to define them. In the general survey at the start of her Formulaic Language and the Lexicon, Wray (2002: 8–10) compiles a list of over 50 different “terms used to describe aspects of formulaicity”, among which a

---

8 As we will see below, other scholars take a far more restrictive view of what an idiom is. Nevertheless, it should be noted Hockett is not alone. Pinker uses word, not only in the “normal” sense, but also explicitly adds a second, technical sense that sounds much like Hockett’s idiom: “A chunk of any size that has to be memorized—prefix, suffix, whole word, idiom, collocation—is the second sense of word...sometimes called a listeme, that is, an item that has to be memorized as part of a list” (1999: 24).
number of familiar labels will be found: collocations, complex lexemes, conventionalized forms, FEI, idioms, lexicalized phrases, phrasemes, multiword items/units, prefabricated routines and patterns, schemata, stock utterances and unanalyzed multiword chunks. Mel’čuk (1995: 170) and Cowie (1998b: 1–8) and sources cited therein add further terms such as phraseological units, lexical solidarities, fixed syntagms, word-combinations, phrasal lexemes and prefabricated units (prefabs), and Gustawsson’s (2006: 11) brief review of such terms adds several others, such as polylexical expressions and set expressions.

Other concerns may also be seen in e.g. the early discussion by Malkiel (1959: 115, also cited in Makkai (1972: 23):

[O]ne does well to steer clear of any reference to the ill-defined category of “idioms” or phraseological formulas. These have been variously spoken of as sequences yielding imperfectly to routine grammatical analysis, as passages strikingly rebellious to literal translation (this phrasing manifests simultaneous concern with more than one language), as semi-autonomous pieces of congealed syntax (a view implying the supremacy of the historical perspective), as word-groups whose aggregate meaning cannot be fully predicted even from thorough knowledge of each ingredient (a semantic approach), and, in stylistic or esthetic terms, as clichés, i.e., as combinations once suffused with fresh metaphoric vigor, but gradually worn thin by dint of use.

Malkiel’s comments clearly indicate that the choice of definition is in a sense at least partially determined by the paradigm the scholar is working within. Scholars of translation studies, for example, are primarily interested in the equivalence problem between languages (and cultures).9

Finkbeiner’s (2008: 21–23) terminological review of the situation in German indicates that this state of affairs is not merely a consequence of disagreements about English terminology:

Die terminologische Vielfalt deutet darauf hin, dass es sich bei den IS [= idiomatische Sätze /DM] um ein äußerst vielschichtiges Phänomen handelt, dessen Erfassung die Forschung vor eine Reihe von Problemen gestellt hat.10

After her review of the various possible labels, Wray produces a ‘working definition’ of formulaic sequence that places her firmly in the psycholinguistic camp:

---

9 For a discussion of this aspect of idiomaticity, as applied to the new world of film subtitling, see e.g. Pedersen 2007.

10 “The terminological abundance indicates that when it comes to IS [idiomatic phrases], we are dealing with an extremely multilayered phenomenon, one that poses a series of problems for research seeking to grasp it.” [trans. DM]
A sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar. (2002: 9, italics in the original)

Such a definition has of course the distinct disadvantage of not being easily made operational: establishing that such a sequence is retrieved whole from the memory brings us into the domain of psycholinguistic research.\footnote{Cf. particularly the work by Ray Gibbs, e.g. Gibbs, 2007, 1995, 1993 the overview in Steen 2007, and the recent comments about psychological and experimental approaches in Gilquin \& Gries 2009.}

It is also patently a psycholinguistically oriented definition, in contrast to, for instance, that of Erman and Warren (2000): ‘A prefab is a combination of at least two words favored by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization’. (Wray 2002: 4)

When accumulating evidence for these sequences, Wray opts in practice for corpus searches for her evidence, which brings her much closer to Erman and Warren’s position than her note from Chapter 1 suggests. She sees her discussion as primarily framed in explicit opposition to earlier generative approaches:

The Chomskian position offers the clearest contrast to the whole notion of circumstantial associations between words, is least tolerant of internally complex units, and holds itself separate from performance and pragmatics, the two axes of the model of formulaicity developed here. (2002: 10–11)\footnote{The discussion is found in Wray 2002: 10–18, and picked up again in her Chapter 14 (pp. 261–81). For an alternate, much more formally developed definitional taxonomy, see Mel’čuk 1995.}

In particular, the question of internally complex units is one that immediately strikes a chord in most earlier research about phraseology. The tension between mental storage as a unit and conventionalization is probably less significant, since these two concepts operate at different levels: mental storage occurs at the level of the individual speaker, while conventionalization necessarily involves the speech community, as well.\footnote{The important role of the speech community is traced in considerable detail in e.g. Croft 2000, particularly Chapter 4. A corpus approach runs a clear risk here: evidence of an idiom in a given domain or geographical/social group may merely mean that only speakers using that domain or in that geographical/social group recognize, let alone use, that particular idiom. This is a more serious threat than for “normal” lexical items, as idioms are so relatively infrequent as a whole (cf. e.g. CCDI 1995: v).}

Both Wray’s formulaicity and Cowie’s preferred term, phraseology, indicate that we are dealing with recurrent patterns in language, patterns that are
associated with some form of semantic and/or pragmatic content.\textsuperscript{14} These patterns can range in size from the morpheme (Hockett’s position) to entire sentences (or even possibly to entire texts, such as the formulaic language in \textit{the Lord’s Prayer} or the American \textit{Pledge of Allegiance}),\textsuperscript{15} and may be available for use anywhere in the range from non-domain-specific language all the way to highly specified pragmatic situations (e.g. the phrase \textit{How do you do?}, which is only used as a formal greeting among non-intimates).

Formulaic language is also seen as normally including or requiring some form of non-literal, i.e. figurative language, as emphasized in the definition used by Naciscione: “The phraseological unit is a stable, cohesive combination of words with a fully or partially figurative meaning” (2001: 20).\textsuperscript{16}

The primary advantage of the cover term phraseology is that it suggests a field or area—although it has yet to be shown that a satisfactory formal and semantic unity of the area does exist.\textsuperscript{17} For the purposes of the present work, it is sufficient to note that there exists a wide range of linguistic phenomena potentially subsumable under the umbrella of formulaic sequences, and—crucially—that idioms are firmly situated within this area.

As a starting point for a definition of idioms, the rather inadequate definition given by the \textit{Oxford English Dictionary (OED)} neatly illustrates both terminological confusion and the gap between the general uses of the word \textit{idiom} and the linguistic use in question here. Of the senses of \textit{idiom} currently offered in the online edition of the \textit{OED}, all five refer to forms or uses distinctively or specifically tied to an individual or group. Thus, senses 1–3 are listed as language-oriented (1 to a given language’s phraseology, 2 to a

\textsuperscript{14} A note of warning about multi-word units (MWUs) and similar terms: In most studies oriented towards phraseology, such as Fellbaum 2007a, MWUs are used for semantically coherent word groups, and thus have nothing to do with the more corpus-oriented concept of n-grams, which are statistically confirmed groupings of adjacent words found within a specified “window”, or range of words in a given corpus (cf. e.g. McEnery & Hardie 2012: 110). N-grams can be (and e.g. trigrams frequently are) semantic fragments: the trigrams \textit{on the order} or \textit{the order of}, are meaningless in and of themselves, but are part of the 4-gram \textit{on the order of}; which clearly does happen to have a meaning. MWU’s, however, are not determined statistically and do form coherent wholes. Individual groups of words can of course be both, as in the 4-gram \textit{on the order of} (+ NP). For a fuller discussion, see e.g. Biber 2006, Ch. 6.

\textsuperscript{15} Neither extreme—the morphemic level (embracing both cranberry morphemes and bound morphemes such as the \textit{un-} in \textit{unhappy} or the \textit{con-} in \textit{congeal}) and the largest-scale levels such as oaths of allegiance or oaths of office—will be relevant for the present work. But Wray’s caveat about her own definition should be noted: “It is clear from this definition that the term aims to be as inclusive as possible, covering any kind of linguistic unit that has been considered formulaic in any research field” (Wray 2002: 9).

\textsuperscript{16} This would probably exclude cases such as the \textit{Pledge of Allegiance}, which is actually a performative, even if pledging allegiance to a flag is clearly some form of metonomy (as spelled out in the following phrase: “and to the Republic for which it stands”).

\textsuperscript{17} For computationally-oriented linguists, where a formal definition is a \textit{sine qua non} for grammar-based searches, the easily identified verb+noun pattern is the most heavily favored (Fazly, Cook & Stevenson 2009, Wulff 2013, 2009, 2008).
people or region’s language), while sense 4 concerns theology (specifically characteristics of Christ), and 5 refers to forms and expressions characteristics of the language of an individual writer or artist.\(^\text{18}\) Only \textit{OED} sense 3 is at all germane to the present discussion:

3. A form of expression, grammatical construction, phrase, etc., used in a distinctive way in a particular language, dialect, or language variety; spec. a group of words established by usage as having a meaning not deducible from the meanings of the individual words.\(^\text{19}\)

This definition specifically focuses on non-compositionality above the word level (‘having a meaning not deducible from the meanings of the individual words’), but also hints at it being relatively fixed (‘a form of expression’) and possibly having a specific pragmatic meaning (‘used in a distinctive way’). As will be seen, all of these factors are involved in a careful definition of the phenomenon of idioms.

For a thorough discussion of what characterizes idioms \textit{per se}, the seminal article by Nunberg, Sag & Wasow (1994) provides a useful starting point, as they are careful not merely to examine the various aspects required for a definition, but also to indicate the province of each, above all whether the aspect is part of a semantically- or syntactically-based approach. They do not specifically offer a definition, but instead note that a judgment of whether a phrase such as \textit{shoot the breeze} is “a prototypical idiom” will be based on “a number of more-or-less orthogonal properties of the phrase” (1994: 492), namely:

1. \textbf{Conventionality} (the language group’s consensus that the phrase has meaning beyond that of the sum of its constituents,\(^\text{20}\) i.e. non-compositionality in a social environment\(^\text{21}\));

---

\(^{18}\) None of the current \textit{OED} senses listed for \textit{idiom}, \textit{idiomatic} or \textit{idiomaticity} fully captures the use of these terms in ESL or language-learning contexts, where, together with \textit{fluency}, idiomaticity is, (controversially, but usually) assumed to be one of the primary goals of language learners. In such contexts, idiomaticity means something like ‘speaking in the same way as most native speakers do’.

In discussions of ELF (‘English as a lingua franca’), idiomaticity is regarded as primarily characteristic of L1 speakers and not necessarily always desirable (cf. the discussion in Björkman 2010: 34).

\(^{19}\)The corresponding definition in previous (written) editions is considerably less explicit and less focused: “A form of expression, grammatical construction, phrase, etc., peculiar to a language; a peculiarity of phraseology approved by usage of the language and often having a signification other than its grammatical or logical one” (V, “I”: 21–22, sense 4). All references to the \textit{OED} in the present work are, unless specified otherwise, to the current \textit{OED Online}.

\(^{20}\) Cf. their list of (primarily generativist) definitions (1994: 498) that insist on non-compositionality.
Inflexibility (typically restricted in its syntactic frames);
Figuration (speakers sense that some form of figurative language is involved);
Proverbiality (typically describing recurrent social situations via a “homey” scenario);
Informality (typically used in informal and colloquial registers linked to popular speech and oral culture);
Affect (typically indicating an evaluative or affective stance, rather than a neutral use of language).

In most discussions of idioms, the primary features are presented in ways similar to (1) to (3); the latter three dimensions are often neglected and rarely seen as decisive, despite the claim of Nunberg et al. that “we have to appeal not just to the semantic properties of idioms, but to the figurational processes that underlie them and the discursive functions that they generally serve” (1994: 494). Gibbs 1993 (echoed in Nunberg et al. 1994: 496, although without reference to his work) provides an early example for the now generally-accepted position that many idioms, such as spill the beans, can be figurative, yet still retain a conceptual grounding in the literal sense of their components (here to spill something, i.e. drawing on a conceptual metaphor such as MIND IS A CONTAINER). This is relevant to dimensions (1) and (3) in the Nunberg et al. list.

The crux of their multi-pronged definition, however, is that the first three dimensions are seen as formal properties characteristic of idioms, yet turn out to be insufficient to fully define them. Property (2) has above all been of interest in a generative framework.\(^\text{22}\) The aspect of conventionality (i.e. whether speakers recognize a given idiom) is an issue rarely addressed in corpus-oriented studies,\(^\text{23}\) whereas the question of non-compositionality has

---

\(^{21}\) Note that this dimension conflates semantics with the sociolinguistic constraint of the speech community, i.e. most—although clearly not all—speakers should be able to recognize this non-compositional sense of the phrase as its basic meaning.

\(^{22}\) Nunberg et al. consider such aspects in detail in their Section 4. In his major work *Foundations of Language*, Jackendoff briefly considers idioms primarily from the point of view of the difficulties generative theories have with coping with them, and makes the memorable observation: “The overwhelming generalization about idioms, in fact, is that they have the syntax of garden-variety phrases” (2002: 169), although he of course recognizes that they usually do not have the full flexibility of “normal” language. Cf. also the extensive (non-generative) treatment of this issue in Gustawsson 2006.

\(^{23}\) “Conventionality” involves the interesting paradox of much of corpus work: the actual number of instances in a given corpus is often astonishingly low, even for what appear to be quite common items. Thus, the infrequent modal lemma *shall* (i.e. including *shalt*) occurs about 30 times per million words in 2005–09 in *COCA*, or roughly twice as frequently as *pea*, and 10 times as often as one of the most common idioms of all, *to break sb’s heart*, yet all of these items are clearly part of Standard English. Cf. note 6, above.
consistently been cited by scholars of idioms and thus the property generally considered the single most important factor in working with idioms.

An explicitly semantic definition is to be found in e.g. Cruse, who points out the circularity of a standard non-compositional definition such as “an idiom is an expression whose meaning cannot be inferred from the meanings of its parts”:

The definition must be understood as stating that an idiom is an expression whose meaning cannot be accounted for as a compositional function of the meanings its parts have when they are not parts of idioms. The circularity is now plain: to apply the definition, we must already be in a position to distinguish idiomatic from non-idiomatic expressions. 24 (1986: 37)

Instead, he offers a two-pronged definition: (i) it must be lexically complex, and (ii) it is a single minimal semantic constituent. He treats idioms as “elementary lexical units” which “display to some extent the sort of internal cohesion that we expect of single words” (1986: 38); the latter aspect is clearly consistent with Wray’s definition, cited above. Note that the lack of formal criteria and his qualifier “to some extent,” allow for a discussion of idiomaticity as potentially fuzzy, rather than absolute, a stance that does not seem to be controversial (e.g. Jaki 2014, Wulff 2013, Finkbeiner 2008, Finkelbaum 2007a, Warren 2005, Moon 1998). 25

While all idioms require both encoding and decoding, an important distinction between idioms of encoding and idioms of decoding was made by Makkai (1972: 24 ff.). 26 As Taylor notes:

To be sure, a full knowledge of the idioms surveyed above will comprise both the decoding and the encoding perspectives. A speaker not only needs to know the conventional meaning of spill the beans, she also has to know the circumstances in which the expression can be appropriately used. A focus on encoding, however, greatly expands the range of things that can appropriately be said to be idiomatic. This is because the conventionalized way of saying something may not, in itself, be in any way ‘idiomatic’, from a decoding perspective. The idiomaticity resides in the fact

Actually testing whether users of English agree on whether a given phrase is an idiom or not would be the province of psycholinguistics, as noted earlier. Cf. Steen 2007, particularly Chapter 4; the curious reader is further referred to works by e.g. Gibbs or Glucksberg (Gibbs 2005, 2000, Glucksberg 2000, Glucksberg & Keysar 1993). A link to collocational studies should probably also be explored.

24 Cruse’s objections notwithstanding, a number of scholars partially or completely agree with this definition; cf. e.g. the list in Gustawsson (2006: 12). The methods of the Pragglejaz group (2007) appear to be relevant here, as a way of breaking out of the vicious circle. For an alternate approach, see Grant & Bauer 2004.

25 Cf. e.g. Moon, who concludes her discussion of fixed expressions and idioms by noting: “All the above-mentioned criteria [for FEIs] are variables” (1998: 9).

that *this* happens to be the conventionalized way to say something, rather than some other, equally plausible way” (2002: 546).

The burden of decoding is eased by the fact that many idioms are at least partially comprehensible via their components (Gibbs 1993): *raining cats and dogs* uses the standard sense of *raining*. A somewhat less obvious fact is that idioms indirectly announce themselves by introducing apparently extraneous objects and scenarios into the text (cf. Steen 2013): e.g. *cats* and *dogs*, rather than the expected forms of precipitation. This in turn usually indicates that the speaker is using an idiom, which immediately eases the task of decoding, by widening the potential range of meanings. This help is not available to encoders.

This is a crucial observation in any discussion of idioms in an educational context, but also of much wider relevance: to start with, when an idiom is presented in a text, it need only be decoded, based on the form actually instanced. The reader (or hearer) does not need to know anything about e.g. limitations on permitted syntactic variation or combinatorial variation. Thus, if she encounters the American phrase *not touch something with a ten-foot pole*, as in [1],

[1] Do you envision a particular audience when you write?  
I always hope to reach people who don’t want to touch religion with a ten-foot pole. (*COCA*)

a British reader might find the idiom a trifle odd, but not difficult to comprehend, and can easily adjust it to the more British form, as in [2]:

[2] Typically, teaching responsibilities go to graduate students and to a new breed of academic [*’gypsy’- non-tenure-track employees on temporary contracts. They are the hardest-worked, worst-paid faculty at many schools, earning not much more than minimum wage while doing the donkey work of teaching introductory courses and other material that their privileged colleagues wouldn’t touch with a barge pole. (*COCA*)

The key point, however, is that if a speaker wants to express something like ‘not want to be involved in any way’, neither of these idiomatic variants will be available for encoding until actually already available through previous exposure and learning. This difficulty applies to both L1 and L2 learners,

---

27 Despite the clear evidence that idioms are relatively rare, as noted above, they loom surprisingly large in educational contexts, not least in discussions of ESL and ELF (see e.g. Kerbel & Grunwell 1997). For idiom analysis of relevance to general teaching, see particularly Grant & Bauer 2004; idiom teaching in EFL educational contexts will be further considered in Section 1.1.1.3, below.

28 All quotes taken from corpora such as *COCA (The Corpus of Contemporary American English)* will be cited in this fashion.
although there is a greater likelihood that L1 acquisition of the idiom will occur at an earlier age (and not in an instructional environment).\textsuperscript{29} As with all vocabulary, idiom acquisition is a life-long task. When an individual speaker encounters an idiom new to them, there is no linguistic distinction between an established idiom and a new coinage (unless accompanied by a metalinguistic comment such as \textit{the proverbial X} or \textit{as the old saw has it}); the only difference is the fact that the idiom has already achieved sufficient circulation and recognition in dictionaries and popular speech, i.e. become conventional, which is something the individual can only know by checking dictionaries, asking others or being further exposed to the idiom.\textsuperscript{30}

Idiosyncratic or nonce uses can of course become idioms as they win acceptance in wider circles. As a case in point, although recorded as a phrase as early as 1819 (the Earl of Munster), and used as early as 1920 to refer to the Soviet Union, \textit{the iron curtain} did not become widespread until Churchill’s use of it in 1946, whereupon it became firmly anchored in the Cold-War rhetoric that lingers on to our day. From that point on, this phrase must be considered an idiom in English.\textsuperscript{31}

One consequence of the encoding difficulty is that students must learn the idioms already in use, rather than being able to simply create their own. A relatively fully worked-out example of just how much can be unpacked from an idiom’s use in a given genre is provided in Section 1.7, below.

Moon’s full-length corpus-based study of \textit{fixed expressions and idioms} (Moon 1998) is an important work for this thesis, not least because she was the Editorial Manager for the first edition of the \textit{Collins COBUILD Dictionary of Idioms}.\textsuperscript{32} The 1998 study addresses the wider range of phenomena called \textit{fixed expressions}, but specifically provides a far-ranging discussion of the subset specifically designated as \textit{idioms}:

\textbf{Fixed expression}, like idiom, is unsatisfactory as a term, since it will be seen that many fixed expressions of these types are not actually fixed;

\textsuperscript{29} Even so, precisely because idioms are often obscure in their composition (e.g. \textit{kith and kin, hoist by/with one’s own petard}), people often exhibit a metalinguistic curiosity about this aspect of their own or the target language: Thus, in its article on \textit{petard}, Wikipedia (http://en.wikipedia.org/wiki/Petard) specifically mentions the idiom, in addition to showing several pictures of petards and citing the relevant passage from \textit{Hamlet}.

\textsuperscript{30} Langacker’s term, \textit{entrenched}, does not as easily evoke the formal aspects of e.g. dictionaries. For a thorough discussion of the difference between the creation of a variation and its acceptance in the language, see Croft 2000, especially Chapter 4.

\textsuperscript{31} \textit{OED}, \textit{iron curtain}, \textit{n}. As the \textit{OED} rather archly remarks in passing about its use in reference to the Soviet Union, “[t]he locus classicus is quot. 1946”, i.e. Churchill’s use. See also the discussion of this and several other politically oriented idioms in Minugh 2007b, as well as the modern history of the term \textit{African-American} (Baugh 1991).

\textsuperscript{32} This subsection is particularly indebted to the detailed discussion in her Chapter 1.
however, I will retain it for simplicity’s sake. I will hereafter refer to fixed expressions (including idioms) as FEIs. (Moon 1998: 2)

Moon cites a number of scholars who have tackled the definitional issue, particularly Gläser 1984, Čermák 1988, Nunberg et al. 1994, Barkema 1996 and Cowie 1998b, noting, as mentioned earlier, both the broad range of phenomena potentially involved and the range of terms. Her investigation excludes four types of phraseological units from her investigation:

These are **compound nouns, adjectives** and **verbs** such as *civil servant*, *self-raising*, and *rubber-stamp*; **phrasal verbs** such as *make up* and *stick out*; **foreign phrases** such as *fait accompli, che sarà sarà*, and *caveat emptor*; and **multi-word inflectional forms** of verbs, adjectives, and adverbs such as *had been lying* and *more carefully*.33 (Moon 1998: 2)

Turning to **idioms**, Moon begins by noting the confusion of levels and applications seen in the *OED* definitions, but then specifies her definition in terms that by now should seem quite familiar:

Narrower uses restrict **idiom** to a particular kind of unit: on that is fixed and semantically opaque or metaphorical, or, traditionally, ‘not the sum of its parts’, for example, *kick the bucket or spill the beans*. Such units are sometimes called **pure idioms** (Fernando and Flavell 1981: *passim*; Cowie 1988: 133).

After a review of various further restrictions or broadenings of the term,34 she concludes without offering a formal definition of **idiom**: 

While I will not use **idiom** as a formal category, I will make occasional use of **idiom** to refer loosely to semi-transparent and opaque metaphorical expressions such as *spill the beans* and *burn one’s candle at both ends*, as opposed to other kinds of expression. In more general contexts, I will subsume **idiom** within the broader category of **FEI**. (1998: 5–6)

Nor is this surprising: more recent studies such as Simpson & Mendis (2003: 420–21), Gustawsson (2006: 12–13) and Langelotz (2006: 2–5), and those dealing with the conceptual nature and models of idioms, such as Egan (2008: 381–83) and Wearing (2012: 502–03), do not offer a more rigorous definition. One possible way forward lies in the probabilistic model Wulff

---

33 The first two are excluded on principle (although e.g. Makkai 1972 includes phrasal verbs in his study), the latter two in order to limit her study to a manageable length (1998: 3). All these areas excluded from Moon’s study are also excluded from the present study, unless specifically indicated otherwise.

34 She cites grammatically ill-formed items such as *by and large* and non-literal *move heaven and earth* as often-excluded categories, and rejects the definitions used by Makkai 1972 and Hockett 1958 as discussions with much wider definitions.
developed at length to handle V NP idioms (Wulff 2008), although for many idioms, the corpora feeding into it may need to be much larger than the BNC, the corpus she used (although not explicitly mentioned, the frequency data from the BNC in Appendix 8 in Gustawsson 2006: 171–74 clearly indicates this, particularly if we wish to consider use per genre, and not just in the corpus as a whole).  

Two additional aspects are worth considering here. First, although idioms are usually envisaged as multiword units, some of them can quite happily be transformed, even into one-word units: break the ice > ice-breaker (also heart-breaker, jaw-breaker). The canonical, multiword form of the idiom is easily recognized in these words. Secondly, phonological considerations are rarely at issue in such discussions, since so much corpus material is in written form.

Moon’s 1998 study was concurrent with work on the first edition of CCDI, of which she was the Editorial Manager. In CCDI, the discussion of what makes an idiom an idiom (1995: iv) is quite brief, and cites (although not in technical terms): (i) non-compositionality, possibly leading to pragmatic misunderstanding; (ii) metaphorical use; and optionally, (iii) obscure origins of the metaphor. However, inclusion also requires some frequency in modern corpora (the—unspecified—threshold being far lower than for normal lexical items, given that very few idioms are frequent).

In the studies found in the present work, phraseology is a general term for the field embracing FEIs (fixed expressions and idioms) or formulaic sequences, while idioms will be reserved for the traditional, “pure” idioms, such as those found in CCDI. Since the thrust of this thesis is not that of yet another attempt to define idioms via form and/or via semantics, we will not attempt to further arbitrate among these various categories, particularly as they are in part an argument about how narrowly, or widely to cast our net when searching for/accepting idioms. Moreover, as the studies in this thesis

---

35 Wulff’s conclusion (focused on V NP constructions) is worth quoting at length: “The present results confirm the hypothesis that idiomaticity cannot be reduced to non-compositionality but that it is a complex meta-concept comprising semantic and formal information. Beyond that, an adequate model of idiomaticity must license differences in the weightings of the different parameters and/or parameter levels contributing to overall idiomaticity. Also, the model has to be able to accommodate the fact that there is no stable correlation between the different variation parameters. In other words, the model must be able to handle probabilistic information. It has to represent what kind of variation is possible, and how likely it is that this kind of variation is instantiated by any given attestation of a particular V NP-construction” (Wulff 2008: 166).

36 Whether written as one word or with a hyphen, they can be considered as one-word expressions. For the record (as it were), one famous Led Zeppelin song is called Heartbreaker.

37 The point is discussed in Makkai, who deals with compounds as idioms (e.g. bläck bïrd v bläckbird), and touched upon in Moon, where she refers to psychological experiments attempting to disambiguate literal and idiomatic uses of sentence pairs (1998: 9).

38 For a discussion of the justification for the specific choice of CCDI as the reference tool providing the list of (current) idioms, see section 1.5.3, below.
are focused on modern corpus material, we wish to include material still in circulation (current conventionality, as evidenced in a modern corpus-based idioms dictionary). We will therefore use a simple operational definition, their presence as entries in CCDI.

In the studies presented here, idioms were therefore considered as:

**any phrase recorded in the Collins COBUILD Dictionary of Idioms or an obvious variation on such a phrase.**

This criterion has the virtue of being transparent and easily checked. In the four studies in this thesis, the CCDI entries were therefore always used as the reference list when searching for idioms via e.g. WordSmith, and the results are always presented as the primary results.

However, a secondary group of idioms turned up during Study 2: other, similar phrases the researcher also regarded as idioms. Since this group consists of idioms that were not searched for, but found serendipitously, it could have an alarming potential for untrammeled elasticity; it is constrained by three factors, however:

(i) These idioms typically introduce alien objects or scenarios into the discourse, e.g. *the carpet* in the sentence *Maybe future generations will be more prepared to answer their child's question about violence instead of pushing it under the carpet*, where the Columbine massacre, not housecleaning, is the issue (The Snapper Online, Millersville University of Pennsylvania);

(ii) this shift in context scenarios triggers the sense of figuration (Nunberg *et al*’s third criterion);

(iii) these phrases are found listed as idioms in one or more learner dictionaries and/or other idioms dictionaries (Nunberg *et al*’s first criterion, that of conventionality).

A number of these additional idioms are specifically discussed in the studies in Part 2. Whenever they are considered, as a further safeguard, results in these studies are always cited for the main group and for both groups together, so that the reader can discount the secondary group, if appropriate.

It is the primary group, the CCDI idioms, which we will focus on.

---

39 This variation includes not merely active to passive, or change in tense, but also use of alternate scenario-related words (*give the green light* > *get the green light*) or hyponyms (*quaking in his boots* > *quaking in his Nikes*), or transformations of a phrase into a single word (*break the ice* > *an icebreaker*).

40 Only in the most recent of idioms need recourse be had to Internet checks such as the Urban Dictionary (www.urbandictionary.com). Direct confirmation via metainformation may occasionally be found in the text itself, such as *It's an old cliche, but you can't judge a book by its cover because it's what's inside that counts* (The Creightonian Online, Creighton University); this is quite rare, as it introduces a complication into the actual text.
1.1.1.2 Compiling Idioms

Idioms and proverbs have long been recognized as a special form of language: Flavell & Flavell (2006: 281–82) note that in as early as 410 BC Aristophanes calls the Greek original of leave no stone unturned “the old proverb”, with reference to the Delphic oracle’s pronouncement to Polycrates in 479 BC. The medieval Proverbs of Alfred (late 12 c.) continue the Biblical, older Latin and Anglo-Saxon traditions in Middle English. In 1530, Jean Palsgrave included a few idioms in his English-French dictionary, e.g. under the adverb when, he included the ‘phrasis’ (= phrase):

[3] Whan all is doone and sayd, pour tout potiage, a phrasis. (814/2)\(^{42}\)

It is hardly accidental that these phrases should turn up in such an early dictionary, as they are an obvious stumbling block in any form of bilingual learning: even when you master the words individually, their conjoined use is not apparent (i.e. we are dealing with non-compositionality), as Palsgrave’s example makes abundantly clear, and of course the encoding problem exacerbates this. Moreover, if we consider frequency, the English lexical items in this idiom are extremely common, whereas the old term potiage is not even found in standard French dictionaries, again a situation that is not uncommon in equivalency situations.\(^{43}\)

Within the area of proverbs, there has been a long tradition of gathering examples in collections such as the long-running Oxford Dictionary of English Proverbs\(^ {44}\) and the Dictionary of American Proverbs (Mieder 1992). One problem with such compilations, particularly in the pre-corpus era, is that there was no way of checking whether a given (recorded) phrase was

\(^{41}\) Cf. Proverbs in the Bible. For earlier English sources, see e.g. Baugh (1967: 152); for the Proverbs of Alfred, see Arngart 1978. For a review of earlier dictionaries, see Moon 2000.

\(^{42}\) This entry is quoted in the OED Online, to illustrate phrase, n. 2.a: “A small group or collocation of words expressing a single notion, or entering with some degree of unity into the structure of a sentence; a common or idiomatic expression.” Interestingly enough the modern irreversible binomial when all is said and done (Malkiel 1959: 151, n. 43) is given by Palsgrave in the reverse order.

\(^{43}\) “Un ancien sens plus general de ‹pitance› (1296) reste vivant dans la locution figure pour tout potage (1478–1480) ‹en tout et pour tout›” [An old, more general sense of ‘allowance, serving (of food)’ is still alive in the figurative expression pour tout potage (1478–1480) ‘all in all’] [trans. DM]. Rey (1995: II, 1594: 2).

The comment by Jackendoff (2002: 169) about idioms having “garden-variety” syntax also applies to their lexis: most idioms have relatively high-frequency words (e.g. miss the boat); there are far fewer idioms with rare words (e.g. hoist by your own petard, put the mockers on sth).

actually at all current in the speech community as a whole, or at least in a major portion of it (say, British English or American English).  

Translation studies have long been interested in problems of idiom equivalency (or lack thereof), and multilingual idiom dictionaries have been produced over the years, such as Norstedt’s Swedish-English-German-French-Spanish equivalency dictionary (Martinsson et al. 2004). Such dictionaries are of course inevitably parasitic on the major idiom dictionaries of the languages in question.

For the purposes of the studies in this thesis, which are all oriented towards English only, the idioms already available in various reference works proved to be sufficient; the specific choices made are presented in Section 1.4, below, and reconsidered in the discussion section (Section 1.5.3, below). By choosing the Bank of English-based CCDI dictionary as the touchstone for idioms, we ensure that the idioms we examine are in circulation today.

1.1.1.3 Recent idiom studies

One important strand in idiom studies concerns the collection of data for analysis. Early idiom studies such as Makkai 1972 were often based on laboriously assembled collections of examples, with little sense of how frequent

—

45 This will also inevitably exclude collections of proverbs and sayings peculiar to various traditional (i.e. more local) dialects, unless an equivalent of the saying is found in Standard English, as when Burn’s famous The best-laid schemes o’ Mice an’ Men, / Gang aft agley (“To A Mouse, On Turning Her Up In Her Nest With The Plough“ [1785]) appears to have transformed itself into our the best laid plans of mice and men often go awry.

46 Cf. e.g. Nida & Taber 2003 (Biblical translation), and more recently, Pedersen’s work on cultural equivalents in subtitling (Pedersen 2007).


For French, recent such dictionaries include Ashraf & Miannay (1999) and Richard (2011). Similar works are to be found for many other languages, as well as numerous bilingual dictionaries focusing on idiomatic expressions.

It should be noted that in recent years internet-based idiom dictionaries have begun to appear, such as the French government site La bas proverb or the Dictionnaire d’expressions idiomatiques (French – Portugese – French) or The Free Dictionary, section on idioms (“The idiom dictionary is compiled from the Cambridge International Dictionary of Idioms and the Cambridge Dictionary of American Idioms”). However, search engines provided by internet sites are rarely optimal for linguistic research.
or varied they might be.\footnote{Cf. Moon’s comment on the weak position of pre-corpus idiom studies, referring to the critique voiced by Altenberg & Eeg-Olofsson in 1990 (Moon 1998: 46–47). This should not be taken as criticism of earlier scholars: they simply did not have the corpus resources we today take for granted.} Introducing a corpus perspective ensures that we are confronted with a much broader range of data (as noted quite early on, e.g. Fillmore 1992). The Collins COBUILD Dictionary of Idioms (CCDI) (1995, revised 2002 and reprinted in 2011 under the name Collins COBUILD Idioms Dictionary) was the first major idioms dictionary using modern data, even though the data upon which it was based, drawn from the Bank of English, was not publicly available. CCDI plays a pivotal role in the studies for this thesis, providing the data base from which the idiom searches in the studies are drawn.

No subsequent work has replaced it yet, although several additional idiom dictionaries have appeared in recent years. Makkai, Boatner & Gates released the 4th edition of A Dictionary of American Idioms in 2004. The Cambridge Idioms Dictionary was released in 2006, as was the 2nd edition of Flavell & Flavell’s brief Dictionary of Idioms and Their Origins. The 3rd edition of the Oxford Dictionary of English Idioms (2009, edited by John Ayto) now includes an online version.\footnote{Part of the publisher trend towards online, rather than printed versions of reference works. Other examples include Macmillan, who in 2007 released a second edition of the Macmillan English Dictionary for Advanced Learners (the British version), but not of the Macmillan English Dictionary of American English (the American version of the same dictionary), as it was to be web-based instead.} While hardly scholarly, the net-based Urban Dictionary (1999–) is of considerable interest for its up-to-date coverage of slang terms and (wannabe) idioms—and obviously to be used with caution.\footnote{The most interesting aspect of Urban Dictionary is its voting mechanism (echoed in the “likes” mechanism of the new social media), which fairly clearly indicates which terms have a reasonable currency in the younger generation, giving a new dimension to the concept of linguistic “consensus”.

Even though idiom studies concentrate on a linguistic phenomenon, they are of course related to other approaches. Katz 2006 reminds us:

\begin{quote}
As noted by many working in the cognitive linguistic tradition (see Lakoff, 1993; Turner, 1998) metaphor and other figures of speech should not be conceptualized as a linguistic phenomenon or as just a linguistic phenomenon, but rather as a cognitive phenomenon, sometimes expressed in surface language … although much of the ongoing research has emphasized language processing per se, one should be sensitive to distinguishing models based on linguistic expression from those that engage underlying nonlinguistic cognitive mechanisms. (2006: 206)
\end{quote}
Moon 1998 may serve to mark the real beginning of corpus-based analysis of idioms. After a pilot study using the *Oxford Hector Pilot Corpus* (the “small” [18M-word] corpus, the 1998 work drew on the much larger resources in the *Bank of English*. Covering a series of topics important to the present thesis, this work remains to this day the single most important, wide-ranging quantitative study of idioms.

Recent studies involving idioms are thus often intertwined with discussions of metaphors (and metonymy), as in Wray 2002 or experiment-oriented psycholinguistic studies, such as Gibbs (1993, 1995, 2007) and Glucksberg (2000). Steen 2007 provides a detailed review of these areas and how they (may) interact; although the term “idiom” does not even occur in the Index, Steen’s work nonetheless provides an excellent review of the methodological and theoretical issues involved in relating these areas to one another.

For *phraseological units* in larger contexts, Naciscione 2001 particularly traces cases where speakers do not merely use an idiom, but echo and develop it, primarily in literary texts. This approach can link well to the pragmatic functions found in academic speech, as sketched in Simpson & Mendis 2003, based on the Michigan Corpus of Academic Spoken English, an article with explicit links to the teaching perspective.

Also from the teaching perspective, Grant & Bauer 2004 attempt to improve on the definition of idioms by dividing them into several different types, based on intelligibility (which boils down to the degree of learner difficulty). Among the various teaching witnesses to learner problems with idioms, Prodromou 2007 seems fairly typical, but with the newly-won awareness that ELF (‘English as a Lingua Franca’) situations may require adjustments on the part of the native speaker, not just the learner.51 The 2009 article by Soler, Hayes & Palomo in the ESP (‘English for Special Purposes’) area of business English is important, in that it provides genre-specific evidence for a higher frequency of FEIs in business press headlines than corpus results in general: there is, after all, no theoretical reason to assume that any given genre will reflect the overall patterns of a more inclusive corpus (one thinks of the legendary bad puns in British tabloid sports headings).52

---

51 For a fully worked-out ELF perspective on discourse problems with idioms, see e.g. Seidhofer 2009; for a more general corpus awareness in ELF, see the new *Journal of English as a Lingua Franca* (2012–). The core problem of difficulty for the non-native learner of English appears to have become somewhat of a cottage industry in Asia, with articles both about teaching English idioms (e.g. Wu, Chen & Huang 2006) and others comparing English idioms and metaphors with those of other languages (e.g. Fu 2005, Cedar 2008).

52 Cf. Jaki’s full-length study of phraseological variation in newspaper headlines (Jaki 2014).
Variation of idioms is one of the new areas that could finally be tackled empirically with the advent of the BNC in the mid-90s.\textsuperscript{53} The careful user of CCDI will note that variation is repeatedly documented in idioms’ lexical makeup, e.g. under \textit{mile}: \textit{go the extra mile}, they add “People sometimes replace ‘go’ with another verb and ‘mile’ with ‘yard’ (1995: 259).\textsuperscript{54} Book-length studies already mentioned include Gustawsson 2006 (on grammar-oriented idiom-breaking in the BNC data) and Langelotz 2006 (the first major attempt, largely based on Langacker’s work, to integrate idioms into the conceptual metaphor approach and produce a typology of idioms). The use of actual data showing some of the variation that exists opens up interesting questions for analysis: is it harder to vary \textit{a pig in a poke} than, say, \textit{at the drop of a hat or a dime a dozen}? Does the variation (or modification, as in Study 3) bring the idiom closer in line with its context, as Soler, Hayes & Palomo 2009 suggest for business English press headlines? We will return to such issues in the individual studies in this volume.

The theoretical model for the structure of V NP idioms in Wulff 2008, which was largely developed via statistical analysis of BNC material, does have considerable potential for the question of variation and how variation works.\textsuperscript{55} It should be considered in connection with Egan’s insight into the recognition-triggering effect of the canonical form (2008: 399–401), which seems thoroughly convincing: a major limitation on variation is precisely the recognition value that the canonical form has; the greater the deviation from the canonical form (and possible ambiguities, such as a literal meaning), the less likely we are to recognize an intended idiom.\textsuperscript{56} One such ambiguous case, the \textit{one-horse town} idiom, is further considered in Section 1.3.5, below, based on the discussion in Study 4.

Although the present work focuses on English idioms and corpora, one useful work covering several aspects of corpus investigations of idioms in German\textsuperscript{57} is the volume edited by Fellbaum (2007a), particularly Stathi 2007 on adjectival modification (relevant for Study 3) and Hümmer 2007 on anal-

\textsuperscript{53} In addition to their lack of empirical data, earlier approaches based on generative grammar were not predisposed to take variation seriously, as that threatened their non-compositional interpretation of what idioms were, a point discussed by e.g. Nunberg, Sag & Wasow (1994).

\textsuperscript{54} A further example of variation in that particular idiom occurs in the editorials corpus examined in Section 1.6, below, where it occurs as \textit{go the extra inch}.

\textsuperscript{55} A quite recent work that I have not yet been able to obtain is Schröder 2013, which appears to be in this same area. Wulff 2013 explicitly links her approach to Fillmore’s Construction Grammar approach (cf. Fillmore, Kay & O’Connor 1988, the famous \textit{let alone} article), indicating how idioms can be absorbed into the CG framework.

\textsuperscript{56} However, this approach can be problematic in EFL contexts: cf. Pitzl 2012.

\textsuperscript{57} This material dealing with German may serve as a reminder that work is of course being done on idioms in other languages; the present thesis does not pretend to be able to survey them, as well. Since, however, the modern corpus tradition developed in English, it may be assumed that idiom research using corpora in other languages will have much more to say in the future.
ysis of multiword units [MWU], particularly their contextual modification (relevant for Section 1.6, below). Both accounts do not merely confine themselves to grammatical variation, but tackle semantic questions as well, which brings them more in line with the present work.

1.1.2 Corpus linguistics and idiom studies: An expanding universe

First, a caveat: this bag, while not quite empty, does not have as many cats as did the man from St. Ives: this section is not intended to serve as an overview of this rapidly expanding field, merely to position idiom studies vis-à-vis this crucial source of material.

Almost from day one, the history of corpus linguistics has been inextricably linked with the phenomenal development of computer processing power and memory. In consequence, pioneering textbooks on corpus linguistics (Bauer 1994, Stubbs 1996, McEnery & Wilson 1996, Thomas & Short 1996, Ooi 1998, Biber et al 1998) are somewhat outdated in the size of the corpora they discuss, although not in their conceptions of search methodology, or what the linguistic implications of their results might be. More recent introductory works include Meyer 2002, Stubbs 2002, McEnery, Xiao & Tono 2006, Lindquist 2009 and McEnery & Hardie 2012, as well as chapters in linguistics survey textbooks (e.g. Culpeper et al. 2009). A useful recent overview for the scholar is O’Keeffe & McCarthy 2010.

For the purposes of the present study, the network of researchers working within the ICAME (International Computer Archive of Modern and Medieval English) framework, with its conferences, journal, corpus archiving and publications, has provided steadily evolving methodological training and an ever-ready forum for corpus work by linguists. Numerous sites of relevance have become available, including the VARIENG site currently hosted by the University of Helsinki. A useful review of the key concepts of representativeness, ‘balancedness’ and comparability may be found in the overview in Leech 2007 and works cited therein. These concepts revolve around the question of how a given corpus relates to the language as a whole, and what kinds of conclusions may be drawn about both the corpus and the language as a whole, a matter of major relevance to most corpus studies, including the present one.

58 “Moore’s Law,” the famous suggestion that computer power (number of transistors per chip, etc—interpretations vary) would double every 18 months has hitherto been proven correct. Happily, it is becoming less and less important to remember that the computing power available to individual researchers and students can be a limiting factor; rather, it is access to programming and statistical methods that looms largest today.

Although the large corpora of recent years have come to dominate corpus and lexicographic work as a whole, advocates of “small” corpora maintain that they also hold a crucial place in providing data, not least for genre and discourse studies. The relevant chapters in O’Keeffe & McCarthy 2010 and the article by Vaughan & Clancy (2013) provide useful overviews of further issues involved in creating and using small corpora. Since even a relatively large corpus such as the BNC, at 100M words, will have a fairly small amount of most genres or discourse types, in effect serving as a collection of small subcorpora, independent small corpora continue to provide a valuable addition to the data available for such studies. The increasing size of modern corpora provides us with ever-greater opportunities to investigate rare phenomena such as idioms, but does not obviate the use of “niched” corpora, particularly when, as in the case of the Coll corpus, the material is not found in the major corpora.

Although recent idiom studies have nearly all drawn upon corpus data, particularly from the BNC (Wulff 2008, 2013, Gustawsson 2006), they have not yet dealt with the twin problems of collecting sufficient instances of idiom use and relating this large collection to the fact that each individual instance was used in a specific context (cf. e.g. Wray 2002: 26–28).

While the studies presented in this volume will not solve the first of these problems, they do present idiom material from specific and more general sources, two of them allowing close study of numerous idioms in specific corpora that represent well-defined genres of English, and one examining a specific pattern of idiom modification in a number of different (but largely parallel) corpora.

The lack of a corpus-data foundation is clearly evident in the theoretical discussion underlying Study 3, the study most directly connected to ques-

---

60 An obvious starting point for genre awareness would be work by John Swales, e.g. Swales 1990. For studies with a direct link to small corpora, see e.g. Ghadessy, Henry & Roseberry 2001. Obvious further examples of specialized corpora include the multilingual CHILDES corpus of children’s language (at http://childes.psy.cmu.edu/), the MICASE corpus of academic spoken English (at http://quod.lib.umich.edu/m/micase/) and various historical corpora of early texts in e.g. medicine. Individual corpora are invariably described in detail by the author(s).

61 Due to collection and processing difficulties, spoken (sections of) corpora inevitably lag behind in terms of sheer numbers. The explosion of interest in spoken academic English (see Biber 2006, and works stemming from the MICASE corpus) is an interesting exception.

62 The largest currently available corpus, the BYU Google-US corpus (at more than 150B words) achieves its size by using printed material from several centuries, so even there a synchronic study would have access to a much smaller amount of material than it would appear at first glance.

63 Briefly: Study 3 focuses on adjectival modification of idioms, specifically “textual anchorings”, as in *[T]*hey were Dutch refugees from the Nazis, but they did not play the *refugee game* [*Time* 1940], where *refugee* links back to the surrounding text, and is not part of the idiom itself, and does not easily collocate with the idiom’s key terms, *play* and *game*. 

38
tions of grammar: Ernst 1980 first poses the problem, but works entirely with intuited examples, rather than actual ones (not surprising, given the early date). However, even Nicolas 1995 uses only a highly restricted set of corpus data (from the Wall Street Journal) to discuss the phenomenon as such (as opposed to the phenomenon in the WSJ). Not until Study 3 (Minugh 2007a) and Stathi’s article (2007) do corpus examples assume a prominent place as the material to be accounted for.

Since these studies were published, the arrival of the new, larger corpora, both those that are purely Internet-based (in the billions of words) and those that cover multiple written sources (e.g. COHA and COCA), will allow us to ask these questions in much greater detail, with the opportunity to make large-scale analyses of individual idioms. Corpora with large samples (the BNC uses 2000-word chunks) are to be preferred over concordance lines, and best of all for pragmatic analysis are corpora that provide texts in their entirety. In this respect, the 100M word Time corpus is perhaps unique, in that every single issue is indirectly available through this search engine, or directly through the Time Archive (http://content.time.com/time/archive/).

A final point. This thesis is not about the language of fiction and other imaginative writing, but such writing does of course form part of the material in the various corpora used here. And for the specific case of imaginative writing by the modern American writer Richard Brautigan considered in Study 3, the thrust of his use of invented idiom Trout Fishing in America is one that makes relevant ideological considerations such as those dealt with by e.g. Charteris-Black (2005) and Goatly (2007).

1.2 Research Aim and Questions

*Take another little piece of my heart now baby!*  
— Janice Joplin

Idioms are both striking and relatively infrequent: they are easily noticed (not least when we first meet them), yet it is difficult to find many examples of a given idiom, let alone to decide when and how to use it. With Sinclair’s 50-examples-per-sense dictum in mind, I found it intriguing to test various corpora for the occurrence of idioms. This in turn led to a growing awareness that they were far more flexible in form and in extension than the generative approach (the non-decompositional, “look up exceptions” stance) and canonical dictionary entries would lead us to believe. The overall aim of the present studies may be stated as follows: through the use of corpus-mining methods, to investigate a series of specific instances and patterns of idiom usage, with respect to variation, both in the data and in the usage itself.

---

64 “Piece of My Heart”, from *Cheap Thrills* (Big Brother and the Holding Company, 1968).
Since a cardinal rule in corpus studies is to know your corpus, the first study in this thesis concerns the creation of a corpus. At the time when the Coll corpus was created (1999), there existed only one freely accessible major corpus, the British National Corpus, which by its very nature concerned only the English of part of the British Isles. Data from the rest of the English-speaking world was lacking. Working as I was with Swedish college students learning to write English, it seemed natural to wish to have access to the writing of their peers in the English-speaking world. Learner corpora such as ICLE (the International Corpus of Learner English) were just getting underway, and in any case could not provide the native-speaker data, particularly from North America, that I wished to consider. At the same time, internet use was beginning to take off at colleges, so the creating of a niche corpus seemed a reasonable way forward. The research question for this first paper may thus be formulated as follows:

(1) How does the data in a coherent, moderately-sized corpus of English written by young people, based on the English-language online college newspapers (then) in existence, compare to data from other then-existing corpora?

In other words, how might the Coll corpus, once created, be used for linguistic studies of various phenomena, and does it exhibit some unusual and/or skewed properties due to its composition, or to the nature of the data? The size of the resultant corpus was actually determined by (i) what college newspapers were online, and (ii) how many issues of each were available, at the time a completely unknown quantity (see Section 1.3.2, below). In order to answer this question, Study 1 details the creation of such a corpus and preliminary findings about its contents, as compared to other available moderately-sized corpora (the Brown family of corpora).

For the other papers in this collection, where already existing corpus/corpora, once selected, must be taken as given, the primary research questions concerned the idioms to be found therein:

(2) How frequent are idioms as a whole in the selected corpora?

---

65 The monumental difficulties—and success—of such a project cannot be overestimated, as may be seen by the fact that a matching US corpus has yet to be produced, the closest currently being the COCA corpus. The Santa Barbara Corpus (of spoken American English), which was meant to be part of the never-realized “American National Corpus”, was just being created (Chafe, Du Bois & Thompson 1991), and even at its current size of 0.25 M words (like the other major spoken corpus, the 0.5 M word London-Lund Corpus), is too small for most idiom studies. The Brown family of corpora, like the ICE family, are all only 1 M words each. Finally, the Bank of English (like the recent in-house corpora produced by major dictionary-makers) is not publically accessible in its entirety.
(3) Are certain subsets of these idioms more common in different parts of the corpus? This comparison can be either geographical or diachronic.

(4) How do these idioms vary in their realizations and textual uses?

To address these questions, the second study examines the idioms in the Coll corpus, i.e. the corpus examined in the first study, composed of the language of young native-speaker writers writing for their peers on the Internet. This especially created corpus was an unknown quantity, in that no similar corpus existed (or exists today). The newspaper CDs of the 1990s could be used as somewhat similar corpora; however, each such CD normally consisted of one year’s output of a specific major national newspaper, which has the advantage of reaching a large segment of the reading public, but the disadvantage of being produced by a relatively small number of individual professional writers and editors.66 The Coll corpus, by contrast, has literally thousands of young contributors spread across the English-speaking world.

The third study focuses on the fourth question, using several different corpora to specifically consider how idioms (which normally are self-contained) may include links (textual “anchorings” or “hooks”) back to the textual and conceptual world in which they are embedded, specifically via adjectival premodification of the main noun in the idiom. To obtain sufficient data for a quantitative study, five different corpora in the 50–100M word range were examined (thus covering both British and American English, with a sizable component of spoken English) for the distributional patterns of idiom modification.

The final study looks at how frequency and variation of idioms may be considered in a diachronic perspective, focusing on one highly specific genre, general-audience American weekly news magazines, as represented by Time, over some 80 years. In particular, by breaking the data down into only two time sequences (1923–1959 and 1960–2006), a sufficient number of idioms can be accumulated to search for change in usage patterns in the 20th century, at least at the aggregate level. By using selected idioms, we may also possibly detect frequency changes that may be related to changes in society itself: one specific example will be whether biblically-based idioms change in frequency over time.

The basic issues involved in these four studies thus concern first the creation and exploration of a corpus that fills a gap in our records of different varieties of English, together with the specific investigation of idiom usage in that corpus, followed by further exploration of how idioms vary in frequency and contextual usage in recent corpora from different geographical

66 One cautionary tale among many: In the USA Today CD from 1990–1994, one sports writer was responsible for 10% of the occurrences of notwithstanding, all of them as a postpositional (as used by other writers, it split as 12% post- and 88% prepositional); cf. Minugh 2002b.
areas and over time. Moreover, the range of idioms discussed will not be restricted to any given grammatical structure or semantic group, but include all types found in the *Collins COBUILD Dictionary of Idioms*.

### 1.3 Method

#### 1.3.1 Overview

The unfolding opportunities provided by the successive release of new corpora and my earlier work on idioms triggered this series of studies in a more cascading fashion than would be ideal. In consequence, the methods used, as described here, must be described and seen together more retroactively than proactively. This is in part unavoidable, as new corpora (not to mention the Internet) emerged unpredictably, providing new material to investigate and analyze. Without these changes, the *Coll* corpus could not have been created, nor much of the subsequent work carried out. Taken together, these studies bear witness to an evolving sense of what we can learn about idioms through the creation and use of corpora.

In corpus research, two questions of fundamental importance need to be addressed before moving on to the analysis of the data:

1. The composition of the corpus itself;
2. The methods used to find data in the corpus.

These are fundamental and crucial aspects that delimit and limit the material for the analysis. Corpus queries can only be successful if the corpus contains the material we wish to ask about: trivially, the *BNC* cannot say anything about Australian English, as there are no samples of Australian English in the *BNC*; nor can the *BNC* inform us about 19th century British English, as it is a synchronic corpus. More subtly, the *BNC* does not have specific samples of caretaker interactions (e.g. nurses and patients in a geriatrics ward), and therefore is ill-equipped for queries about this genre.

As for the second aspect, retrieval software only finds what the query asks, not necessarily what the researcher intends. When extracting data from a corpus, the two classical problems are those of false negatives and of false positives (type I and type II errors, respectively): the search procedure can either fail to find a given string (false negatives) or find too many examples (false positives); the latter case is usually less important, as false “hits” can usually be weeded out by manual inspection (although *play the game* ‘play fairly’ is almost invisible among all the other references to people taking part in various games). The former is a much more serious problem, as the only
way to recover instances the search has missed is by enlarging the search procedure, which often conflicts with constraints of time, in addition to frequently producing further large quantities of false positives. For idiom studies, false negatives often consist of non-canonical instances that are nevertheless recognizable to human intelligence, but not as a computer search string: they may be only partial quotes (the fat lady, as opposed to *It ain’t over till the fat lady sings*), or misspelled (*More than Meats the Eye* [The Sun, cited in Jaki 2014: 21]) or have rhyming (*A friend with weed is a friend indeed*) or semantically linked (*Put on your Doc Martins!*) substitutions.

In the present thesis, the various papers have dealt with these questions in somewhat different ways, so that the answers will vary with the paper concerned.

The first question is of direct relevance for Study 1, “The Coll Corpus”, which reports on the corpus created by the researcher. For the other three studies, this question becomes one of judging existing corpora (or subsets thereof), given the purposes of the investigation, as it is not possible to influence their actual composition. The second question is not without relevance for the first study, but is much more important for the other three, as they are primarily concerned with the idioms to be found in the respective corpora. It should nonetheless be remembered that this work is not focused on data extraction, knotty as that occasionally may be, but on the data itself.

When creating a corpus, the primary question is that of selection and potential skewing. Since no corpus of a living language can ever even remotely approach being complete (unlike, say, a corpus of extant Old English or Classical Latin texts), such a corpus is necessarily a subset of the language in question. How the selection is made will clearly have an impact on the extent to which the corpus results can be considered generalizable to the language as a whole. Nearly all currently-available corpora are skewed, in the sense that they primarily (or solely) are based on written data; even when spoken data is included, it frequently reflects at least partially scripted material.67 The corpora used in the present thesis are no exception to this general restriction.

On the positive side (for linguists, at least), most such corpora have been created by linguists for linguists, and are therefore relatively balanced in their sampling techniques, ensuring a good representativeness of domains, styles, and so on (cf. e.g. Davies 2012, 2010 on COCA, Gilquin & Gries 2009, Hilpert & Gries 2009, Tognini-Bonelli 2001).68 This is only partially

---

67 “[W]e at Longman have always felt it essential to have a principled approach to corpus design, rather than following the more opportunistic path, which is usually overdependent on newspapers in electronic form or scripted or semiscripted radio broadcasts” (Summers 2006).

68 The near-mythical ANC, the American National Corpus, currently at www.anc.org (where the free-access part is called the Open American National Corpus), could provide a fascinating study in corpus creation nightmares. Started before the BNC was released (Chafe, Du Bois & Thompson 1991), it has still only reached 22M words (15 M in the OANC), and appears to
true of the corpora used in the present work’s studies, particularly the *Time Corpus*, which is manifestly not representative of anything other than itself and the American journalistic domain it forms an important part of (together with its post-1932 twin *Newsweek*).\(^6^9\) However, it has other virtues which make it suitable for its task.

For studies 2–4, which primarily involve the study of idioms, research has long been hampered by a search for ways of identifying idioms through either formal or semantic means. For corpus linguists, who primarily work with unparsed and semantically untagged material, the latter approach is rarely applicable.\(^7^0\) However, formal approaches have not been particularly successful, either. The fundamental hallmark of idioms, non-compositionality, is not a formal characteristic, and therefore not searchable. To take a simple example, consider the identical utterances by B in utterance pairs (1) and (2):

(1)  A: Where are you?  
     B: I’m at sea.  

(2)  A: How’s it going?  
     B: I’m at sea.

Utterance pair (1) is probably a request for speaker B’s current physical location, with the answer that s/he is somewhere out on the ocean, whereas pair (2) is probably a request for the status of his/her work, with the answer indicating that speaker B feels lost or confused about how to proceed. Additional material can at times make this clear, as in (2’):

(2’) A: How’s it going?  
    B: I’m all at sea.

Here, the adverbial *all* is hardly possible if the sense is ‘physical location’, but is quite appropriate for the sense ‘feeling lost and confused’. However, the problem is that such markers are *optional*—both (1) and (2) are completely adequate responses, and are identical in form. Nevertheless, if *I’m at sea* is retrieved, it can be a type II error (a false positive); tokens

---

\(^6^9\) Cf. the strong emphasis Leech gives to exactly this point in Leech 2012.  
\(^7^0\) One thoroughly parsed exception is the British component of the *ICE* corpus collection; the other corpora vary in the extent and types of tagging, with the most recent corpora increasingly being tagged relatively efficiently, thus allowing much more structured queries.
where the literal (non-idiomatic) sense is involved must be weeded out by inspection.\textsuperscript{71}

The situation is more complex for type I errors, as idioms turn out to be more variable than was originally thought.\textsuperscript{72} Specific examples of this variability are discussed in all four papers, and they all require a balance between searching for a given string and searching for strings with increasingly comprehensive “wild cards”, with a steadily decreasing rate of return for more and more extreme variants. In addition, the further the variant is from the canonical form, the greater the risk that the researcher is making connections that other readers will fail to see.\textsuperscript{73}

It is not least for this reason that the choice of the \textit{CCDI} list was made, as providing a large, stable base for searching for idioms. An apparent difficulty is that a number of these phrases also have a literal meaning (Moon 1998: Ch. 7). In recent empirical work, for example, Fazly, Cook & Stevenson 2009 examined 60 VNICs (verb+noun idiomatic combinations) found in the \textit{BNC}, such as \textit{hold (your) fire}, \textit{shoot the breeze}, \textit{make a face}, noting that “close to half of these also have a clear literal meaning; and of those with a literal meaning, on average around 40\% of their usages are literal” (2009: 63).\textsuperscript{74} Disambiguation is, however, rarely a problem, as their collocational patterns are normally quite different; in the present studies, the disambiguation is done manually, not automatically.

If one takes the context into account, extremely few of these examples turned out also to have a plausible literal meaning, so that in practice the search did not produce large numbers of ambiguous readings (exceptions and minor difficulties will be discussed subsequently). As a useful consequence, the results of these searches can be directly compared to those discussed by Moon (1998) and others using a similar approach.

\textsuperscript{71} Greater intersubjective reliability can be obtained by having more than one judge. This is, however, rarely a major issue for identifying idiomatic uses. Cf. the Pragglejaz Group’s work on identifying metaphors (Steen 2010, Pragglejaz Group 2007).

A more difficult problem is where other factors, such as a specific spoken intonation, can produce an idiomatic reading: for example, \textit{Well, excuse me!} can be said with emphasis, meaning roughly ‘I strongly disagree with your statement or wish’, and approaches idiomatic status with the extension: \textit{Well, excuse me for breathing!} \textsuperscript{72}

\textsuperscript{72} Cf. e.g. Wulff 2008, Gustawsson 2006, Croft & Cruse 2004: 229–47.

\textsuperscript{73} Cf. Egan 2008: 399–401. To illustrate the problem with a lexical example, while most people appear to connect \textit{breath with breathe}, and \textit{life with live}, the similar etymological link between \textit{grass} and \textit{graze} appears to be not at all as transparent. Cf. the discussion of variants found for \textit{a one-horse town} in Paper 4, below. Also, as noted earlier, this question can be quite problematic, or even counter-productive, in EFL contexts (Pitzl 2012).

\textsuperscript{74} Given the title of the present work, it is pleasing to note that Moon (1998: 181) cites two studies of informant perceptions of phrases with both literal and idiomatic readings (Popiel & McRae 1988, Cronk \textit{et al.} 1993); both rated \textit{let the cat out of the bag} as the phrase which informants perceived as having the strongest distinction between frequency and familiarity of the literal and idiomatic meanings.
Let us now turn to the individual studies.

1.3.2 Creating a corpus: *The Coll Corpus*

Paper 1, *The Coll Corpus*, details the creation and potential uses of a corpus of web-based college student newspapers. The technical details for creating the *Coll Corpus* are explained in detail in Paper 1, and need not be rehearsed again. Several points about the corpus design need, however, to be considered.

First of all, *Coll* was created to fill a data gap that other corpora available at the time did not address. In terms of geographical distribution, this corpus was largely American in origin (some 77%); at the time, there existed no large-scale contemporary corpus of American English (although the 1M word *Frown Corpus* released in 1999 [the same year *Coll* was compiled] at least partially filled that gap). More importantly, no corpus of writing by younger adults existed, either; the relatively uncensored writing found in student newspapers presented a significant untapped resource for examining the language of the rising generation. Looking ahead to the present (2014), no similar corpus has been created as yet.

The question of balancing the corpus was less successfully dealt with, for two reasons. First, the corpus was collected over a short time period, just before the end of the spring semester; collection over an entire academic year would have produced a thematically more balanced corpus. Secondly, many of the colleges included had relatively few issues to choose between, or even only one trial issue, while others had updates every weekday, so that they were inherently unequal. In the interest of balance, the decision was to take one issue from each newspaper, rather than simply collecting all available material. This of course also had the consequence (and advantage) that no single editor or writer would have a significant impact on the corpus as a whole. Even so, a given newspaper issue could vary from more than 30,000 to only 12 words.

The question of inclusiveness vs. representativity was in principle the same as that faced by e.g. *The Bank of English*, where the problem was solved by the creation of a so-called monitor corpus (Sinclair’s term, cf. Clear 1987, Sinclair 1991: 24–26). However, the *Bank of English* had a nearly inexhaustible supply of almost any written genre at their disposal;76 although the *Coll Corpus* in 1999 had relatively few online student newspapers with daily, or even weekly, issues, to include them all would nevertheless

---

75 This gap also existed for other non-British-Isles dialects. The *ICE* project would ultimately provide 1M word samples of English world-wide (Greenbaum & Nelson 1996), but was largely unavailable then.

76 Based on the information from Mark Davies at [http://corpus.byu.edu/coca/compare-boe.asp](http://corpus.byu.edu/coca/compare-boe.asp), it would appear that the *Bank of English* ceased to be updated a few years ago.
very rapidly have turned it into what amounted to a thoroughly American corpus.

As it was, the slight geographical imbalance (see Paper 1) was a direct consequence of the relative paucity of non-American sites on the whole. On the brighter side, American Title IX demands for equal investments in female sports resulted in a more equal gender distribution of sports reporting than in regular newspapers.

The resultant corpus, all .txt files, is fully searchable via e.g. WordSmith software. A minimal amount of metadata is available through the file names and headlines: they are organized by newspaper, (state) and country, and can usually be separated into subject areas. The original intention was to feed the corpus into a better structured database, but copyright restrictions were a potential problem, and the legal status of internet material unsettled at the time, as it still is. No paper approached by email objected to its material being used in the corpus; the vast majority simply did not reply, and those few that did willingly gave permission, although not enough for publication at the time.

Turning to the actual data in the corpus, the primary difficulty in using its material was the question of what this corpus could be compared to. The then-available corpora of near-equivalent size were primarily the Brown series (Brown, Frown, LOB, FLOB, at 1 M each), or the much larger BNC (or at least the 90%, or 90M words, that represented written English), which was British only. All of these corpora covered a far broader range of domains and genres, and nothing specifically by people in this age bracket. There also existed a series of annual newspaper CDs (The Independent, The New York Times, The Los Angeles Times, all in the 50M range, and with a relatively good match for genres and domains, but all written by relatively few people, all of whom were professional reporters and checked by even fewer (and presumably older) editors and sub-editors (who still existed in the 1990s). In other words, no direct equivalent to Coll existed, a fact exacerbat-ed by the irreverent and/or self-mocking attitude that turned up in numerous Coll texts, far more so than in other corpora, particularly Coll’s closest equivalent, regular newspapers.

Given these limitations, it could not be a simple matter to draw general conclusions from the Coll figures. In addition, just under 4M words is

---

77 Historically, a similar imbalance between languages of the world has always existed in corpus linguistics, with English corpora being at any given time larger by several orders of magnitude than equivalent corpora for other major languages, and with many languages totally without such resources even today.

78 For an introduction to this contentious gender issue in American colleges and universities, see e.g. Lopiano 2000.

79 Technically, the comparisons were not difficult: Searches were carried out via WordSmith (for the Coll corpus and the Brown series), or by the corpus’ proprietary search engine for the
simply not enough material for many lexical comparisons, particularly if the item in question is infrequent in college domains (e.g. a search for nap-py/diaper produced only a handful of instances in Coll, not because the search did not find them, but because such matters were rarely discussed in college student newspapers, a variation on the “toothbrush effect” (Volodina et al. 2013), but is even true for items that could have been expected to be much more frequent in both college and general domains, such as date rape (completely missing from the Brown corpora series). Thus, many traditional geographical comparisons could not be carried out; others, such as the relatively frequent modal verbs, can provide robust figures. Morphological and grammatical matters can generally also be relatively successfully compared: pairs such as older/elder and while/whilst were considered; the way construction produced 231 examples from Coll, which is only marginally less than those in the BNC (matched for an equivalent corpus size, the BNC would have had 237 tokens).  

1.3.3 Examining an entire corpus: The College Idiom

For this follow-up study, the Coll Corpus was utilized as the raw material for searches of various types of idioms, as used by the student online newspapers that comprise the corpus. The question of corpus composition is therefore not of immediate relevance here. Instead, the primary design issues were:

(1) Which corpus/corpora to draw on to answer the research questions;
(2) Which idioms in the selected corpus/corpora to search for and analyze;
(3) Which specific strings to search for to operationalize these research questions.

As regards the choice of corpus, given that the goal of the article was to examine the unsolicited writing of younger adult native speakers writing for an audience, the Coll corpus was (and still is) the only corpus available for English; nearly all the writers were in the 17–23 age group, and they covered a wide range geographically. The corpus size (3.7M words) was far from sufficient for idioms, as many of the idiomatic phrases searched for were only rarely represented, if at all, but they were new data genuinely used by

---

other corpora. Given the very different sizes of the corpora involved, numbers were normalized to per 1 million words, except where indicated otherwise.

Such searches border on the tedious, due to the number of false positives. The search can be narrowed down by restricting it to the string [poss.pron.]+way, but still must be examined item by item, to remove false matches such as lost her way.
the target group of writers, and the metadata was sufficiently detailed for an examination of e.g. geographical distribution.\footnote{Corpus size does matter: “Nearly one third of the idioms in [CCDI] occur less often than once per 10 million words of the corpus” (CCDI 1995:v), a point reiterated in Moon 2005. In Gustawsson’s investigation of 300 “verbal idioms” in the entire BNC, only some 33 occurred 50 times or more, which suggests that we would need a corpus on the order of 10 to 100 billion words to truly fulfil Sinclair’s criterion.}

Turning to the question of which idioms to search for, we have already mentioned that the identification of idioms by formal or semantic means is not a simple problem.\footnote{A useful relatively technical discussion specifically about searching for idioms (based on examples from German) may be found in Herold 2007. For practicing teachers who are budding corpus linguists, a detailed hands-on presentation of the technical problems involved in searching for idioms may be found in Grant 2005 (a follow-up to the approach in Grant & Bauer 2004). A relatively extreme example: if you search through Coll for ass (or arse), alone or with kick*, you will probably miss this variant: \textit{Are we forgetting that The United States of America kicks some serious a**?} (California State University, Chico State, \textit{The Orion Online}).} Instead, for the purposes of this investigation (as in the thesis as a whole and my previous idiom studies; see p. 31, above), the definition of idioms is an operational one: to qualify as a searchable idiom, the (canonical form of the) phrase had to be among the idioms found in \textit{CCDI}, (2002, 2\textsuperscript{nd} edition).\footnote{The “canonical form” is simply the form listed as the main form of the idiom, as defined by \textit{CCDI}. In terms of frequency, the most salient form of the verb need not be the infinitive, which is what \textit{CCDI} lists by default. Cf. Wulff 2008, Ch. 4.} By using \textit{CCDI}, a relatively large, modern well-known baseline could be established for these searches, with the caveat that \textit{CCDI} was a selection made on the basis of the material in the \textit{Bank of English} in 2002, then about 220M words.\footnote{\textit{CCDI} deliberately excluded phrasal verbs and fixed expressions such as in fact and excuse me. See the brief, but enlightening discussion in \textit{CCDI} 1995: iv–vii and the much more thorough discussion in Moon 1998.} The strings searched for were designed to include all specified variant forms of the idioms, and occasional subdivisions into separate subsenses.\footnote{For example, \textit{CCDI} lists \textit{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”. Here, the corpus context was not always sufficient to disentangle these partially overlapping meanings.} Each \textit{CCDI} idiom was then searched for in turn via WordSmith software, with a generous use of wildcards to cover variants.\footnote{For example, \{land/lands/landed/landing on, fall/falls/fell/falling/fallen on\} within three words of \{feet/feet/feet/feet\?\} would cover the expected range of variation for \textit{fall on your feet} in WordSmith; in subsequent searches via Word, the punctuation could be ignored.} It was then necessary to review all the results, in order to categorize them for variation such as geographical origin, type of text, and...
so on. Duplicate hits (such as the second hit for an eye for an eye) were eliminated. For phrases with both a literal and a metaphorical sense (e.g. tear someone limb from limb), literal uses were disregarded, except where the metaphorical sense was also invoked at the same time.

One specific problem area considered in the paper is that of proper names that draw on idioms, such as a coffee shop called The Common Ground, or references to the film called Drop Dead Gorgeous. Particularly items in the public discourse at the time (what today would perhaps even be called memes or be said to be going viral on the Internet), such as The Full Monty can in one sense be said to be merely literal, a reference to a film and no more, but at the same time these tokens are part of what keeps the idiom alive in the language. For this reason, they have been accounted for separately, rather than merely being absorbed into the aggregate data.

A much more difficult problem area is non-CCDI idioms accidentally found in Coll. To take only one striking example, good old/ol’ boys is a phrase with major resonance throughout modern American history, embodying as it does both the American South’s racist past and a potential for a deep sense of speaker-reader solidarity (depending on how it is used). It is probable that the corpus evidence upon which CCDI was based was insufficiently global, so that this Americanism was missed. However, that can scarcely be said of another idiom found eight times in the Coll corpus, to take the reins, which the OED cites from British English from at least the 1600s on, and which is also absent from CCDI.87

It would have been possible to just ignore all such “extra” idioms, but given that the overall extent of idioms in modern corpora was inherently interesting, another approach was chosen. The results from the searches of the Coll Corpus were divided into:

(1) **standard** idioms (essentially the CCDI canonical forms or specified variants);
(2) **variant** idioms (e.g. to gather dust, rather than the canonical form cited in CCDI, to collect dust; or to clam up, as opposed to canonical to shut up like a clam);
(3) **other** idioms (e.g. the long arm of the law, an idiom not found in CCDI, but confirmed as such in other dictionaries, in this case The New Oxford Dictionary of English 1998: 90).88

Group (3) thus consisted of a number of “missing” items found while searching for the “given” idioms of CCDI; in all, they comprised just under 9% of the total number found in the Coll corpus. The relatively large size of group (3)—which was found serendipitously—suggests that further examination of

87 *OED online, rein n, 1, sense 2b, fig., “Freq. with verbs of holding, taking or transferring”.
88 For a discussion of this group vis-à-vis the CCDI idioms, see 1.1.1.1, above.
the corpus, using other keywords, should produce additional examples, considerably boosting the total numbers of idioms in *Coll*.

It is important to note, however, that no claim was being made that this tripartite division reflected a division into types, nor of what is possible: many of the “standard” idioms presumably have variants that were not recorded in *Coll*, and the “other” idioms are “other” only by virtue of their not being in *CCDI* (which is an indirect claim about their lack of frequency—in the then-*Bank of English*, at least).

The results of the searches were further examined for various factors, such as geographical location, size of the newspaper, type of article, position in the article, use in quoted texts (frequently indicative of a spoken origin). They were also compared to “regular” newspaper corpora (on annual CDs). Comparisons were also made to other aspects of idiom formation, such as idiom-breaking and contextual anchoring (Gustawsson 2006, Langlotz 2006, Minugh 2007).

1.3.4 Examining an idiom structure in corpora: *The Filling in the Sandwich*

*In this first chapter I will put my methodological cards on the table so that my game will be clear from the start.*

(Steen 2007: 4)

Study 3 focused on how a specific semantic and grammatical phenomenon was realized as part of idioms: the insertion of *context-oriented modifications* into above all the premodifier position of a salient noun in the idiom (as in Steen’s example, above, where the word *methodological* links back to his chapter on methodology, not to some odd form of Tarot deck). For this investigation, the primary design issues once again were:

1. Which corpus/corpora to draw on;
2. Which idioms to search for;
3. How to search for and analyze the material.

Preliminary background readings indicated that idioms could indeed be modified (Ernst 1980, Nicolas 1995), but that this was not particularly frequent. This in turn meant that a rather large corpus base would be needed to find

---

89 A copy of Jaki 2014 arrived just as this thesis was going to press; dealing specifically with newspaper headlines, it should provide valuable input for future comparisons.

90 The quote by Steen introduces his discussion of various types of research on metaphor, classified along three dimensions: in grammar/usage, as language/thought, as behavior/symbolic structure.
sufficient data. Even as late as 2007, when this study was published, the range of corpora for idiom work was still quite limited: the British National Corpus (released in 1994) was the only large-scale (100M) linguistically-informed corpus freely available. It was thus an obvious choice as a component of the study, but not sufficient by itself: although it contained a wide range of genres, it was geographically restricted; worse, an American equivalent did not exist at all. The other components therefore comprised annual newspaper CDs, all of which were in the 50M range.91

With the BNC and the 2002 Independent, 150 M non-overlapping words of relatively modern BrE were available for the search; the equivalent amount of AmE was obtained from two quality newspapers from opposite coasts, the New York Times (1996) and the Los Angeles Times (1990), together with the final component, the 1995 Broadcast News (BN) CD-ROM, which consists of transcripts of radio programs. In all, 300 M words from the late 1980s to 2002 were available for the search. In addition, since BN consisted of (transcribed) spoken AmE, this partially balanced the BNC’s 10 M word (transcribed) spoken component, although not its range of genres.92

The idioms selected were once again all found in the Collins COBUILD Dictionary of Idiom (2nd ed., 2002). For reasons of convenience, the majority of the 55 idioms ultimately chosen had a keyword from the beginning of the alphabet (e.g. babe in the wood(s) was listed under babe); nothing suggests that this (as opposed to a fully random selection, or, say, every 50th word) would skew the data. One criterion was that these items should be reasonably easy to search for and to disambiguate from literal uses (so that for e.g. fresh blood, sentences like Traces of fresh blood were found at the crime scene could easily be filtered out); another was that they not generate a large number of irrelevant hits (thus, headwords such as heart, hand or foot were not chosen).93 More importantly, even idioms that a priori appeared to be difficult to modify, such as full circle, gray/grey area, couch potato and a

91 Their inclusion is discussed in the Study itself, but it should be noted here that newspaper CDs are not linguistically designed corpora, so that relatively few studies have been made of their material (Nicolas 1995 is an interesting exception). This issue will resurface in Study 4.
92 The selection of the years for the CDs was fortuitous: those were among the years for which our Department had managed to obtain CDs. Newspaper CDs were a bit of a Wild West market: some, like the NYT, were marketed for years by a specific company, while others came and went, vanishing without warning. One CD, USA Today (1990–94), I found remaindered for $1 in a book outlet store. None of them appear to be commercially available any longer, as lawsuits about proprietary rights to articles crippled them and net-based search engines for the individual papers have since taken over their function of information retrieval, to the detriment of linguistic research.
93 The CCDI idioms exhibit a range of grammatical types, whereas (like the study by Nicolas) most recent large-scale studies concentrate on only verb+noun constructions (preferably in the BNC) like fly the coop, spill the beans (Wulff 2008, Gustawsson 2006; and for German, Fellerbaum 2007b, Storrer 2007—Finkbeiner’s 2008 much broader study of German idioms is an interesting exception). Cf. note 91, below.
dime a dozen were included, to ensure a broad range of constructions. The actual searches were via WordSmith when possible, otherwise via the CD’s proprietary software; in no case were there any real technical difficulties.

The analytic framework was created primarily in response to previous work by Ernst (1980) and Nicolas (1995), with further input from Moon (1998). What is at stake is idiom modification via the addition of more information (and thus not grammatical modification such as active > passive); the interest in this phenomenon can be traced back to the issue of compositionality in idioms. Contextually oriented modifications (Ernst’s “domain delimiters”, which we call textual “anchorings”) are usually semantically anomalous (He (really) put his diplomatic foot in it this time) if considered only within the idiom itself.

All 55 idioms were then searched for in the corpora, and the results examined. The analysis used four relatively clear categories of modification, plus the simplex, or unmodified, idiom: (i) [“regular”] modifiers (a small drop in the ocean); (ii) anchorings (His bankruptcy was just another drop in the financial ocean); (iii) names (The Full Monty), and (iv) exemplifications (the acid test of a gentleman: respect for those who can be of no possible service to him). This classification then formed the basis for further discussion; the focus being on the category of premodifying anchorings or textual "hooks", i.e. premodification of the key noun by some word or phrase that is not related to the idiom proper, but to the text within which it has been embedded.

A contemporary example of this phenomenon may be seen in the citation from Gerard Steen at the start of this section; another example, from Study 3, was found in a program about changes in TV, where media mogul Rupert Murdoch was cited as suggesting a multiplexing approach to improving quality, the result being that


---

94 We may disregard modifications such as We’ve finally come full circle, a thoroughly gray area. Couch potato seems difficult to modify, although a postmodification like ?a couch potato in the grand style might be possible, and it seems unlikely that we would encounter ?In that castle, ghosts are a dime an eerie dozen. Still, stranger utterances have been heard, which is one reason why empirical data is so important.

95 In a useful paper about adjectival modification in German idioms, published the same year as Study 3, Stathi 2007 points out that Burger 1998 fails to note the distinction between semantically normal adjective modification and those that we call textual “anchorings”. Normal adjective modification is straightforward, and refers to the noun in the idiom, not to the surrounding context, e.g. He (really) put his big/ugly foot in it this time.

96 Category (iii) is arguably part of the simplex group, but it is also a name; categories (i) and (ii) are formally identical, but semantically distinct; the collocational semantics of (ii) are anomalous, as the modifier collocates with the surrounding discourse, rather than the idiom.
Clearly, these “hooks” cannot be searched for in a principled lexical way, but only by inspection of all of the results of the search; the classification was made during this inspection.\(^{97}\)

In an attempt to assess the range of possible anchorings or textual “hooks”, further examples were obtained via the net-oriented software Web-Corp, developed at Birmingham by Antoinette Renouf and colleagues.\(^{98}\) Since WebCorp is able to trawl the entire web, it excels at finding further examples of linguistic phenomena, thus making it a useful tool for this particular investigation, even though it cannot provide frequency information.

1.3.5 A diachronic investigation of idioms in a corpus: Is Time A-changin’?

Study 4 is concerned with how idiom frequencies in English have changed in recent English: Which (sorts of) idioms have become more frequent, less so, or even dropped out altogether?

To place this final study about idioms and corpora in its historical perspective, one should note that the release of the Time Corpus in 2007 marked the advent of first freely accessible large diachronic corpus in modern English. Ranging from 1923 to 2006, it contains about 100 M words. Another unusual feature of this corpus is that it contains the complete articles for all of its texts.\(^{99}\) Its only drawback (and a major one) is that it represented a highly specific domain, i.e. modern American weekly journalism, so that all attempts to generalize to English in general must be done with extreme caution (Leech 2012, Minugh forthcoming).

Given that the research question was diachronic in nature (how idioms in English have changed in the 20\(^{th}\) century), and that idioms occur relatively

---

\(^{97}\) If the corpus is parsed, or your software will do it on the fly, searches can be made for e.g. the ADJ N (e.g. specify N as apple cart and see which adjectives turn up), and if collocation information is added, phrases like digital apple cart can be flagged as semantically anomalous, and thereby detected, but since there are so many different forms idioms can take, people prefer to look for only predictable constructions, as in Wulff 2008, which concentrates on verb+noun idioms.

\(^{98}\) Now called WebCorp Live, and with considerable improvements, it may be found at [http://www.webcorp.org.uk/live/%20](http://www.webcorp.org.uk/live/%20).

\(^{99}\) Most corpora do not contain full texts, either for copyright or sampling reasons: either there are legal restrictions on what can be made available, or the text mass would simply be so large that linguists prefer to sample (as in the BNC or ICE corpora). The Time texts are not immediately accessible through the database, but can now be retrieved a few lines at a time, for those who so wish, or through the Time Archive ([http://content.time.com/time/archive/](http://content.time.com/time/archive/)). A number of subsequent corpora, however, do contain entire newspaper and magazine articles—but not entire novels or the like. See, however, the brand-new Google-US and Google-UK corpora.

The list of Time articles, with metainformation, has been made available by the compiler, Mark Davies, of BYU: on the corpus homepage, under “Help”, go to “Corpus Composition” to download a zipped list.
rarely, the corpus was the only option available at the time. The smaller Brown/Frown and LOB/FLOB corpora could admittedly have provided glimpses of the language in 1961 and 1991, respectively, but at 1 M words each, they could not have been expected to contain many idioms (as discussed in this article).

The twin limitations of the Time Corpus are obvious: its extremely limited provenance (AmE journalism only, heavily slanted towards contemporary news) and its extremely limited authorship (Time journalists and editors). On a more positive note, the diachronic comparison is genuinely genre-specific, with the caveat that Time itself may have evolved over the decades.  

A corpus with 100 M words sounds like a large corpus, indeed, but if it is divided into ten-year units, each decade has only 10–15 M words. For this reason, some of the comparisons over time were made between the earlier decades and the later ones, thus a two-point comparison.

Given Time Magazine’s well-known penchant for playful language (explicitly commented on in e.g. Leech 2012), considerable efforts were made to include a broad range of searchable potential variants of the target expressions in this study. A detailed check was therefore made for potential variation of idioms, using a one-horse town as the example, and the much larger contemporary Oxford English Corpus (OEC) as the comparison database. The basic idea underlying this idiom is presumably that the location in question possesses only one X, where X is something that could normally be expected to be plural; in the canonical form (according to CCDI), X is a horse. This question is not quite as simple as it looks: there is a fuzzy border between variants of this idiom and similar structures that merely describe the situation in that location, such as a one-crop economy; cases like a one-paper town could be either literal or figurative, depending on the context (either that the town is dominated by one political party [a political context], or that it’s too small to support two papers [an economic context]). However, the context will usually be sufficient to determine which aspect is being evoked.

A distinction should additionally be made between variation using synonyms (or scenario-related equivalents that allow e.g. passive/active variation) and variation using hyponyms: a main idiom might be e.g. /shake,
quake, tremble, quiver} in one’s {boots, shoes}, while the hyponym variants would specify in this case the type of shoewear: {loafers, size nines, Manolo Blahniks, jackboots}. The interesting point is that the hyponyms are more likely to be nonce uses, and less well-known (i.e. no consensus exists, or in Langackerian terms, they are not entrenched); such nonce uses are, however, more striking (and thereby probably more noticeable). A similar pattern of variation involving verb choice is presented in Section 6 of Paper 1, where the Coll data for the way construction is examined.

The idioms I selected for examination in Paper 4 were, as in the previous article (and for similar reasons), all drawn from the Collins COBUILD Dictionary of Idioms, 2nd ed. However, this time several specific subgroups were created, as follows.

The first group consisted of 46 idioms that were labeled (fairly) old-fashioned in CCDI. Here, the question was simply whether these terms were more common earlier, and in recent decades had become as rare as suggested. The CCDI label seemed out of place to me for some of these items, as I know and use them, so that actual large-scale testing seemed appropriate.

To create a relatively robust basis for comparison (as noted above), the Time data, although available per decade, was divided up into only two time periods, the 1920s–1950s and the 1960s–2000s, the assumption being that the earlier period would have higher values (“old-fashioned” in 2006 would presumably not be so in 1936). It is also to be expected that a number of these idioms are so rare (or nonexistent) in the corpus that comparisons must be made on a more aggregate level than that of individual idioms, particularly since some of them are primarily BrE, such as the idiom not have a bean, which does not even occur in the Time corpus. With more data, the selection of the divisions can become statistically important (cf. the discussions in Hilpert & Gries 2009, Millar 2009), but the Time numbers are too low for multiple break points.

Another way of investigating variation was by using the construction as adj as X, which allows us to track the variation in the X part of the structure.

---

103 The variations listed here (from Moon 2005) were all attested in the Bank of English. The suggested analysis is by the present author.
104 In keeping with the COBUILD approach in general, their entries did not use the actual label old-fashioned, but baked the information into the discussion of the entry’s meaning. Thus, for not have a bean, they write “This is a fairly old-fashioned expression, which is used in British English” (2002: 23b).
105 We leave aside the classic question of apparent time, which will be considered in 1.5.2, below. There is after all no reason to believe that people are not aware of how common a given idiom is at a given time (entrenchment must be seen as updateable). It is possible that they are more current in AmE, and that this was not adequately reflected in the Bank of English.
106 E.g. be in apple-pie order, a little birdie told me, snug as a bug in a rug, catch 40 winks.
This latter part could be either the canonical form (as scarce as hen’s teeth), an uncomplicated variant (as slippery as noodles) or extended via (often comic) exaggeration ([mall] parking space as scarce as Cabbage Patch dolls). Three such patterns found in CCDI were selected: as scarce as X, as slippery as Y, as happy as Z.

The next group of idioms selected was chosen to investigate whether the Time corpus would reflect the putative secularization of America: it consisted of idioms that were Biblical in origin. As CCDI did not actually provide such labels, the entire list of CCDI idioms had to be examined, including checking that the origins actually were Biblical. Some were easily identified (beat swords into plowshares, from Isaiah 2:3–4, or fall from grace, which invokes Christian theology), while others were more dubious (the olive branch can be identified with Noah, but also with Greek and Roman symbolism, including Pax Romana). As the number of such idioms, as well as their frequencies, was expected to be quite small, the results were aggregated by decade.

The final group selected was based on the semantic groupings provided by CCDI2. The area chosen here, DECEPTION, contained 38 easily searchable idioms, such as a shell game, to sell someone down the river and tell a white lie.107 Again, the expected figures were so small for many of these items that the aggregate results were again used, to search for overall trends. There is a certain potential for geographical skewing, as a number of these idioms appear to be British (as throughout CCDI), but their absence merely lowers the scale, rather than distorting the distribution.

It is again important to remember that the results of these investigations can provide an adequate view of the usage patterns in Time itself, but do not necessarily extrapolate to American English, let alone English as a whole. At the same time, it should be conceded that Time occupies a prominent place in middle-class American reading habits of the 20th century (Leech 2012, 2007), so that at least some receptive awareness of these idioms should have been obtained by its many readers.

### 1.4 Summary of the main findings

#### 1.4.1 The Coll Corpus

The first paper in this study, “The Coll Corpus: Towards a corpus of web-based college student newspapers”, was first presented at the ICAME 2000 conference, and subsequently published in the edited volume by Peters, Collins & Smith 2002. Although investigated some 15 years ago, it addressed

---

107 Five further idioms were excluded, due to the poor signal-to-noise ratio in the searches.
several issues about corpus linguistics that are still relevant today. Of particular interest are two:

(1) The practicality and usefulness of creating numerous small corpora with a relatively narrow focus;
(2) The usefulness of examining the written language of the next generation.

The first issue, that of smaller corpora, may seem strange to mention, given that today’s English megacorpora are measured in billions of words. However, as discussed above (pp. 37–39), narrowly-focused corpora enable us to carry out e.g. domain-based studies in detail, in a way that large corpora rarely do. Retrieving entire articles (let alone entire issues of e.g. a magazine) from older corpora was normally not possible, not least because of copyright restrictions, but also for reasons of storage space; samples were the norm for the *Brown* series, the *ICE* series and the *BNC*. The *Coll Corpus*, by contrast, contains one complete issue of each newspaper sampled, and retains all names of authors, editors and so on.

By 1999, a new form of empowerment was taking place, thanks to the Internet: individual scholars and students could directly retrieve electronic texts from the Net, creating medium-sized corpora with far less effort than in the past (the standard procedure in the past, typing in and proofreading texts, had been mind-bogglingly slow). This point was specifically addressed in the article, as it was written at a time when the *BNC* bid fair to be the “only game in town” for linguists. At that point, there was no large corpus of e.g. American English, and the only other corpus of British English (the Collins COBUILD *Bank of English*) was, and still is, not publicly accessible. The *Coll Corpus* paper thus was intended to demonstrate that a moderate-sized corpus (3.7 M words) of reasonable coherence could be relatively easily assembled from the Net and could provide interesting, domain-specific comparisons to the material in the standard corpora.

The second issue focuses on the fact that the *Coll Corpus* article deals with the written language of young people. The writers of the *Coll Corpus* were nearly all articulate users of the Internet at that relatively early date (the data was largely collected during spring 1999). Today, they are 35 to 40, which places them solidly within the age brackets for mainstream text producers, but at the time, they were clearly much younger than most writers with access to the reading public. The texts they produced for their student newspapers were produced *for their own sake*, not for term papers and tests, and were both longer and more complex than most contemporary internet texts such as blogs or tweets.

Today’s typical corpora with younger contributors are international learner corpora, i.e. collections of short essays written by non-native speakers (usually with a BrE university student segment as a control group), usually
under test conditions and for an English department, whereas the Coll Corpus texts were nearly all produced by English-speaking young middle-class adults writing for their peers about a wide range of subjects. It is important to note that these texts were relatively free of “guidance” (i.e. censorship) from older editors, as regards both subject matter and language. This was thus a window into reasonably sophisticated text production during the authors’ formative years. In that respect, the Coll Corpus remains unique.

Ideally, the data in the Coll corpus should be weighted to match the demographic spread of the geographical areas contributing to the corpus (which in 1999 in practice were identical to the British Isles, North America and the erstwhile settler colonies, i.e. Australia, New Zealand and South Africa, online student newspapers not yet being available for other English-speaking areas). In practice, there was a clear skewing in favor of North America, but not so drastically as to invalidate the distribution altogether.\(^{108}\)

To gain a sense of how the Coll corpus compared to other corpora, it was primarily compared to the four members of the Brown family then available: Brown, Frown, LOB and FLOB. Modal verbs provided quite robust statistics, and largely confirmed the changes detected in e.g. Leech (cf. also the subsequent discussion in Millar 2009, Leech 2012), but also indicated that will was much more in use, while the more informal ‘ll was surprisingly little used.

A series of lexical items probed for various well-known differences, particularly British-American, most of which were as expected, although sometimes hampered by insufficient data. Among the surprises were that these younger AmE users were much more willing to use towards and amidst than expected (even if toward and amid still dominated). All Coll writers were more reluctant to use the mutated forms elder, eldest than those in the Brown series.

Moving on to more grammatical questions, the past forms of the verb sneak were not much used, but snuck outranked sneaked in all geographical areas, indicating that its spread continues. The UK use of prevent without from was also seen here, although not as strongly as expected—and not at all in the US data; the change does not seem to have been accepted in the US as of 1999. The AmE double-marked hypothetical if-clause (if we would have had preliminaries we would have done better) was clearly underrepresented, presumably indicating that it has not yet been accepted in more formal contexts, whereas the mandative subjunctive (cf. Hundt 1998: 159–75) is clearly on the rise throughout the Coll areas.

The way construction (Goldberg 1995) was also examined for frequency and variations; no less than 85 different verbs were found, with the general

\(^{108}\) Cf. the discussion in Section 1.4.2, above, on balancing corpora.
pattern of a few key items, and a large number of variations. A series of formal markers such as *lest* and *albeit* were also considered. Of the further areas examined, perhaps the most striking was the BrE error rate for *its/it’s*, which reached a surprisingly high 11% for incorrect use of *it’s* instead of *its*, as compared to the US 2%.

Overall, the *Coll* data is clearly useful for checking on patterns and serving as a potential contributor to discussions about ongoing language change, at least for relatively frequent items. For lexical items, particularly those outside the domains familiar to college students, *Coll* did not seem adequate.

### 1.4.2 The College Idiom

In *The College Idiom: Idioms in the Coll Corpus*, first presented at ICAME 2006 and subsequently published in *ICAME Journal 38* (2008), the material in the *Coll* corpus was systematically examined with a focus on how college-age student writers used idioms. There are two interesting aspects about these writers: they are nearly all middle-class speakers in the 18–25 age bracket (although geographically distributed throughout most of the English-speaking world), and their youth suggests that they may not have mastered a fuller range of idioms, so that the ones they do use should be quite frequent (or at the very least, be “up and coming”).

This paper is based upon *Coll*, the corpus described in Paper 1, with a total of 3.7 M words, dominated by the US and Canada (a reflection of the then state of internet access more than anything else). A significant difference, as compared to *ad hoc* corpora created by commercial newspapers, such as the *Independent* or *New York Times* annual CDs (let alone the singleton *Time* corpus), is that the *Coll* corpus is not dominated by individual regions, policies, writers or editors; the single largest newspaper contribution was in total only 1.4% of the corpus, and that paper in turn had numerous contributors, whereas major newspapers will represent a much smaller group of writers and editors, often primarily working in the same building and writing for the same audience.

When this study was published, there had not been much corpus work based on actual idiom occurrences (see section 1.1.1, above), not least because they are not all that easy to search for mechanically, and because of the need for fairly large corpora to find them in. If we follow Sinclair’s dictum that “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), even the *Coll* corpus is far

---

109 For a discussion of the frequent linguistic pattern with a very few core variants and a large number of infrequent alternatives (primarily worked out with phonetic data, but applicable not least to lexical variation), see e.g. Kretzschmar 2012.
too small for this, so that this article should be considered exploratory, rather than definitive.

A total of 5,439 idiom tokens were found (out of its 3.7M words), which corresponds to a probability of any individual idiom occurring at about 4.17 times per 10M words (4951 idiom tokens, or 3.80 per 10M words, if the non-

CCDI idioms are removed). These figures are in line with previous research, indicating that the Coll corpus is not unusually rich in idioms. However, the number of idioms per newspaper ranged from 56 to 3 per 10,000 words, which is an appreciable spread; furthermore, the text size of the newspaper appeared to be irrelevant in this context.

An interesting pattern emerged as regard geography: The Coll material indicates that North Americans actually used a number of the idioms labeled BrE in CCDI, whereas the idioms CCDI labeled NAm were indeed primarily found only in the NAm texts; this discrepancy is probably an artefact of the corpus used for creating CCDI (the Bank of English). Geographical gaps are, however, easily bridged if they come into the international limelight, as the history of the idiom It isn’t/ain’t over until the fat lady sings indicates: unknown to BrE in 1990, it became well-known to British political commentators through the 1992 Clinton-Bush election, and is now considered common goods (NODE 1998: 1029).

As expected, body-related headwords like eye, hand, face, finger, head were all prolific producers of idioms, but rather surprisingly, door was also highly productive, primarily through the idioms open/close doors, which were clearly related to the sense of a college education opening doors for future successes; another frequent headword, edge, produced the marketable idiom cutting-edge (+ noun).

The first 50 newspapers were examined for the placement of idioms, particularly whether they frequently occurred in the headline, or at the beginning or end of the text (cf. Moon 1995). Surprisingly, only 4% were in headlines, with another 2% each at the beginning or end: since most articles are quite short, this is a strikingly low figure, and suggests that their function is not the obvious one of being a rhetorical flourish in those positions, a point explored further in Section 1.6, below. If one instead considers the text type, editorials clearly dominate, with roughly twice the number of idioms, as compared to sports articles. Another idiom-rich area is that of quotes (particularly interesting, since editorials have practically no quotes at all).

---

110 The details: 5439 idioms in 3,738,946 words means that 5439*10M = 3,738,946*x, so that x = 14,547 idioms per 10 M words. Since there are 3485 different idioms in CCDI, the probability of any given idiom occurring is 14,547/3485 per 10M, or about 4.17 times. Naturally, the actual occurrences will vary, as idioms are not equally frequent.

111 Since the papers themselves vary enormously in the extent and way they designate their articles as belonging to one genre or another, there is a fairly large margin of error, although hardly enough to threaten the editorial-sports distinction. See Section 1.6, below.
One point that is difficult to resolve in idiom studies is the status of names that encapsulate idioms, such as the film *The Full Monty*, the album called *Salad Days*, the band called *Ten Foot Pole*, the song called *My Name is Mud*, the bar called *Mick E Fynn’s*, the coffee shop called *The Common Ground* and the like; together, they comprise 4.7% of the idioms found. The key point is that users have no choice about using the idiom, as it is the name of the object; in an extreme case, the user may not even be aware of the idiom underlying the name. Local uses, such as for the coffee shop mentioned earlier (*The Common Ground*), may be considered simply a local echo of the human fascination with language, but a successful film or book title like *The Full Monty* or *Catch-22* has the potential to introduce and keep an idiom alive for an entire generation or more.

Finally, we should mention that there are a considerable number of linguistically creative variants of idioms, including puns (*to Kerry the torch*), elaboration (*all dressed up and nowhere to sing*), contextual specification (*wake up and smell the bird seed*) and more.

### 1.4.3 The Filling in the Sandwich: Internal modification of idioms

First presented at ICAME 2004, this paper was subsequently published in Faccinetti 2007. It engages with the issue raised by Ernst 1980 and Nicolas 1995: how idioms can be overtly linked with the text they occur in. Consider the following pair of sentences:

[5] Mr. Major really let the cat out of the bag this time.
[5a] Mr. Major really let the cat out of the NATO bag this time.
[5b] This time, Mr. Major really let the cat out of the bag that NATO was holding.

Assuming that Mr. Major refers to the former head of the Tory party, we expect the word *NATO* to link back to a political discourse involving the PM, rather than a local cat and some bag with *NATO* stenciled on it. It is this sort of specification, or **anchoring**, in the larger discourse that is at issue here, more specifically in the form of premodification, as in [5a].

A primary characteristic of idioms is that they are self-contained and complete stretches of discourse, and can be inserted into most texts with almost no restrictions beyond those of style and relevance. They require no further explanations or linkings to function well: a sentence like [5] is fully adequate as it stands. The pattern in [5a] or [5b] is optional, not obligatory. However, previous scholars had claimed that anchoring modification as in [5a] was “fairly common in newspaper and magazine articles” (Ernst 1980; cf. also Nicolas 1995). To investigate this, corpus material from the *BNC* and
four different newspaper CDs (The Los Angeles Times [1990], Broadcast News radio transcripts [1995], The New York Times [1996] and The Independent [2002]) was examined in search of such anchorings or ‘hooks’ into the context (Ernst’s ‘domain delimiters’). A selection of 55 easily searchable idioms were examined in these five sources; variants were included, such as new/fresh blood ‘new participants, players, etc’.

For Ernst and Nicolas the theoretical question, ultimately derived from generative grammar, revolved around the grammar of idiom-breaking, the extent to which an idiom can be changed. This issue was previously considered controversial, but it is now generally agreed that most idioms can be more flexible than first thought (cf. Gustawsson 2006). The question this paper asks instead is how and to what extent idioms can be anchored in the discourse around them (cf. Stathi 2007).

In a typical modification, as in [6], the adjective collocates properly with its noun:

[6] “Yet we cannot get visas and we are such a small drop in the ocean compared with the businessman.” (BNC: HJG 1067)

In the variant we call anchorings, as in [5a], the adjective fails to collocate with the rest of the NP. Another, somewhat similar group of modifications, may be called “exemplifications,” as in [7], below. They too are linked into the context, not the idiom:

[7] Critics of Gorbachev ask why he does not bite the bullet and move directly to a market economy — as Poland has. (LAT 1990)

The use of a parallel clause referring to the market economy links the bite the bullet idiom into the economic text surrounding it. The major difference, then, between anchoring and exemplification is the formal difference between premodification and coordination; note also that the coordinated material can sprawl out at great length, while processing constraints keep the premodifiers short.

In sum, Study 3 distinguishes five patterns: simplex (as in [5]), idiom-related modification (as in [6]), anchorings (as in [5a]), exemplifications (as in [7]) and names (as in the film The Full Monty).

Although Ernst 1980 and Nicolas 1995 indicated that contextual anchorings are common and almost never impossible, the larger empirical evidence does not support such a standpoint: the overwhelmingly most common patterns in Study 3 are either the simplex, or canonical idiom, with no additional information (47.9% of the idioms investigated), or with modifiers that are

---

112 Altogether some 300M words, about half written, half spoken (the 10% spoken segment in the BNC and the entire BN CD are transcriptions of spoken English); geographically, they are half British, half American.
themselves located within the world of the idiom (43.7%), as in [6], where small is a normal modifier of drop. Anchorings turn out to be an infrequent variant, comprising only 2.7% of the instances found, a frequency almost identical to that of exemplifications, which comprise another 3.0% of the instances in these corpora. Names are also only 2.7%, so their dubious status will have no major effect on these findings.

Although generally unusual, anchorings do seem to occur with specific idioms: 3 of the 55 investigated were astonishingly high, in the 18–23% range, and a total of 7 had 10% or better. Two of the highest, (not) have a leg to stand on (23.1%) and babe in the woods (18.8%), are relatively old: 1594 and 1595, respectively, according to the OED. Rain on your parade, at 23.1%, is first recorded in 1964, but in our corpora is as common as (not) have a leg to stand on.

Other items practically never seem to accept anchoring, such as a dime a dozen (97.8% simplex) or bats in the belfry (88.9% simplex). The first may possibly be explained by the unusual syntax, making it difficult to alter it, but that can hardly apply to the second: in case she had some unresolved bats in the belfry (COHA fiction, 1971) does not appear to be any more deviant than many other acceptable modifications. Corpus frequencies may play a role here, as high-frequency idioms such as catch sb’s eye (899 instances) are immensely more likely to exhibit variation in a corpus than low-frequency idioms such as out with the ark or a blot on sb’s escutcheon (5 instances each). Just under half fall below Sinclair’s cutoff point of 50 instances, so that even larger corpora would be needed for the lower-frequency items.113

The WebCorp results from the Internet were intriguing in this respect: they produced an entire series of anchorings, e.g. a drop in the {budget, casino, celestial, conceptual, financial, fiscal, funding, income, linguistic, Microsoft, ratings, sexual, virtual} bucket, providing further confirmation that this form of premodification is actually possible and used, although not at all to the same extent as pre-corpus research predicted.114

The final section of the paper uses a chapter in Richard Brautigan’s Trout Fishing in America to provide a reading that illustrates how anchoring can be used in literature to unmask the use of hidden ideology (cf. e.g. Goatly 2007). In this case, the invented phrase Trout Fishing in America is inserted adjectivally, in defiance of normal grammar, to obfuscate the ideological content of the actual premodifiers. The anchorings discussed in the body of

---

113 This is the obvious difficulty with all corpus work: failure to find a phenomenon may mean that it does not exist, but is not proof; a larger corpus may well include such a sought-after item. Trivially, there may also be a problem with the search techniques used.

114 WebCorp does not provide statistics for its searches, merely examples that it has found. Our main analysis, with the corpora as its base, can provide such statistics.
the paper can thus be seen in a more problematic, less value-neutral light in certain types of text.

1.4.4  Is Time A-changin’?: A diachronic investigation of the idioms used in Time

First presented at the Stockholm 2007 Metaphor Festival, this study was subsequently published online in Johannesson & Minugh 2008a and in book form in Johannesson & Minugh 2008b. It bears the hallmarks of another exploratory situation, where the scholarly community for the first time gained access to a major corpus of American English, albeit one with specific restrictions: the texts stretch from 1923 to 2006, but are of course only those published in Time Magazine. The major drawback of this well-defined and narrow domain is that it reflects the writings of only a few journalists and editors, with a particular emphasis on breaking news; the major advantage (beyond its being the first easily and openly accessible diachronic source for AmE) is that Time has long enjoyed high status in middle-class America, and is thus widely read and influential. The Time corpus size (c. 105 M words) was deemed sufficient to allow exploratory investigations of even such low-frequency items as idioms, and it therefore formed the base for this investigation. Again, the idioms investigated were drawn from the CCDI, but this time not the entire range, but only idioms selected for specific properties.

It is difficult to ascertain the extent to which Time itself has changed (see the discussion in 1.5.2, below), but from the figures on words per decade, it seems likely that color printing lies behind the post-60s drop of almost 30% in number of words per issue, as it becomes more of a multimodal magazine.

Idiom variation was first investigated via the idiom (to be) a one-horse town. During this period, the development of the automobile industry changed what could potentially be a literal description of a small town into a purely metaphorical description; interestingly enough, the idiom does not occur in Time with reference to a human settlement until a 1968 example. Instead, the main, sometimes dismissive use from the 20s on is about activities that are surprisingly small, as in [8]:

For a more theoretical discussion of an item’s potential linguistic impact on the speech community, see Leech 2007: 144–45.

See the discussion in 1.3.5, above, for the OEC data and semantic complications with this idiom. It is furthermore clear that the CCDI definition is not quite correct, as the dominant use in both Time and OEC is of a human activity, not a town, a point not made clearly enough in Study 4.
[8] Then he put up $10,000 to finance a one-horse airline which operated one single-engine Stinson cabin plan from Oklahoma City.\textsuperscript{117}

The results confirm that Time indeed likes to give priority to word play and semantic extension, although perhaps not as drastically as the British tabloid press headlines (Jaki 2014). A further check appears later in the study, via the as X as Y idioms (see below).

For “old-fashioned” idioms (the CCDI label for them), a division was arbitrarily made into two roughly equal time periods:\textsuperscript{118} 1923 through the 1950s and from 1960 to 2006, which places the breaking point around 1960 (the end of the Eisenhower presidency and the election of John F. Kennedy), a point at which American society began to experience major cultural shifts such as the Civil Rights movement, the onset of the Vietnam conflict (the first seriously-questioned modern war) and the first stirrings of the modern feminist and gay movements.

For the ten most frequent of these idioms, the difference between the two periods was statistically significant, with \( p < .001 \), so that they have clearly dropped in frequency during the 20th century. The aggregate frequencies for all 46 items (Figure 2) show the same trend, with the breaking point in Time apparently occurring somewhere between the 1960s and 1970s.

To arrive at some sense of how willing Time was to use creative idiom-breaking, the as X as Y construction was examined, using three idioms: as scarce as hen’s teeth, as happy as {a clam, Larry, a pig},\textsuperscript{119} and as slippery as an eel. Here, the canonical form, although present, was actually less common than the shifted idioms. Thus, as scarce as hen’s teeth\textsuperscript{120} (6 occurrences) was paralleled by 22 changed instances, as in [9]:

[9] Hollywood is finding that adaptable novels are as scarce as cheap real estate

[10] In questions of air safety, definitive answers are as scarce as anti-gravity screens

where [9] has a straightforward comparison that is related to the context (the highly urbanized Hollywood area), whereas [10] permits itself a flight of fantasy. The same patterns can be see for as happy as X, which has both simple and complex variants:

\textsuperscript{117} All the examples in this section are drawn from the Time corpus.

\textsuperscript{118} The Time corpus provides data by periods, which Study 3 used, but by looking at individual examples, it is possible to ascertain the date of each example, and thus create different temporal classifications.

\textsuperscript{119} Type II errors (false positives) predominate for as happy as, no less than 49 of 84, or 58%, such as I am as happy as I can be, without you (Time 1942). They must be eliminated by inspection.

\textsuperscript{120} Curiously enough, the canonical form has hen’s, not *a hen’s or *hens’ (cf. OED Online, hen, n., CCDI hen).
[11] as happy as a grig ['a happy person']\textsuperscript{121}
[12] as happy as a Teletubby on tequila

In particular, it should be noted that the extension in [12] uses alliteration.
Some of these examples, however, are arguably not idiom-breaking, or idiomatic at all, but merely informative, as in [13]:

[13] good generals were as scarce as good shoes in the Continental Army,

the point being that good shoes were scarce in the American army of 1778, so that hen’s teeth is actually irrelevant here and may very well not be part of the meaning.\textsuperscript{122}

The next issue, that of secularization, was examined through the possible loss of Bible-based idioms. This is of course not a straightforward link, as the origins of an idiom are not automatically activated by the idiom; people can learn the idiom to fall from/out of grace without understanding the theological background, and the Biblical sense can even be entirely lost, as a comparison to the history of a sea change ‘a major change’ suggests: most instances of its use in modern texts carry no hint of its Shakespearean origins (cf. Minugh 2007b).

The actual data chronicle the occurrences of such idioms, and do not clearly suggest a major drop, particularly since the data from the 60s to the 00s exhibit no real change. Also, as pointed out in the article, Biblical phrases are always available for revival.

As regards Study 4’s final question, the semantic field of DECEPTION, it examined only 38 idioms in this area specified by CCDI\textsuperscript{2}; it may be assumed that other expressions can represent these same functions, so that a drop in frequency may not mean a change in the overall usage, only in the usage of those specific expressions. As the number of instances was not large (some failed to occur for decades at a time), the results are aggregated. Their peak uses are in the 50s and the 90s, but the most interesting decade is the 70s, which appear to be the lowest results since the Depression. The jury remains out on this issue.

Although the Time corpus no longer is the biggest or the best, when it comes to American corpora, it still can provide insight into how language interacts with society over nearly a century. The possibility of checking findings from it with those of other corpora merely adds to its interest as a corpus source.

\textsuperscript{121} The OED hedges its bets (grig \textsuperscript{n1}, sense 5, ‘merry as a grig’), the suggested origins being ‘grasshopper’, ‘(young) eel’ or (improbably) ‘a Greek’.
\textsuperscript{122} See also the similar discussion of one-horse town and one-crop economy, in Section 1.3.5, above.
1.5 Discussion of the Findings

In this Section, it is (perhaps somewhat illogically) assumed that the reader has in fact already read the papers in Section 2, and summarized in Section 1.4, so that we may now proceed to consider the overall aim and the research questions they dealt with, together with results from other recent research.

The first section (1.5.1) deals with the issues involved in the creation of a new corpus, such as selection, balance and representativeness. Section 1.5.2 considers the problems inherent in comparing corpora as such, and the specific problems involved in the comparison of the Coll corpus with the corpora available at the time of its creation (1999). Section 1.5.3 deals with the selection of material to be investigated: since our specific interest is idioms, and they cannot be exhaustively identified, how is the selection to be made, and what consequences stem from these choices? Section 1.5.4 reviews the findings in terms of frequency, distribution and functions of these various idioms, while the final section considers how they may vary, both in form (“idiom-breaking”) and diachronically.

1.5.1 Corpus creation

Research question (1) asked, “How does the data in a coherent, moderately-sized corpus of English written by young people, based on the English-language online college newspapers (then) in existence, compare to data from other then-existing corpora?” Since no such corpus (or subcorpus) existed, the author created one, taking the same advantage of the increasing use of the Internet that these newspapers themselves had begun to do—at that point in time, most universities and colleges had not yet produced electronic versions, but a sufficient number had taken the plunge to make such a corpus possible.

The question of corpus creation is thus central for Study 1, which details the creation and exploration of the Coll corpus, an early attempt at utilizing material on the Internet for linguistic purposes, by identifying a specific group of contributors whose writing we did not know all that much about: university students writing for their peers (as opposed to writing term and exam papers for their professors). In subsequent years the explosive growth in chat sites and blogs (at least the more serious ones) has opened up alternate channels, but on the whole, these latter sites are less productive of reporting and argumentative texts of moderate length, as well as being far more socially heterogeneous.\textsuperscript{123} We shall return to this point in Section 1.8, below.

\textsuperscript{123} The new social media such as Facebook, Twitter and Instagram are far shorter still, and therefore rather different again, as well as being a heterogeneous genre.
Comparing the internet situation of 1999 with that of the present, the basic social, demographic and geographical facts underlying the Coll corpus seem not to have changed significantly, with the North American hegemony unchallenged in terms of college and university online newspapers. Wikipedia currently (2014) lists 718 student online newspapers in the US and 89 in Canada, or roughly triple the figures of 2000: 215 and 35, respectively; Yahoo currently lists 784 for the US and 43 for Canada. These figures are almost certainly too low, given that the US alone has over 5,000 accredited university and college campuses.\textsuperscript{124} All of this suggests that online college newspapers continue to flourish and expand in numbers.

However, in spite of the ever-growing use of English in tertiary education in a global context, this is not yet reflected in internet-based university newspapers run by students. The textual dominance of the traditional English-speaking populations continues unabated. Although the first stirrings of a more global student online use of English are beginning to be visible (Wikipedia currently lists 17 student online English newspapers in India), their number is still tiny in proportion to the populations involved. That was not the case in 2000; they were simply nonexistent.\textsuperscript{125}

The technical side of corpus creation (its collection and structure) is, if anything, easier today, with far more rapid Internet access and much greater computer storage space and processing power, but the single aspect that the corpus would most have benefited from remains unchanged: the need to put it into a standard mark-up language (cf. e.g. Reppen 2010b, Nelson 2010 Koester 2010 and Clancy 2010). This is not particularly challenging for those with programming skills, and adequate semiautomatic solutions have already been developed. Nevertheless, it should be remembered that the first research question merely assumed the rapid compilation of a not particularly large corpus, and that accessibility via e.g. WordSmith was sufficient to allow a basic analysis. This proved to be feasible, even if more sophisticated methods of collection and archiving are to be preferred whenever possible. The choice not to attempt creating a better-structured version of the Coll corpus was primarily determined by the question of copyrights: in 2000, it did not seem likely that enough permissions could be collected to make the corpus freely available.

The question of corpus balance and representativeness (Leech 2007) has several dimensions: finding and retrieving as many different newspapers as

\footnotesize
\textsuperscript{124} See http://ope.ed.gov/accreditation/ for a list of postsecondary accreditations. The current list contained over 10,000 schools, many of which, however, were vocational, offering only a narrow range of short-term applied courses and programs.
\textsuperscript{125} See e.g. Graddol 2012 and references therein. There is clearly also a connection here to the expanding field of “English as a Lingua Franca” (ELF).
possible; how many issues per newspaper were to be included, and when to collect the data (different times of year yield different seasonal themes). The most important factor was the number of issues per newspaper, as that could vary widely. The optimal solution was sketched in the paper itself (and discussed in Section 1.3.2, above): create a smaller, more geographically balanced “monitor corpus” (one issue per newspaper), which thus became the Coll corpus, and create another version that includes much more (a project that was not realized due to lack of time): given the large number of newspapers involved, it would still be unlikely that any one newspaper would publish so much material that it could dominate analyses.

However, the choice to limit the Coll corpus to this balance implies that its limited size would not be adequate to investigate its idioms in great depth, so that our research questions 2 and 3 could only be given partial answers directly through Coll. In addition, there was the as-yet unknown relationship between professional journalistic writing at the regional or national level, and the cub reporters creating the online college newspapers in the Coll corpus; it was by no means certain that Coll would be a miniature copy of e.g. the New York Times. On the positive side, contributing to a college student newspaper is far more easily done than getting hired by a national newspaper, so that the number of writers per newspaper is likely to be larger, and the range of their backgrounds larger, than that for professional newspapers.

1.5.2 Comparing corpora

Comparing corpora is rarely easy and foolproof. The best matches are between corpora that are designed and executed according to the same model: the most obvious example is the Brown series: the original Brown corpus of American English triggered the creation of the LOB corpus of British English (and also with material from 1961), and the twin Freiburg Brown and FLOB corpora; this latter pair matches the original Brown design, but with input from 1991, 30 years later.\footnote{126} Even here, however, Leech 2007 notes that a series of potential objections can be made. To take just one obvious example, genres are not necessarily stable over time (Leech 2007 speaks of “genre evolution”), even if their labels may be, so that an important genre in 1991 may not have existed in 1961, not to mention the obvious fact that there are strikingly few blogs from 1961…

A series of papers about English modals considers this problem, beginning with Leech 2003 (about the Brown family of corpora), continuing with...
Millar 2009 (using a detailed statistical analysis of the *Time* data), and concluding with a detailed, convincing extension in Leech 2012. The crux of the matter is that *Time* represents *Time*, and cannot automatically be assumed to be representative of English in all respects (see also Section 3 in our Paper 4), a trap briefly discussed above.\(^{127}\)

Leech’s second important point is the question of whether a cultural artifact at time T\(_1\) is identical to the same artifact at time T\(_2\): Section 5 in Leech 2012 takes up this question in detail, using the frequency shift in *may* in *Time* towards its epistemic meaning as an example of how the 1920s *Time* may not be the same as the 1990s *Time*.\(^{128}\)

Returning to *Coll* and the *Brown* family of matched corpora, it is naturally possible, by normalizing to figures per 1 M words, to compare the *Coll* material with these four corpora, always keeping in mind that *Coll* is a decade later than *Frown* and *FLOB*, and four decades later than *Brown* and *LOB*. This was done in our Paper 1 for e.g. the modals. When doing so, however, one ignores the salient fact that these four contain a much wider sample of English than college newspapers do. A more fine-grained analysis than in Paper 1 would have been more revealing for the modals, in particular.

To test this, the *shall* and *will* modals have now been examined for *Brown* and *Frown* categories A–C (journalism samples) and compared to those in *Coll*. In *Brown* A–C, *shall* occurred 25 times (about 118 per 1 M), which is somewhat less than half the *Brown* average of 255 per million, but almost seven times that of the *Coll* US figure of 17 per million.\(^{129}\) More interestingly, in the *Frown* journalistic texts, it occurred only 8 times, or about 45 per million, which is less than a third of the *Frown* average, and much closer to the young Americans’ use of *shall*. The *COCA* newspaper results and those of *Coll* are almost identical for *shall*, which suggests that the gradual decline in *shall* continues, but is not accelerating among younger writers. The figures involved, however, are quite small in absolute numbers, and should therefore be treated with a certain caution. If we instead briefly look at its rival, *will*, we see that the *Coll* writers were willing to use it more often than in any of the other corpora (*‘ll* forms, not considered here, were surprisingly infrequent in *Coll*), as summarized in Table 1. This may suggest that *will* has taken over as the appropriate formal marker of the future in the *Coll* world.\(^{130}\)

---

\(^{127}\) For a more general review of methodological problems in corpus comparisons, see Leech 2007, and more recently, Davies 2012.

\(^{128}\) This has immediate bearing on our Paper 4, where *Time* is treated as a single artefact.

\(^{129}\) *Brown* and *Frown* had 88 journalistic texts (categories A-C); on the ICAME CD used here, there were a total of 212,090 words (the exact figures for *Brown* change, as adjustments are made over the years).

\(^{130}\) A more detailed investigation of course also needs to include *is/are going to* and other alternatives.
Table 1. A domain-sensitive comparison of shall and will.\textsuperscript{131}

<table>
<thead>
<tr>
<th>Corpus</th>
<th>shall / 1 M</th>
<th>Will / 1 M</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All (A–C)</td>
<td>All (A–C)</td>
</tr>
<tr>
<td>Brown</td>
<td>255</td>
<td>2145</td>
</tr>
<tr>
<td></td>
<td>118</td>
<td>3211</td>
</tr>
<tr>
<td>Frown</td>
<td>148</td>
<td>1996</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>3115</td>
</tr>
<tr>
<td>Coll (US)</td>
<td>17</td>
<td>4244</td>
</tr>
<tr>
<td>COCA</td>
<td>38</td>
<td>1855</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>2450</td>
</tr>
</tbody>
</table>

The distribution of shall, in particular, is highly sensitive to domains: an examination of its domain distribution in the COCA corpus indicates that newspapers have the lowest level (13.4/M), while e.g. fiction and academic texts have 68.3/M and 59.8/M, respectively. This suggests that a more detailed examination of the subgenres involved would indeed improve the value of the comparison. For the Coll corpus, this would require assigning a subdomain to each article, as the newspapers themselves do not classify them reliably (and often not at all).\textsuperscript{132}

Comparisons of individual lexical items are invariably problematic: how does one select the specific items examined? Is there a principled way to select areas of potential change? In Paper 1, there were several classes chosen, two with morphological markers (e.g. -st in amid(st), among(st), whil(e/st), and mutated vowels in older/elder, oldest/eldest), one with well-known geographical differences (e.g. windscreen/windshield), and one where the lexical item has hitherto been found only in US or UK-oriented areas (blindside/gazump). Here, the question of domain-specific use would appear to be somewhat less relevant than with e.g. shall. What instead is problematic, particularly for the lexical items, is that the numbers are very low, so low that an individual film name, for example, produced most of the instances of the verb shag. For the adjectives with mutated vowels, on the other hand, the information is sufficient to suggest that the younger generation, even in the UK, does not really use them.

The grammatical patterns do have enough information to suggest that the changes investigated have been picked up by the younger generation faster than the population at large. At the same time, there is always the possibility that as they grow older, these students will adopt the traditional variant—

\textsuperscript{131} Shalt and modal wilt are also included; shalt, in particular, is frequent enough to noticeably affect the overall frequencies of shall. The figures for will are for the verb only. COCA currently includes up to 2012, and is still growing (Davies 2010).

\textsuperscript{132} The problem of subdomains is equally applicable to the newspaper CDs used for comparison in Paper 1, as they too only sporadically label their subgenres; see, however, Section 1.6, below, where this approach has been applied for the editorials in Coll.
even if this seems much less likely than that the change is working its way through the generations.\textsuperscript{133}

The search for the \textit{way} construction (\textit{to X one's way + PP}_{loc}) in the \textit{Coll} corpus clearly produced enough material to gain a sense of how it can be used, which is the primary goal of the corpus. Interestingly, the data are of the same magnitude as for the \textit{BNC}, suggesting that this construction may be of general, rather than domain-restricted use. For the items chosen to investigate formal phrases, we again have the problem of how to select the most revealing items: the use of \textit{lest} in particular seems to signal that some such items may increasingly be being used only in a parodic sense.\textsuperscript{134} This again touches on the fact that as a genre, student newspapers are likely to have a higher proportion of parody than almost any other genre.\textsuperscript{135}

1.5.3 Choosing your idioms

The choice of corpus is not the only choice to be made: there is also the choice of the items to search for. Precisely because one cannot formally identify idioms (the way one can identify, say, the \textit{way} construction), there is no way to search for all the idioms in a given corpus. Instead, some form of selection criterion must be imposed upon the data, which in this case meant an idiom dictionary that was representative of modern English.

From the outset, one serious problem was how little various idiom dictionaries actually agree on what is worth listing. For English, the most recent idiom dictionaries at the time (with later updates) were:\textsuperscript{136} 

1. \textit{CCDI} \textsuperscript{137} \textit{The Collins COBUILD Dictionary of Idioms} (2002 [1995])
2. \textit{ODEI} \textsuperscript{138} \textit{The Oxford Dictionary of English Idioms} (1998 [1983])
3. \textit{LID} \textsuperscript{139} \textit{The Longman Idioms Dictionary} (1998)

\textsuperscript{133} This issue is related to the apparent-time hypothesis (e.g. Tagliamonte & D'Arcy 2007, Chambers 2002, Labov 1994).
\textsuperscript{134} The link of \textit{lest we forget} to World War I (specifically the Gallipoli Campaign and Anzac Day) makes this somewhat less likely in Australia and New Zealand.
\textsuperscript{135} Fortunately, the \textit{Harvard Lampoon} and other such publications were not included in the \textit{Coll} corpus. On the other hand, they too are language, and must at some point be reckoned with.
\textsuperscript{137} The 2011 edition is simply a reprint of the 2002 (2\textsuperscript{nd}) edition.
\textsuperscript{138} Now replaced by \textit{The Oxford Dictionary of English Idioms} (online, 3rd edition) (2009).
\textsuperscript{139} Replaces \textit{The Longman Dictionary of English Idioms} (1980). There is also an American spin-off: \textit{The Longman Dictionary of American English Idioms} (1999); as always, the American edition has fewer entries, which seems to be a marketing choice, as there is no reason to assume that over 300 million speakers would have fewer idioms available to them than the 65 million in Great Britain.
Since then, several additional idiom dictionaries have appeared:

(5) **DAI**  

(6) **CID**  

As with regular dictionaries, online equivalents are appearing—but it seems more likely that they will be merged into learner dictionaries online, which means that the frequency information available in e.g. *CCDI* will probably be lost.\(^\text{141}\)

*DOI*, by Linda and Roger Flavell, was by far the least comprehensive, with slightly less than 600 items, all with historical explanations. *CCDI* and *ODEI* provided a more traditional dictionary approach, with about 4400 and 6900 entries.\(^\text{142}\)

What is relatively startling, however, is how little these dictionaries overlap. To test this overlap, two comparisons were made. First, using the first item on every fifth page of *CCDI* (pages 5, 10 and so on), an 85-item list was generated and then matched against *DOI* and *ODEI*. Since *DOI* contains so few items, it is not surprising that only 10 *CCDI* items were also found therein (and only 5 in all three dictionaries),\(^\text{143}\) but the surprising aspect was that *ODEI*, which had some 58% more entries, contained only 37 of the 85 random items from *CCDI*, or 27.6% of the possible matches.

Next, the same procedure was applied to *ODEI*, generating a 121-item list for matching against *DOI* and *CCDI*. Here, too, the disparity was striking: only 9 and 28 items (7.4% and 23.1%), respectively, overlapped, although that might in part simply be a consequence of the larger size of *ODEI*. In total, only 6 of the 121 *ODEI* items were found in all three dictionaries.

These discrepancies cannot simply be explained by types of idiom, or stylistic level, or traditional vs. more modern, as quite traditional idioms in

---

\(^{140}\) Replaces *The Cambridge International Dictionary of Idioms* (1998). It has not been used in these investigations, primarily because *CCDI* worked so well, but Grant & Bauer 2004 and Grant 2005 chose to start their investigation from the *CIDI*, which may be seen as a strong vote of confidence in its scholarship.

\(^{141}\) Since learner dictionaries now give information about general usage frequencies, they are unlikely to additionally contain idiom frequencies.

\(^{142}\) *CCDI* (1995: iv, and 2002: back cover) specifies “approximately 4400 current British and American English idioms”, while *ODEI* (1998: back cover) lists “over 14 000 references recorded and described”. A check of the first 100 pages of *ODEI* produced the estimate of 6933 main entries, corresponding to the roughly 3485 in *CCDI*.

\(^{143}\) Mathematically, if the *DOI* size is extrapolated to match the 4400 in *CCDI*, there could be a putative match of as high as almost 75 items, or 87.9%, but this is quite misleading, as *DOI* primarily deals with idioms “with a story”, as opposed to relatively transparent idioms such as *bad blood, not the end of the world, pull strings, white as snow, don’t rub it in* or *play with fire*, all of which are found in *CCDI* and the first four of which are also found in *ODEI*.
CCDI, such as separate the wheat/grain from the chaff (Biblical in origin) and up with the lark (proverbial), are not in ODEI, as well as much more modern idioms, such as be in business ‘ready to go/start something’, go nuclear ‘go ballistic, get very angry’ and not rocket science ‘not demanding intelligence’, are all missing from ODEI.

Rather, at least part of the explanation lies in the much wider concept of idiom in ODEI, which includes what might more properly be seen as lexical items, such as cover address (‘an address, eg of an agency, one’s employers, etc. through which mail will reach the addressee’), a hit and run (driver), adverbials such as dead (on time), of necessity, on balance, or phrases such as clear one’s throat, do as one is told. All of these items clearly limit the space allocated to idioms proper; a detailed justification of their choices is spelled out in their introduction (1998: ix–xvi).

In all probability, even the number of “proper” idioms available for inclusion is considerably larger than any one book publisher chooses to print, and since idioms do not have a watertight formal definition, and are relatively infrequent, not even (partially) corpus-based dictionaries can be fully decisive on this point. In addition, it is obvious that regular 21st c. learner dictionaries now seem as a matter of editorial policy to include idioms under the relevant keywords: the Oxford Advanced Learner’s Dictionary (2005) contains 77 of the 85 selected CCDI items, or 90.6%, an inclusiveness which appears to be typical of modern learner dictionaries.

In the present work, for Papers 2 to 4, the choice fell upon the idioms listed in CCDI, the Collins COBUILD Dictionary of Idioms (cf. Section 1.1.1). Particularly for Study 4, however, the selection of idioms was further constrained, as when the investigation concerned only those with the label “old-fashioned”.

1.5.4 Idiom distribution in the corpora

The second research question, “How frequent are idioms as a whole in the selected corpora?”, has a direct bearing on Papers 2–4. The first step in each case was to search for these idioms.

Identifying the CCDI idioms in was in most cases straightforward. Paper 2 classified them into two groups: standard or variant idioms on the

---

144 See the objections to such a wide scope, as voiced in e.g. Gustawsson 2006: 36–38.
145 Many of these items are metaphorical, in the sense of the Pragglejaz group (2007), but e.g. hit and run is only used to describe a driving offense, and thus is not as idiomatic as, say, cut and run. Pragglejaz, however, would probably regard both hit (with a car) and run ‘(illegally) drive away’ as metaphorical, since using a car as an instrument is not a primary use of either hit or run.
146 A technical note: The search and retrieval processes were time-consuming, but not groundbreaking, and hence have been backgrounded in this thesis. See Grant 2005 for a description of a similar retrieval process.
basis of their mention in CCDI (e.g. to gather dust was standard, while to gather dust used a different, but clearly related lexical item, and hence is a variant idiom). As specified in Section 1.3.3, above, the difference between them was essentially whether the variants could be plausibly identified with the standard idiom, a process that admittedly calls for personal judgments, but which did not seem to seriously threaten the analysis. The third group also used in Paper 2, other idioms, was more problematic, as it threatened to be open-ended (but note the restrictions discussed in Section 1.1.1.1, above) and depended upon noticing them while weeding out uses that were non-idiomatic. Particularly for BrE idioms, there could have been further idioms that this winnowing process failed to catch. Identification of the CCDI idioms in Papers 3 and 4 was also relatively unproblematic, and there, no attempts were made to keep track of further idioms unrecorded in CCDI.

The distribution of idioms in the Coll corpus revealed an aggregate range of idiom density in its newspapers from 55.7 per 10,000 words to as low as 1.2 per 10,000 words, which is a large range, indeed. In order to examine this range in greater detail, an analysis of 50 newspapers (Table 10) looked at the various genres, to indicate whether this is a general phenomenon of newspapers (a result of staff preferences?), or one that is genre-determined. The results strongly suggest that there is a clear hierarchy by genre (editorials had the highest frequency, and news the lowest), and that would in turn affect the overall figures for various newspapers.

Paper 3, however, additionally used the Webcorp search engine (thus drawing on the total resources of the Net) to gain a sense of the potential range of the data beyond that of the actual corpora investigated (as discussed in Section 1.3.4, and commented on below, in Section 1.5.5). The results clearly indicate that there are a great many potential variants, although the main results also indicate that for most idioms, the canonical uses unquestionably dominate.

In Paper 3, however, where corpora were combined in order to obtain a sufficiently large number of tokens instancing the relatively low-frequency “anchorings” of idioms in their context, the BNC material was far more heterogeneous than the rest (all newspaper or radio program texts), and therefore should have been split up and its non-journalistic parts compared sepa-

---

147 Gustawsson’s examination of “idiom-breaking” (2006) focused on syntactic shifts, rather than lexical variation, which is the primary focus in the present studies. Both aspects are covered in Moon 1998. Langlotz 2006 provides a cognitive-linguistic model including (primarily grammatical) idiom variation, based on Langacker’s work, which appears to be an interesting development in conceptualizing what can occur with idioms.

148 A full classification by genre would require a classification of all of Coll. As many Coll papers only provide headlines, and not labels, this would be a major, and somewhat tricky undertaking. See the discussion on creating an editorials subcorpus in Coll (Section 1.6, below).
rately.\textsuperscript{149} It is not necessarily the case that the texts from other domains behave in the same way as newspaper texts, a point that should have been investigated, at least for the more common idioms.

There was, however, one question that appears in Papers 2 and 4, where frequency is a major issue, that still seems to be unresolved: should proper nouns (whether of TV shows, movies, books, stores, theaters, computer games, or even history) be counted as instances of idioms? Most tagging systems would treat the name of a shop, \textit{The Common Ground}, as a phrase that is a proper name, but if it at the same time is a pun on an idiom, should it be counted as an idiom each time that coffee shop is mentioned?\textsuperscript{150} Steen 2013 mentions this as regards metaphor, seeing the solution in whether the context actuates the conceptual basis of the metaphor or a literal reading, a process that also is seen in puns (cf. Alm-Arvius 2008). An utterance like

\begin{quote}
[14] I’ll go to Washington about this!\textsuperscript{151}
\end{quote}

can mean either [14a] or [14b]:

\begin{quote}
[14a] I’ll physically go to Washington (and talk to my Congressman)  
[14b] I’ll contact influential people in the government (without necessarily leaving my office)
\end{quote}

Similarly, the coffee-shop name \textit{The Common Ground} can be mentioned without activating the underlying pun, particularly if students have gone there for upwards of three years! The solution employed in these papers was to keep track of proper names, looking at the results with and without including them, a Solomonic solution that as yet lacks a formal basis. In general, people are surely not equally aware of etymological, social or historical clues as they use language, and this can obviously be applied here, as well.\textsuperscript{152}

This brings us to our next point: Distribution can be not only over geographical origin, but also over time, which is the major subject in Paper 4. The focus here is on usage, not on whether people understand the idiom in its original sense (which in any case may not be widely known, or even be completely obscure).\textsuperscript{153}

\textsuperscript{149} Just under 10.5 M words of the \textit{BNC} are classified as “newspaper text” in the \textit{BYU-BNC}.
\textsuperscript{150} Cf. the discussion on headings and headlines, which also touches on, but does not resolve this, in Moon 1998: 290–93.
\textsuperscript{151} An alternate formulation, “I’ll take this to Washington”, can normally be disambiguated, usually by whether the reference is to a physical object or not.
\textsuperscript{152} One obvious starting point is offered by Gibbs 1993, about “dead” metaphors, although not about names.
\textsuperscript{153} Two classic examples will suffice:
To establish which entries were old-fashioned (and thereby might be expected to be more common in the older issues of *Time*, the CCDI label “old-fashioned” was chosen as the criterion. This can be problematic for the individual researcher, particularly when an idiom that you consider perfectly normal is labeled “old-fashioned”. However, this is no different than many other lexical and grammatical items with variation—as an individual, you may consider the older variant the best or even the only acceptable variant. Here, the corpus provides the evidence, but unfortunately, the number of items found in this 100M word corpus was too small to test them individually. This precludes a more detailed analysis of individual items, although the numbers are high enough to justify the conclusion that many items CCDI labeled as (mainly) BrE probably had little or no currency in AmE even in the 1920s–1950s (further evidence from *COHA* in Minugh, submitted). The most intriguing result in this analysis is the aggregate change in idiom use in *Time* that is unquestionably visible in Figure 2, and which appears to have taken place in the (late?) 1960s. Clearly, at some point in that period, a change in idiom use in *Time* took place—but it is dangerous to draw a direct line to social change without a clear mechanism,—or directly from *Time* to all (American) English.

The question of secularization raised in Section 5.3 in Paper 4 relies to some extent upon the indirect chain of reasoning that people familiar with Biblical narratives will also know (and use) the idioms that derive from those narratives—but as noted above (and exemplified in Section 1.1.2), an idiom can break free from its origins. In the section on “deception”, examples can be found both of older idioms (*to sell someone down the river*, which of course had a literal basis in the New Orleans slave market and the Deep South cotton plantations before 1865) and of revived coinages (*to be economical with the truth*).

In terms of research question (2), how frequent idioms are as a whole in these different corpora, it is clear that these corpora can be used to find various idioms in use. The overall frequencies of occurrence seem to match those of previous findings (Moon 1998, Minugh 1999, 2001).

1. The proverb *For want of a halfpennyworth of tar, the ship was lost* (‘don’t be cheap, or you can lose something valuable as a result’), which apparently originally referred to tarring wounds on sheep (not ships) (cf. *OED Online*, *halfpennyworth*, n.)
2. The idiom *to kick the bucket* ‘to die’, which has numerous explanations, none definitively accepted (cf. *OED Online*, *bucket n.2*).

154 For a wonderful series of graphs and charts presenting ongoing change in pronunciation, and how people perceive this, see Wells 2008.
155 One example of such a mechanism may be discerned in the data and discussion presented in Baugh 1991, about the labels African-Americans apply to themselves: when the Rev. Jesse Jackson publicly endorsed the term African-American, he thereby legitimized it for America, and corpus results confirm that it had an immediate effect.
156 Paper 4 cites its use when it flared into prominence in the UK in the *Spycatcher* case, but the *OED Online* (economical, adj.) actually records this phrase as early as 1848.
Although the Coll results indicate that there are quite a few idioms that are frequently used (see Table 9 in Study 2), low frequencies for most idioms are again and again found in idiom studies. To take a concrete example, the 3.7 M word Coll corpus has no instances of to shoot the breeze, which seems common enough; data from the COCA corpus suggest that the phrase would be likely to occur only 0.78 times in a corpus of that size, or less than once (as is actually the case). This in turn suggests that one reason why people frequently do not recognize a given idiom is simply that they are not exposed to them sufficiently often (in Langackerian terms, entrenchment fails to occur throughout the speech community or at a young age). This is a reasonable explanation as to why it takes time to acquire many of these phrases, and strongly suggests that idiom acquisition should be a prioritized subject of future idiom research.

Research question (3) asked, “Are certain subsets of these idioms more common in different parts of the corpus? This comparison can be either geographical or diachronic.” For many idioms, the data is still uncomfortably sparse for the more detailed analysis that will tell us how they are distributed, particularly between different genres. The Coll corpus data does, however, clearly indicate that various sections of newspapers can have quite different overall idiom distributions. This will of course have repercussions for idioms studies that treat e.g. the BNC as a single source; it may in fact have multiple patterns in its subsections, patterns that cannot yet be detected, as there simply are too few idioms for the patterns to emerge statistically.

An examination of the rhetorical uses of idioms did not produce any major patterns (such as at the beginning or end of a text). However, text type clearly was relevant, with editorials and profiles (feature articles) being twice as likely to use idioms as most other types, and news scarcely using them at all. The quoted material in the Coll corpus, however, most of which is probably of spoken origin (interviews), showed a much higher frequency than normal (p < .001), indicating that future investigations should definitely attempt to include more spoken text and pay more attention to this dimension.

From the Coll corpus alone, it can be seen that the geographical labels in e.g. CCDI are in need of revision. This was in any case to be expected, in an Internet-aware ever-changing world, but it is clear that the roots of the problem go back to the earlier lack of large enough American corpus material.

---

157 Cf. the introductory section on “British and American idioms” in CCDI (1995: vii), which adumbrates the increasingly international nature of current-day written English.
1.5.5 Idiom variation in corpora

Research question (4), “How do these idioms vary in their realizations and textual uses?”, actually contains two separate questions. The first will be dealt with in detail here, while the second will be answered via two new sub-studies in Sections 1.6 and 1.7 which deal with further questions raised by the present Studies.

As noted in section 1.1.2, above, the question of idiom variation (“idiom-breaking”) was theoretically more worrisome to generative grammarians of the 70s than to today’s cognitive linguists. It has clearly been established that such variation does occur (cf. Langlotz 2006, Gustawsson 2006, Grant 2005, Moon 1999, 1998, as well as the results in the present Studies), but what remains to be worked out in detail is which types of idioms encourage this variation, and in which text genres.158

As its goal was to obtain a large number of examples to investigate a relatively uncommon, but recognized phenomenon, that of anchoring idioms in their surrounding text, Study 3 used four different corpora: four annual newspaper CDs were added to the BNC, to give access to a total of some 300 million words, about 50 of which were from spoken radio transcripts, with another 10 million words from the BNC’s spoken component. Newspapers dominated, but the breadth of the BNC gave some representation to other genres, hopefully allowing us to detect whether this potential distributional weakness was badly skewing the results.159

In the five-category classification employed (cf. p. 63 and Study 3), the emphasis was on premodifier anchorings within the idiom. Simplex forms and post-modification accounted for 90% of the examples, suggesting (i) that anchorings in the text are in no way required of idioms, (ii) extensive anchorings will occur as a post-modifier or via exemplification (to ease listener/reader processing). This fits in with what we know about idioms in general: as they are almost always orthogonal to the actual main discourse (introducing unexpected information, as in She stayed there until the cows came home—what cows? How did they get into the story?), they are processed on their own, and not integrated into the text in a literal way (cf. Steen 2013 for several extended analyses of this pattern with the use of metaphors).

158 An early signpost was work on speaker intuitions by Gibbs et al. 1989. A comment in Moon’s book indicates one genre to examine further: “In OHPC, the interruption of FEIs occurs most commonly in journalism” (1998: 177). In this context it should be remembered that the corpora used in Paper 3 were largely newspaper- and radio-derived.
159 In point of fact, this question of potential skewing (i.e. greater use in specific domains) was never thoroughly followed up, as the phenomenon of “anchoring” was so infrequent that BNC subsections were not large enough to do so: for e.g. rain on sb’s parade, discussed just below, only one single instance was found in the BNC (in fiction).
Interestingly, while the overall use of the anchoring option was only about 2.8% (2.7% if proper names are included), a few items were far more likely to attract this option, with *not have a leg to stand on* and *rain on your parade* both having anchorings for 23.1% of their occurrences, whereas others, such as *a dime a dozen* or *the whole ball of wax*, were 97.1% and 84.6% simplex, and yet others favoring post-modification or raising the question of which factors determine such different patternings.

It could be that this is merely random variation, particularly for those with relatively low numbers (*the whole ball of wax* was under 20 instances as a whole). However, phonetic, semantic and grammatical factors almost certainly play a role. For example, *a dime a dozen* (98% simplex) contains both alliteration and matching stress, both of which would be disturbed by the addition of an adjectival modifier: *a dime a political dozen* or *a political dime a dozen*. Its syntax is also unusual, something found in several of the highest-frequency simplex items (*pace Jackendoff*), while others seem quite normal (e.g. *to die with your boots on*).

For the top ten idioms with an exemplification (thus specifically tying the idiom into the discourse), 8 out of 10 have a noun or PP that can be specified as an example, e.g. *a bull in a china shop*, with 81% simplex, 8% anchorings. From a corpus-linguistic point of view, the crux here is that such questions need to be answered on an empirical, as well as a theoretical, basis. The data in Paper 3 is a first step in that direction, and again points to the need for detailed examination of idioms at an individual, rather than an aggregate level.

A related issue is seen in the variation of idioms in the Coll corpus (Paper 2): some 270 different tokens (5.0%) of the idioms found contained a shift not documented in *CCDI*. Potential motivations for this variation can often be found: some are clearly inventive substitution (*easy as torte*) or puns (*crash and learn, to Kerry the torch*), others reverse the idiom (*the buck starts here, to add injury to insult*) or produce twists from the text (in a text about pigeons: *wake up and smell the bird seed*). Again, 270 is too small a number for us to track the variation of individual idioms, particularly as there are not many instances of most of these idioms (*open/close doors* is the most obvious exception, as the corpus was collected around graduation time; it is perhaps the only idiom in the entire thesis where there is a clear and persistent link between an idiom and its textual context).

---

160 *Birds of a feather* (14% anchorings) has an optional alliterating and rhyming extension: *flock together; in this longer form, feather*…*together* can still work as a rhyming pair even if *feather* is premodified. It might also be argued that *to rain on X’s parade* is semantically deviant, as humans do not normally rain on anything (although humans can of course rain down blows on sb).

161 In fairness to *CCDI* and other dictionaries, no reference work can cover all of the potential variants for such expressions, and *CCDI* is careful to point to further variation for numerous entries. Cf. Kretzschmar 2012 for a theoretical discussion of this linguistic phenomenon.
To gain some sense of how much an idiomatic phrase can vary, Paper 4 investigated three constructions of the *as ADJ as X* type (all recorded in *CCDI*), as found in *Time Magazine*. There are 6 examples of *as scarce as hen’s teeth*, but 22 instances of *as scarce as* with other NPs, many of which are quite elaborate. Some are straightforward, as in *signs of a Sino-Soviet thaw are about as scarce as palm trees in Peking and Moscow*, or more unusually, *In questions of air safety, definitive answers are as scarce as anti-gravity screens*. As a similar pattern is found for the other two constructions, it appears clear that *Time* writers deliberately avoid using the standard idiom.\(^{162}\) Here, the major split seems to be between the simple comparison with an object known to have the characteristic in question and which thus can lexically vary the idiom, while retaining its semantic identity (*as slippery as noodles* rather than *as slippery as eels*) and the much more complicated and inventive use of scenarios intended to illustrate the emotion (*as happy as a five-year-old with his curls cut off*). Naturally, the more inventive NPs have no chance of becoming entrenched, but remain nonce uses. Lacking repeatability, they are quite unpredictable, not merely variants. For a variant such as *as easy as torte*, *torte* is in the same semantic field as ‘pie’, whereas for the nonce use *as happy as a Teletubby on tequila*—alliterative and striking as it is—, the scenario is hardly likely and there is no semantic support for the link. Since this particular construction can be extended through a locative (*as happy as a clam [at high tide]*), the extension into startling nonce uses is perhaps more tolerated: *as happy as a hayride down the middle aisle of Oklahoma!*

This situation can also be seen in terms of encoding and decoding: the nonce uses allow encoders considerable leeway in language fireworks, as they are usually easily decoded; at the same time, their probability of entrenchment in the speaker community as a whole drops sharply as they become longer or more esoteric. The construction *as ADJ as NP* also seems to be an important component here, as it is this frame that is predictable and thereby learnable.

We have previously noted researcher reservations about the extent to which *Time* is representative of anything but itself; however, judging by the wide range of nonce idioms that Paper 3 found via WebCorp searches, the *Time* data with nonce idioms does not seem quite as extreme as at first sight. Perhaps the most obvious reservation should be about its longer flights of fancy: since the NP branches off to the right, it is possible to have phrases

\(^{162}\) Recall the distinction by Katz on p. 34, above: the lexical form (as in *as scarce as hen’s teeth*) has an underlying conceptual scenario that can be invoked via a sufficiently similar lexical expression, presumably also invoking the canonical idiom, thus helping the interpretation; the more far-fetched or tenuous the link, the greater the danger that listeners or readers will not make the connection. Hence the value of retaining the grammatical construction as one way to support making the connection.
such as a hayride down the middle aisle of Oklahoma! However, it should also be noted that a clam at high tide is both longish and well-known.\textsuperscript{163}

The final aspect of our research question about variation and textual use can also be clearly seen by regarding these idioms as constructions. The three picked for examination are different, in that the first is canonically a comparison to a non-existent entity (hen’s teeth), the second to a physical property (a slippery feel) and the third to an emotion (happy). All three, however, largely avoid the physical reality described in the text (a brick path worn slippery as slate is one of the few exceptions), or invent other imaginary “happy” etc. scenarios. The invented variations, the majority of which are context-triggered, function within an extremely limited textual space; also, as most of them are noticeably longer than the original, they are not suitable for headlines. They are in aggregate more common in the earlier period (47 v. 28, which corresponds to 35.2 and 22.6 per 10 M words), but the numbers are too low to be trustworthy.

The Time idioms of Biblical origin break up suggestively into decades of “normalcy” (the 1920s and 1950s), with a high use of idioms, a decade of troubles (the Great Depression of the 1930s), and the other decades, all of which are at about the same level of 9 to 10 instances per million words. For such an analysis to be convincing, we will need detailed analyses of how these idioms work individually, not merely as an aggregate of idioms.

The great danger with such an analysis is that it facilely identifies each decade with (specific) social changes. For such a link to be convincing, one must postulate that the Time writers are – consciously or unconsciously – aware of the religious background for these idioms, and a certain degree of deliberateness must be assumed in their choices.\textsuperscript{164} This is easier for political use of idioms: there can be clear microclusters around a given idiom that assumes political significance in a given election year. During the presidential campaign of 1928, for example, a circular published by the Republican Party claimed that if Herbert Hoover won there would be a chicken in every pot and a car in every garage,\textsuperscript{165} and in 1984 it was Walter Mondale’s turn to ask Where’s the beef? (cited in Paper 3, below); these uses can easily produce microclusters in the data. Another external (i.e. non-linguistic) factor that may enter into this is the historical dimension: it is possible that to sell

\textsuperscript{163} This idiom is not in CCDI, but a COCA quote from NPR (National Public Radio) picks up a 1992 interview with Christine Ammer, author of Have a Nice Day--no Problem!: A dictionary of clichés, where she explains it.

\textsuperscript{164} Of course, once even well-educated writers and reporters are no longer aware of the roots of an idiom, this link can be considered broken. As an easily accessible written source, the Bible has hitherto resisted this. It is noticeable, however, that overtly marked Biblical quotes in the Coll corpus turn up in overtly Christian colleges, such as Walla Walla College in Washington State, USA.

someone down the river could become taboo, because of its increasingly disconcerting roots in the American slave trade, whereas to sell the pass, with its non-American origin (the OED suggests Ireland, with pass meaning ‘a place of defense’, first cited in 1850), is not as potentially controversial in American English.\footnote{This has happened with speak with forked tongue, which is impossible to use in North America today without “scare quotes”.
}

Moreover, in the interpretation of the data, there are considerations of the sampling process itself: the choice of intervals affects the amount of mathematical smoothing in the data, and sharper shifts may be seen if the appropriate intervals are selected, and there is as yet no agreement on what intervals function best (cf. Hilpert & Gries 2009).

As a final specific comment, it may be noted that for the semantic field ‘fraud’ (investigated in Paper 4), the occurrences of any given idiom per 10M are roughly equivalent to the actual numbers per decade, which means that they are too few in number to bear investigating at the level of the individual idiom. For the aggregate data from the 40s to the 00s, the dip in the 70s is merely part of normal variation (p = .80).

In the next two subsections, we will consider idiom use in detail, rather than in aggregate form, via two different approaches. First, in a new short investigation, we will examine all the \textit{CCDI} idioms found in a specific sub-genre of the \textit{Coll} corpus, namely all 132 instances of idioms found in \textit{Coll} editorials, to see what can be discovered about patterns of idiom usage in a given genre. In the following section, we will instead select a single idiom, the glass ceiling, and examine all 36 instances of its occurrence in a single spoken corpus, the \textit{Broadcast News} 1995 CD (used in Study 3), both to see its features in detail and to indicate how such an investigation can be integrated into teaching about idioms, using a corpus-data approach.

\subsection*{1.6 A Qualitative Substudy: Idiom use in Coll editorials}

For this investigation, the entire \textit{Coll} corpus was re-examined, in order to more accurately pinpoint and extract its editorials. Although editorials are a traditional component of newspapers, this does not seem to be the case for the \textit{Coll} corpus as a whole: of its 288 online newspapers, only 78, or 27\%, actually have identifiable editorials. On the other hand, those papers that do contain editorials often have several in the same issue, resulting in a total of 112 editorials available for the present substudy.

Identifying them turned out to be somewhat tricky, as the definition of what an editorial is turns out to be rather fuzzy: a number of different names are employed, ranging from obvious cases such as (Staff) Editorial, Web
*Editorial, Editorial Page, Letter from the Editor(s)* and authorial imprimatur such as *Editor-in-Chief*, *(Staff) Editor(s), Staff, Editorial Board*, through *Opinion(s), Viewpoint(s), Our View, Editor’s Comment, Commentary Editor* and less certain cases such as *Op/Ed, Ed/Op Editor, The Last Word*. Interestingly enough, the BrE term *Leader* did not occur at all. An additional dozen editorials with no designation, from some six newspapers, were also included among these 112 editorials, on the basis of (i) their being signed by *Staff or Editor(ial) Staff* and (ii) their paper having no other overtly labeled editorials.

In terms of subject matters, these editorials primarily revolve around campus life, either conflicts with or support of the administration and/or student governing bodies, or directed towards student attitudes to their studies, their college or student life. However, there are also a number of articles about national and international politics (51 and 3, respectively), notably including the Balkans conflicts, the Columbine shootings and ongoing political or personal musings or fulminations about e.g. Monica Lewinsky’s role in the White House at the turn of the century.

Since editorials normally weigh and judge, and often argue for specific standpoints, we might expect these texts to exhibit a heightened use of the two modals *must* and *should*, compared to their overall distribution in *Coll*, and that indeed is the case. Of course, modals have numerous meanings, but the deontic meaning (which is relevant here) clearly dominates for both (Millar 2009: 202–04). Comparing modal frequencies in the *Coll* editorials with those cited in Study 1 for *Coll* as a whole, we find that *must* and *should* occur at about 167% and 250% of their respective overall frequencies in the *Coll* corpus. The higher rate of these modals does indeed suggest that this

---

167 This lack, coupled with a noticeably lower frequency of editorials in the British part of *Coll*, may in part be a consequence of a general trend for these student newspapers to be far more tongue-in-cheek and self-mocking, with the editors’ ingestion of dubious and/or extralegal substances and lack of sexual prowess given far more space than in the American editorials, where articles written out of a sense of communal loyalty to their school and its mission were much more dominant.

168 Letters to the Editor, independent opinion pieces and articles signed by *(Staff/Contributing/Sports) Reporter* and the like were excluded, as were *Opinion* articles in papers that had articles directly labeled as editorials. From their content (particularly the presentation of arguments why the college or student body should act in a certain way) and form (particularly the use of modals like *must*), it was nevertheless clear that many of these articles were not straightforward reporting, but editorial-like in nature. This appears to be more the case than for newspapers in general, and is probably worth an investigation of its own.

169 Since the majority of the *Coll* material was gathered in May, i.e. shortly before the end of the academic year and graduation, another dominant theme is reflection on the paper’s performance during the year, and on the overall meaning of a nearly-concluded college education.

170 Direct comparison with the *Brown* family is not possible, as *Brown* category B, Editorials, also includes e.g. Letters to the Editor.
subcorpus corresponds to an editorial stance, so that we shall consider the label “editorial” appropriate. But which idioms, then, may one expect to find when it comes to the use of idioms, particularly given that the range of subjects is potentially very large, indeed? Can the range of idioms be in any way predicted, or their uses?

Idioms are to a large extent self-contained, carrying their own reference domain with them, as it were: when we say Now you’ve let the cat out of the bag, or Uh-oh, you’ve spilled the beans, we immediately know that neither cats nor bags nor beans need to have been part of the discourse. In consequence, like attitudinal adverbials, they can be “slotted into” almost any discourse. Nor does the discourse have to adjust to this addition of cats, bags or beans: we process the underlying message (the “meaning” of the idiom) and slot that into the discourse, instead. The writer can of course choose to overtly work the idiom into the discourse (see Study 3, below), but need not do so. As a result, we should not be too optimistic about predicting which particular idioms turn up in a given genre, in our case, editorials.

Given that editorials express an opinion, i.e. are normally judgmental, it might be assumed that they are somewhat richer in idioms’ “encapsulated wisdom” than other text types. This does appear to be the case, but not strikingly so, as Study 2 had already suggested. Considering the editorial material and Coll as a whole, we obtain the following:

Table 2. Percentage of idioms found in Coll editorials

<table>
<thead>
<tr>
<th></th>
<th>Coll</th>
<th>Editorials</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>3,738,946</td>
<td>59,632</td>
<td>1.59%</td>
</tr>
<tr>
<td>Idioms (Study 2)</td>
<td>5439</td>
<td>98</td>
<td>1.80%</td>
</tr>
<tr>
<td>Idioms (found)</td>
<td>(5572)</td>
<td>133</td>
<td>2.39%</td>
</tr>
<tr>
<td>Idioms (+ new)</td>
<td>(5719)</td>
<td>148</td>
<td>2.59%</td>
</tr>
</tbody>
</table>

The first row provides the baseline for comparison: it lists the total number of words in the Coll corpus and the “Editorials” subcorpus, which is 1.59% of the total corpus. The second row shows the number of idioms (tokens) in the “Editorials” subcorpus identified during the original search in Study 2 and their percentage of the idioms found in Coll. The third row also includes additional tokens of the CCDI idioms found during the present re-examination. The final row also includes an additional 15 instances of

The Coll editorials comprise 1.59% of the entire corpus (based on Study 2), but contain 48 and 143 instances, or 2.67% and 3.98%, respectively, of Coll’s total instances of must and should (in all their various forms).

The searches for Study 2 were primarily focused on the keywords given in CDI. In the present instance, the Coll editorials were converted into a single Word file, and examined with the help of the Word 2010 search engine. The latter proved to be a rapid and accurate
idioms not in CCDI. Mathematically, this is of course a case of ‘moving the goalposts’, since it can be assumed that further searches of the entire corpus in this manner would proportionately increase the original Coll idiom figure of 5439. However, here we simply add in the extra idioms actually found and accept that these pseudo-figures are only part of the story.

Unsurprisingly, this re-examination also suggests that searches for idioms are trickier than first assumed (cf. especially Grant 2005). For comparisons with the original study, the lowest figure should be used, but for the present substudy, we shall deal with all 148 idiom tokens actually found this time.

The first point to be made is that no idiom dominates: only 9 occur more than once (the most frequent, point the finger at ‘blame’, occurs 6 times, 3 of which were in the same editorial), with 4 occurring in two articles, and 3 in three separate articles. None occurred in four or more articles. One major implication of these figures is that idioms are in a very real sense independent of the texts they are inserted in, at least as far as the editorial genre is concerned: they do not seem to be predictably used for specific editorial topics. Nor, given the warnings about cliché writing found in writing guides, does it seem likely that other genres will exhibit a more predictable pattern. They are always available, but rarely to be expected.

If we consider an approach such as Conceptual Metaphor Theory (e.g. Lakoff & Johnson 2003, 1999, Kövecses 2010, 2007), it is not hard to demonstrate that many of the idioms found in these editorials incorporate standard conceptual metaphors: for example, the PATH metaphor may be seen in the home stretch (The Aumnibus, Auburn University), while Congress is dragging its feet when it comes to federal legislation (The Creightonian Online, Creighton University) is clearly connected to both PATH and ACTION IS MOTION; even more obviously, Sonnenschein and Dean of the College John Boyer thought they had won the war. Since then, however, they have lost every battle (Chicago Weekly News, University of Chicago) connects to ARGUMENT IS WAR. However, since there is no one-to-one connection between conceptual metaphors, basic or otherwise, and idioms, this approach does not provide any further guidance about what may be used in

way of finding incomplete idioms, such as gray/grey for gray/grey area or ticket for meal ticket, both of which would have been missed in the earlier search, which was confined to the keywords (e.g. area and meal, respectively); variants were not as thoroughly searched for in the earlier studies. In addition, the editorial texts were this time read in their entirety, with the specific purpose of detecting idioms, not least such fragmentary repetitions.

Some were clearly American (and thus less likely to be included), such as [American as] apple pie, hit a grand slam (a baseball metaphor), while others simply seem to be not in CCDI, such as stand up for sth, urban legend, since day one. Of these 15, only 4 were found in the ODEI: do your homework, go West [young man], the plank [=beam] in your eye, and broaden your horizons.

A personal favorite is We’re not in Kansas [any more] (‘We’re [back] in reality’, ultimately derived from The Wizard of Oz), but here used in a British editorial.
editorials (cf. also Katz 2006), and will not be pursued further in the present work, except as an aid in interpreting individual idiom uses.

Considering the idioms in a given genre as a whole, one possibility would be to categorize the context within which the idiom is used as positive (‘laudatory’), neutral or negative (‘critical’), where ‘positive’ means that the editor is in favor of the action or judgment, and ‘negative’ the reverse. Typical straightforward positive, neutral and negative examples, respectively, might be:

[15] Danny [Wuerffel, an American football player] was a class act [‘a really good role model or admirable person’] and a gentleman while being bombarded [with questions] by 20 or so college students. (The Aumnibus, Auburn University)

[16] It is much more important that this person we have in office [Bill Clinton] is checking out FBI files (never mind the fact that no one in this administration has security clearance to be able to look at these files) on the past administrations and selling our souls to the Communists for illegal campaign money… (The Aumnibus, Auburn University)

[17] The 1998–99 school year at Auburn will forever be remembered as the year of the catch phrase. Terms such as betterment of Auburn, personal vendetta, micromanagement, the Auburn family, cowards, integrity and closure have dominated headlines and conversations since almost day one. Is anybody else tired of this? (The Auburn Plainsman Online, Auburn University)

It is considerably more difficult to decide on the prosody of an idiom like in the works, which suggests action (as opposed to stalling), but ultimately relies on the context to determine whether the action is good or bad.174 The following example was therefore regarded as negative, not least due to the negative force of backhand and hide:

[18] The Colorado Supreme Court ruled Nov. 23 any documents that are part of a government decision-making process need not be released to the public. Hypothetically, [our college] could use this ruling to backhand students with another increase in fees or to hide any other decision that could be in the works. (The Metropolitan, Metropolitan State College of Denver)

and this one as positive, due to the positive force of promising, doesn’t take it lying down (another idiom), count on [an action] and care and show it:

173 It might be argued that the overall content is negative, but since the idiom appears to work perfectly well with positive sentences such as He’s been a star since day one, this can be regarded as neutral. For a way to examine this more closely, see section 1.7, below.

174 The key introduction to semantic prosody remains Hoey 2005.
But one promising trend we noted about OSU is that when the bad things happen, our community doesn't take it lying down. Count on a protest or a community forum, a candlelight vigil or a march, a letter writing campaign or a petition. The bottom line: OSU students care, and show it. (The Daily Barometer Online, Oregon State University)

A more intricate example is one use of the idiom to follow suit:

Yes, both of these new appointees probably have Auburn's best interest in mind, but it is unfortunate for this University that their opinions may follow suit with [Head Trustee] Lowder, rather than be reflective of their own thoughts. (The Auburn Plainsman Online, Auburn University)

The original sense, deriving from card games, is that one plays the same suit as the first player (spade follows spade, etc.), which is a form of cooperation. However, here the cooperation means following Lowder’s lead, rather than being an independent thinking Trustee, i.e. an abrogation of their position of trust, and hence ultimately negative.

This judgment introduces a subjective element, of course, and in addition involves dealing with multiple layers of discourse (a “positive” idiom can be negated in the surrounding text, as in the complexly layered example [30], below), so that individual instances may be somewhat contentious, but in aggregate should give us an indication of how idioms may be used.

Taken all together, the distribution of idioms in positive, neutral and negative contexts is positive 30%, neutral 5% and negative 65%, with a fairly large degree of uncertainty, probably at least on the order of 5% (such as Example [22], where the context is playful). Nonetheless, the results clearly indicate a preponderance of negative contexts. This disparity may, however, not reflect the idioms themselves (as noted above, they can to a considerable extent be inserted into very different contexts), and merely reflect the nature of editorials, which seek out polarized issues, and—particularly in a student-administration context—tend to be suspicious of the “other side”, i.e. administrators who restrict their freedom, increase tuition fees and all too often remind them of the adult world they have not yet fully accepted.

Another issue of relevance here is the extent to which these idioms form an integral part of the text. Earlier in this section, it was noted that idioms are

---

175 An interesting example of how a linguistic form can be in contrast to the standard conceptual metaphor DOWN IS BAD. There are of course others, such as to have deep roots, where DEEP IS GOOD, presumably (following Langacker’s layerings approach) relying on the more abstract MUCH IS GOOD.

176 More generally, Nunberg, Sag & Wasow 1994 include affect as one of their six dimensions of what a prototypical idiom consists of (cf. p. 23, above), which tallies well with the results here.
self-contained, but can be incorporated into the core of the image structure in the article (cf. the theory of Conceptual Blending, as espoused in e.g. Faucconnier & Turner 2008). But does this happen often?

On the whole, the answer appears to be a qualified “yes”. There appears to be a cline from deliberate echoes of the idiom further on in the text (quite rare) through strong to faint echoes of the conceptual metaphor underlying the idiom (e.g. ARGUMENT IS WAR) to idioms that merely function as temporal adverbials or the like. If we consider examples [15] to [20], above, we will find that in [15], a class act is not in any way picked up or elaborated further in the article, which is about athletes that have found support in being active Christians. In [16], venal crimes are at the center of the article’s accusations, which certainly fits the idea of buying and selling, while the potential religious dimension introduced by souls is not followed up. Thus class act is simply an alternative to gentleman (general praise), and not integrated into the text, while selling our souls resonates with charges of illegal money, and thus is partly integrated. Example [17] functions as a time adverbial (‘since the beginning’) and is not relevant to the article, while [18] in the works is about the decision-making process, and thus partially relevant. Example [19], on the other hand, functions as a metacomment clearly focusing our attention on the main message of the entire editorial: “OSU students care, and show it”; by virtue of its prominent position, the idiom is considered as thoroughly integrated into the editorial. In example [20], the article focuses on whether the new trustees will be independent or simply agree with the Head Trustee, so that follow suit is clearly linked into the discussion.

Using this approach, we can establish a categorization along a cline from unintegrated, through partial integration, to core integration, in other words a roughly tripartite division. Again, this is a partially subjective categorization, but it can serve as a rough estimate of how idioms are used in these editorials. The results are approximately 45% core, 40% partial and 15% unintegrated (see Examples [20], [16], and [17], respectively); alternatively, one can say that about 80–90% are more than casually related to the text they are integrated into, not primarily through the domain they derive from, but rather by fitting into the theme of the editorial.¹⁷⁸

What is particularly striking is that one claim of Study 2 is confirmed, in that not very many of these idioms are found in the key positions of the headline or the final sentence, less than 5% and 6%, respectively. If we relax

¹⁷⁷ Given the contextual tired of, however, it can be argued that since day one is being used to mean ‘all the time, forever’ and thus is negative.

¹⁷⁸ Several caveats apply: some editorials are not unified, but take up several themes; a few take e.g. the academic year’s events as a structuring device for fragmented editorials that are a series of otherwise unconnected brief comments.

Also, it is clear that not all idioms have an original physical domain accessible to most speakers, the classic kick the bucket being only one such case in point.
the restriction to the first and last paragraph, we still do not reach 10% for either slot. A number of these roughly two dozen headlines and concluding lines involve the use of humor (which is otherwise surprisingly rare):

[21][Headline:] Greener Pastures: Overnight parking in lots should be allowed. (The Michigan Daily, University of Michigan)

[22]The licensing commission should reconsider its decision and give a legitimate business the right to expand. It's a move that may be uncomfortable for some city officials, but they'll just have to grin and bear it. (Daily Free Press, Boston University)

[23][Header:] “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.” - Edith Hubbard (The Gauntlet, University of Calgary)

The headline in example [21] refers to a potentially better use of university parking lots named by color (the Gold Lot, the Blue Lot, the Green Lot); pastures are not involved. In [22], the editorial’s final sentence, the business referred to is a strip club, hence the deliberate misspelling of bear; this joke could be seen as undercutting the seriousness of the editorial, which is in support of the business’s right to expand, but it does not seem to have worried the editors. The Devil’s Dictionary-like definition in [23] introduces a serious editorial about what should be considered newsworthy, so the Biblical idiom (Matthew 3:12) is clearly relevant.

Perhaps the most serious question that can be raised about the use of idioms as a whole can be illustrated in example [24]:

[24][Headline:] Shenanigans steal the show
It only takes one bad apple to ruin the bunch, or, in the case of a few immature Missouri Southern students, a whole row of them to steal the show. While Southern deserves its fair share of brownie points for trying to be sophisticated in hosting a lecture by Hazel O'Leary, the College gets docked in the bonus round for failing to administer an IQ test at the door.179 (The Chartnet, Missouri Southern State College)

179 Docked in the bonus round sounds like a potential idiom, but the WebCorp search engine failed to turn up any examples of it from the Net, although in the bonus round seems to have a tenuous hold: “Heading Into the Bonus Round – in Space” A bonus round is something one usually associates with the likes of a TV game show, not a pioneering deep space mission. ‘We are definitely in the bonus round,’ said Stardust-NExT Project Manager Tim Larson of NASA's Jet Propulsion Laboratory in Pasadena, Calif. “This spacecraft has already flown by an asteroid and a comet, returned comet dust samples to Earth, and now has almost doubled its originally planned mission life. Now it is poised to perform one more comet flyby.” (http://www.nasa.gov/mission_pages/stardust/news/stardust20110209.html, Sept. 2, 2011, accessed June 29, 2014)
The editorial is two-pronged: this first (idiom-laden) half concerns the failure of some students to observe the rules of decorum when listening to an invited speaker; the second (idiomless) half critiques the apparently rather vapid talk by the former US Secretary of Energy. All of the idioms are to be found in the headline and first few lines (with shenanigans echoed in the final sentence). Aside from continuing the tone set by the headline’s shenanigans, however, it is difficult to see what the idioms rhetorically add to the text. The potential echoes set up by these idioms are not at all exploited. One conclusion – which is really a disguised judgment of taste – is that this is just bad writing, or writing seeking a little flair, but given the results of this substudy, it seems much more plausible to accept that idioms need not function in a more integrated fashion. Indeed, it is hard to imagine that a short text such as [24], with only 262 words, would be able to integrate each and every one of a rapid series of idioms.

Since these idioms are so many and varied, we cannot treat them all individually (nor do we have enough material, as the overwhelming majority occur only once in this entire subcorpus). Nevertheless, to give some sense of the pragmatic functions idioms may fill in these texts, a few examples will now be considered in more detail. The functions considered are: metalanguage, evaluation, description, emphasis and paraphrase. These functions frequently overlap, as Simpson & Mendis point out (2003: 427, 428), an aspect referred to by Moon as cross-functioning (1998: 240-42).

A striking first point is that the editorial idioms are almost never metalinguistic, i.e. flagged in the text itself. Example [25] is one of only three clear such cases:

[25]It’s an old cliché, but you can’t judge a book by its cover because it’s what’s inside that counts. (The Creightonian Online, Creighton University)

---

180 Steen (2008) takes up the question of deliberate v. non-deliberate use of metaphors, an aspect which can of course also be applied to the use of idioms, and which may be instrumental in understanding more about the choice to include idioms in texts, but one which would take this work in an entirely different direction.

181 This discussion follow the analysis suggested in Simpson & Mendis 2003, with the reservation that they were considering spoken interactions, not written one-way communication (even if the Coll editors are clearly aware of an audience that may react). Thus, their function of collaboration is omitted, as not being of relevance here.

182 It can be argued, however, that the deictic in Example [36], that big war chest, should be construed as meaning ‘the one we all know about’, which is close enough to ‘the proverbial’ to be considered marginally metalinguistic. Cf. the old Glenn Miller hit That Old Black Magic (1942).
Here, it has both a metalinguistic marker (an old cliché) and a follow-up explanation, so that it can in addition be seen as a paraphrase. Direct *periphrasis* is also quite rare, with only 8 clear examples.

A much more common function is *evaluation*. Roughly 80% of these idioms have a fairly clear evaluative function, although this evaluation can clearly work in tandem with components from the surrounding text: 183

[26] Besides, who would want a scholarship to an Alabama institution whose programs have already been *slashed to the bone* by James’ asinine proposals? (*The Aumnibus*, Auburn University)

[27] Greedy people with big bucks are *pulling everybody’s strings*. (*The Orion Online*, Chico State University)

[28] When he found dissenting voices - including some of his own colleagues, among them the Head of Retail Services - he *shifted the goalposts*. (*Warwick Boar*, Warwick University)

[29] Last week, The Pendulum published the results of a four-month inquiry into who on this campus are the “movers and shakers.” (*The Pendulum*, Elon College)

[30] But NATO leaders should not be duped into believing the Yugoslav president has had a *change of heart*. (*Daily Free Press*, Boston University)

[31] The truth of the matter and what neither the president nor anybody else searching for answers in the wake of the Littleton tragedy is telling them, is that there’s no reason for students not to be worried. (*The State Press*, Arizona State University)

[32] Christmas eve service at church where families come, people are in a good mood and a sense of tranquillity is *in the air*. (*The Marcolian*, Marietta College)

The first three examples are clearly negative in their evaluations: [26] has the violent verb *slashed*, and its central image is of a deep chop or wound; [27] suggests that people are marionettes, and the operators are *greedy, with big bucks*; and [28] indicates that the rules of a (university politics) game are being violated. The next two are examples of positive evaluations: [29] identifies a list of important people (and continues by identifying further such leaders in the paper itself, praising its own); [30] has a primarily positive sense, although here the *not be duped* clearly indicates that some Dickensian conversion is not about to happen. 184 Finally, [31] and [32] are examples of

183 The spreading of information over more than one unit is a commonplace in construction grammar and Langackerian thinking, with analogs in areas such as grammaticalization and language change and even phonetics.

184 For the positive semantic prosody of *change of heart*, cf. e.g. *Oxford ALD7* under *change n*: “usually making you feel more friendly, helpful, etc.” The *OED Online* is more coy, merely noting that a change takes place—although all its citations are of a positive change.
Inherently neutral, non-evaluative idioms; they adapt to their surroundings.\textsuperscript{185}

Again, about 80\% of these idioms (not necessarily the same ones) are also in some sense \textit{emphasizing}, either by their very nature, or by their position in the text:

[33] This all comes in the home stretch though, and we should view Fob’s proposal as what it truly is: a lousy, last-ditch effort to turn this election around. (\textit{The Aumnibus}, Auburn University)

[34] They were cold-blooded, ruthless killers who methodically carried out a pre-mediated plan for mass murder and maximum chaos. (\textit{The Georgetown Independent}, Georgetown University)

[35] We will only try rehab when we hit rock bottom. But then again, only 13 brutal deaths are not the rock bottom in America. (\textit{The University Register}, University of Minnesota [Morris])

[36] Today, wealth isn’t on the books as a requirement of running for office, but without that big war chest you’re not going to be elected for dog catcher. (\textit{The Orion Online}, Chico State University)

[37] The University hit a grand slam in the softball game of public relations with a gauzy Newsweek story on college admissions that depicted the College as an Eden where only the most worthy could hope to be admitted. (\textit{Chicago Weekly News}, University of Chicago)

[38] Metro’s Director of College Communications, Carrie Schafer, said she bent over backwards to accommodate the request, but apparently she didn’t bend far enough to accommodate state law, which says such requests should be filled immediately if the records are not in active use or storage. At most, the law allows three days for compliance. (\textit{The Metropolitan}, Metropolitan State College of Denver)

In [33], the temporal aspect is backgrounded by lousy and the image inherent in the process suggested by last-ditch (i.e. just before the enemy seizes all your defense lines and you are overrun); [34] also relies on conceptual metaphors such as VIOLENCE IS HOT, where premeditation (COLD) is seen as more evil (as is its legal status: premeditated murder is considered murder in the first degree), so that ruthless is backed up by methodically and premeditated; [35] the addition of rock indicates bedrock, as deep as you can sink, while the verb hit indicates a rapid descent. Citation [36] is a belittling example of hyperbole: the election is to about as trivial an office as can be imagined, and even there a big war chest is necessary. All of these are negative, in some sense, but [37] offers an example of a positive emphasis: a grand slam is the best result possible for a baseball or softball hitter, bringing in the maximum number of runs for the team; it is backed up by Eden, while contextually being undercut by the suggestion of fluffiness inherent in gauzy.

\textsuperscript{185} The phrase \textit{in the wake of} is probably already on its way to grammaticalization as a multi-word preposition. By the same token, it may serve as an example of a non-emphatic idiom, unless reinforced by a noun like \textit{disaster: in the wake of the Estonia disaster}. 

94
Emphasis by position has already been mentioned above, but repetition, although infrequent, should also be noted, as in [38], where the idiom’s literal meaning is exploited to suggest an unsatisfactory degree of bending, i.e. non-compliance with the law. Here, repetition is humorous, but in the following excerpts from one and the same editorial, the idiom—quite unusually—becomes a text-structuring device:

[39] We have taken a stand on many issues. We have pointed the finger when it’s appropriate. …
[40] What feedback am I hearing now that we have made this transition? The Technique is too sensational. The Technique is only out to get people and only points the finger.
[41] Most of the comments that I hear about the Technique’s “sensationalism” are coming from some of the people that the paper has pointed a finger at. (The Georgia Tech Technique, Georgia Tech)

Turning to other aspects of these idioms, we may first note that all 148 idioms are used in their standard sense, with the possible exception of [20], above, which stretches the meaning of follow suit. There are, however, a few clear examples of semantic specification, as in [42], or idiom-breaking, where the idiom is deliberately inverted, as in [43] and [44]. Irony is occasionally found, as in [45], and occasionally an elaboration, as in [45]:

[42] The same First Amendment that protects journalism and even student journalism was under assault, and [Dr. Jerry] Brown fought fire with sharp intelligence and facts. (The Auburn Plainsman Online, Auburn University)
[43] [Headline:] Jimmy Rane goes with board’s grain. (The Auburn Plainsman Online, Auburn University)
[44] Even the silver clouds seemed to have a dark lining. (The Cavalier Daily, University of Virginia)
[45] But remember most of all, like everything I’ve written in the past, every word is gospel. (The Falcon, Seattle Pacific University)
[46] The NRA, now in hiding to protect its public image, will soon saddle up even higher on its horse, negating any chance of reasonable gun reform. (The State Press, Arizona State University)

Example [42] is an interesting example of replacing the marker, or shell noun, fire with a much more detailed list of (in this case) good qualities, as he defended the First Amendment; [43] is from the same editorial as [20], and uses the idiom in a headline to indicate that a newly-appointed university trustee will probably fail to think for himself, but the idiom of ‘going/cutting against the grain’ is not followed up in the subsequent text. Example [44] is of course inverted, here in order to indicate that there will be further trouble in the Palestinian-Israeli conflict; like [43], it is not used further in the text. And [45] may serve as a humorous example of irony, since the text is about gullible people, whereas the humor in [46], with cowboy
and posse images being called forth by *saddle up*, is sharply dampened by the writer’s hostility to the National Rifle Association (in connection with the Columbine shootings): phrases like *in hiding*, *protect its public image* and the concluding *reasonable* all indicate a distancing from the NRA.

In conclusion, the study of the idioms found in this subgenre of newspapers, the editorial, reveals several aspects of their usage. These editorials are somewhat idiom richer than other newspaper subgenres, but not strikingly so. They do not exhibit striking patterns of position within the editorial, such as its headline or introduction, but have a relatively random distribution within these fairly short texts, and are rarely repeated. Roughly half of them have a strong conceptual and metaphorical link to the central theme(s) of their editorial, and only a very few seem to have no such link at all. They are normally used both normatively (with a strong tendency to have a negative valence) and for emphasis (by being striking images). In sum, they remain an optional resource, one that can be skillfully exploited in the larger context, or the text as a whole, but need not be. It is for this reason that we now turn to a different type of investigation, one where the individual idiom itself is studied in greater depth.

1.7 A Pedagogical Application: Exploring the glass ceiling

Since the application of idiom and variation to teaching situations has been one of the driving forces behind my interest in corpus work, this section will provide an extended example of how idioms from corpus material can be used to help students sharpen their analytical skills in several areas of linguistics by working with authentic texts. We will work with a single idiom, *the glass ceiling* ‘invisible barrier to women’s promotions’ as it was actually used with second-semester students at the English Department, Stockholm University for a number of years. The students were given fairly minimal instructions: they were supplied with a two-page list of text snippets using the idiom, without any explanation of the idiom itself, and told to examine these instances and be prepared to discuss what these extracts suggest about the idiom.

This exercise uses what is essentially a typical corpus or data-driven approach: the search phrase *glass ceiling(s)* was used to retrieve all instances of that string from a given corpus, in this case the *Broadcast News* CD, with transcripts of all its American radio programs from March 1995 to February 1996. To simplify matters (and fit on a two-page handout!), the additional three or four instances that involved a physical ceiling made of some kind of glass (with no idiomatic sense implied)

---

186 To simplify matters (and fit on a two-page handout!), the additional three or four instances that involved a physical ceiling made of some kind of glass (with no idiomatic sense implied)
publishers, who package the CD as a one-year reference CD, with its own built-in search engine.\textsuperscript{187} What is at issue here is what to do with the instances that such a search turns up. Only the phrase itself was considered; no attempt was made to look for variations.\textsuperscript{188} In addition, in a few cases, the citation was expanded from the standard number of characters, so as to provide sufficient context for an unambiguous interpretation.

As for the idiom itself, one reason for choosing it was that it was relatively frequent during this time span, ensuring that we would have a large number of examples to work with: 36 instances provides a fair range, without becoming unmanageable in a classroom session. As an accidental bonus, 36 items can be evenly parceled out to 4 or 6 groups, as you wish, for intensive small-group study before a discussion with the entire class.

If we consider the idiom itself, its origins have been traced back to the mid-1980s.\textsuperscript{189} It is listed in CCDI (1995: 163) under glass, but is not listed in e.g. ODEI. Learner dictionaries now routinely list it (Oxford ALD7, Longman LDOCE4, Macmillan MED, Cambridge ALD3), and it has become a sufficiently technical term to be found in Wikipedia.\textsuperscript{190}

Here are the instances from Broadcast News 1995:

dwffd

---

\textsuperscript{187} This is normally the case with newspaper and similar CDs; they were designed as information-retrieval devices, not as linguistic corpora, so that metainformation about e.g. the total number of words on the CD is often impossible to obtain directly. On some newspaper CDs, the text is stored as .txt files, so that it is possible to use e.g. WordSmith or AntConc; others have compressed files, so that only the proprietary software can be used to query them. For a contemporary discussion of what such CDs offered in this field at the time, see e.g. Minugh 1999 or Granath 2007. For more modern reviews of corpora in the classroom, see e.g. Bennett 2010 and Reppen 2010a.

\textsuperscript{188} All of the instances listed here contain the string glass ceiling(s), but in their proximity some additional “idiom-breaking” does occur, as in the steel ceiling of example 28 or in example 34, where glass ceiling is paired with sticky floor, which in this context is clearly part of the scenario envisaged by the glass ceiling metaphor.

\textsuperscript{189} OED Online Draft Addition 1997 cites an instance from 1994; Merriam Webster Collegiate Dictionary (1993: 495a) lists it from 1986. In 2009, Wikipedia had a comment that the term was common in Hewlett-Packard circles in the late 1970s, but that comment is now (April 2014) no longer present in its article about the “glass ceiling”.

\textsuperscript{190} Scandinavian readers may note that the idiom was not found in Norstedts stora engelska-svenska ordbok (1993), but appeared in Norstedts engelska-svenska ordbok — professionell (2010), as ‘glastak’ (literally “glass ceiling”), with the note ‘mainly transferred’, without any further explanation. Svensk ordbok utgiven av Svenska Akademien (2009) does not list it, although it has occurred in Swedish since at least the early 1990s, and is easily documented: 4 of the first 10 (out of 238) occurrences of glastak in Dagens Nyheter’s search engine (accessed April 24, 2014) used it in the idiomatic sense. Collins Robert French Dictionary, 9th ed. (2010: 1423b), on the other hand, merely provides the gloss “niveau professionnel où les femmes ont tendance à plafonner” [“professional level where women have a tendency to wind up”].

97
The Case of the Glass Ceiling

(1) a 15-piece, all-woman band named DIVA, which is making its mark in jazz and breaking through a musical glass ceiling.
(2) but they affirm the reality of closed doors, tight doors for blacks and browns and glass ceilings for women. How can you just kind of ignore those facts?
(3) but we’re not seeing parity in managers and supervisors. We have a long way to go, and there’s still the glass ceiling.
(4) Commentator Laura Archer Pulfer says her generation of women has been trying to shatter the glass ceiling for a long time, but she has high hopes for her daughter.
(5) even when she considers what might go wrong for her in the business world - encountering the stubborn glass ceiling for instance, moving up only so far and no farther.
(6) Glass ceiling Not Quite Broken: Laura Archer Pulfer compares the ‘discomfort’ of childbirth to the ‘discomfort’ of bumping into a glass ceiling.
(7) Grant says British institutions have a de facto glass ceiling that bars advancement by blacks.
(8) Guests Jo Ellen Allen, Celinda Lake, and Sherry Bebitch Jeffe discuss the so-called `glass ceiling’ for women, especially in Congress, and whether affirmative action is still a necessary program for women.
(9) He found in that review proof of race discrimination and sex discrimination, which is illegal, artificial glass ceilings for women and closed doors for blacks and browns.
(10) her head against an armor ceiling. Maj. LILLIAN PFLUKE, U.S. Army: It’s not a glass ceiling because you can’t even really see through it.
(11) I think the best thing about it is that it has made a- it’s sort of shattered that glass ceiling that Steve Forbes was under, that he had the ideas that appealed to Americans, but he couldn’t win any elections.
(12) Injustice is not a black thing, you know? Talk to a woman who has approached and hit that glass ceiling. That woman does not have to be a black woman.
(13) JESSE JACKSON: It seems that the glass ceiling, basically for white women, and a lower ceiling for blacks and the hole dug by the government begs for some remedy.
(14) LARRY KING: What do you mean by the glass ceiling? 1st CALLER: Well, the glass ceiling in business is that they only
let women go up so far- LARRY KING: Oh, middle management.

(15) LOIDA NICOLAS LEWIS: There is a glass ceiling, except that in my case, I inherited it. But for many women, for many people of color, it exists very definitely.

(16) I get some of the flat tax out there. So he’s done that to a certain extent. But that’s almost a little bit of a glass ceiling. You can only get so far and you get stopped.

(17) men and minorities are grossly under-represented in top management positions. The so-called ‘glass ceiling’ is apparently difficult to push through at the top of the corporate ladder.

(18) Most of those [chief executives] interviewed said they believed the glass ceiling was no longer a problem; it would be shattered in a matter of time.

(19) Now we just have to explain to our daughters about the discomfort of banging your head against a brick wall and a glass ceiling.

(20) PETER JENNINGS: When we come back, why so many women and minorities still bump into the glass ceiling.

(21) ROSEANNE BARR: -the true fact of it is, for you women out there who are listening, there really is no glass ceiling, you can go as far as you want to go, and nobody can stop you, as long as you have it in your head that that is possible and you can do it

(22) Sen. ALAN SIMPSON: He was the man who fought for job equity and the crashing in of the glass ceiling for women in this country.

(23) So the debate is still at that level. So do you feel you have to get out there and smash that fist through that glass ceiling

(24) The Glass ceiling Report, which came out earlier this year, clearly indicates that gender discrimination persists in this country.

(25) the invitation to the political dance, and in the next segment, decide whether they’re ever going to be able to dance on that glass ceiling, so thank you

(26) The Labor Department’s own glass ceiling commission reports that women and minorities make up about 57 percent of the senior managers at Fortune 500 and Fortune 1000 companies.

(27) There’s a real corporate culture change reflected there, and at the same time, although we haven’t broken the glass ceiling, women are opening small businesses in record numbers.

(28) There’s that glass ceiling you hear about all the time. It’s really here, and it’s more of a steel ceiling.
(29) they remained at low and mid-level positions. But as soon as they moved higher, or tried to, they ran into the infamous glass ceiling and, sometimes, outright hostility.

(30) those of us who’ve gotten into the political arena. Are we ever going to be allowed to dance on that top ceiling, that glass ceiling? Will we ever see a woman president or women in key political roles, and would that make a difference?

(31) Three years ago, then Secretary of Labor Elizabeth Dole commissioned a report on the glass ceiling.

(32) Tonight with Peter Jennings: Commission Issues Report on Glass Ceiling. The federal government releases a report on the glass ceiling concluding that it still exists. The glass ceiling essentially keeps women and minorities out of the top ranks in most corporations.

(33) we do not need this glass ceiling that prevents upward mobility; and we do not need this sticky floor that keeps so many women- women of color, poor women, down at the bottom

(34) Will women accept or be enthusiastic about women in politics ever getting to dance on top of that glass ceiling - will we ever get to the top?

(35) Women Refs May Crack Glass ceiling in NBA Basketball Two of the best referees in women’s college basketball were invited to the NBA’s rookie ref camp.

(36) women to come up against the institutional church. And break that, you know, it’s a glass ceiling in business. It’s been a glass ceiling in the church.

There are a number of different ways to approach the data, but I typically choose to begin by asking for a definition of the idiom itself, plus specific examples which support such a definition; the interested reader is invited to pick out appropriate examples. The first approximation to a definition of course focuses on gender at the workplace, but students rapidly realize that there are other examples, such as (20), (26) and (32), which, by specifically including minorities, indicate that ethnic origin can also be involved, even if the primary focus remains on gender. Further testing for exceptions to this definition will turn up a few examples, perhaps the most spectacular being instance (11), where the term is applied to a millionaire white male politician and his struggle to be taken seriously as a Presidential contender, in other words, an excellent example of how a term can become generalized. This also invites us to consider the difference between a term’s core meaning and
a peripheral example: if only 9 of 36 instances do not overtly mention gender within these short citations, and only 4 make it clear that they are not referring to women’s situation, how seriously should we take the evidence that it has once been used about Steve Forbes? Is it a nonce use, or a first sign of a generalizing pattern to come? The transition from a slang phrase to a publicly (and politically) recognized term used in public discourse can be seen in the government report referred to in (26), (31) and (32), which brings in the question of official language versus the language of public debate and that of slang.

A technical point about learning to use corpora can also turn up when students begin to consider example (17), which appears to contradict all the other uses of the term: “men and minorities are grossly under-represented in top management positions.” The overwhelmingly most probable solution to this strange example is of course that the software extracts a sample with a specified number of characters (centered on the search string), and the original text “Wolmen and minorities…” has been truncated to the extracted “men and minorities…”, thereby obscuring the meaning. Students need to become aware of such pitfalls when using corpora.

Returning to the texts themselves, another approach brings in sociolinguisitics, since quite a few of these instances contain overt references to middle management (14), senior managers (26), top management positions (17) or quite simply, the top (34). Citation (33) also creates the sticky floor, which thereby makes explicit that what the glass ceiling is about is of course upward mobility for above all professional (i.e. middle-class) women: women who do not get tapped for even low-level managerial positions have no problem with the glass ceiling, as they are on the sticky floor. Suddenly, gender politics is intertwined with economic issues.

An analysis of the concepts underlying the words themselves suggests several interesting semantic aspects: glass is normally transparent and strong, yet brittle. Its transparency makes it invisible, which is useful to those who want to use it without admitting it exists (invisibility and deception come to mind)—it is thus possible to deny its existence as a barrier. Its strength is limited, however: beyond its tensile strength, it will break or shatter (both interesting verbs in this context). A ceiling is in turn normally a


192 A semantic analysis using characteristics or qualia such as ± WOMAN functions rather nicely here, in that loss of the characteristic + WOMAN, but retention of +RESTRICTION indicates how the generalization occurs. For a brief introduction to this area, see e.g. Saeed 2009, Ch. 9; more detailed presentations may be found in e.g. Pustejovsky 2000, 1995.

192 Citation 9 also mentions closed doors, a metaphor that is a faint echo of the signs that used to be posted at the door of various firms in New York, saying “No Irish need apply” and the like.
protective, delimiting element, the uppermost boundary of a room; it is not normally thought of as being too low—but there is also the aviation and weather sense of a cloud ceiling, the height beyond which you cannot see. This is activated when the ceiling is invoked in its limiting aspect. The semantics for both terms can be varied by semi-repetition: we find a steel or armor ceiling and a sticky floor or brick wall directly linked to a glass ceiling.\(^{193}\)

If we look at collocations and semantic prosody, we find an interesting set of verbs used with the glass ceiling, none of them obligatory: we find both neutral \{encounter (and possibly bump into)\} and relatively violent or painful verbs \{break, shatter, hit, bang, crash, smash\}; latent in all of the latter is some form of energy in conflict with the glass. In addition, verbs like break or shatter can have an agent ("We’ll break that glass ceiling") or be in the passive ("the glass ceiling will be shattered"); analysis of active/passive, transitive/intransitive is particularly interesting if we consider the role of the agent from a social point of view.

Working with scenarios,\(^{194}\) there are clearly three involved: (1) the glass is invisible, and hence its proponents can claim it does not exist (and therefore cannot be attacked or removed); (2) the glass is strong enough to function as a ceiling, i.e. a barrier that holds women (and minorities) down; (3) the glass is brittle and can be destroyed, thus liberating women (and minorities) from its constraints. The idiom’s metaphoric world and its implications are thus available for discussion.

In addition, there are a small number of items with a rather different situation: scenario (4) involves dancing on the ceiling, where the glass element is basically ignored and the ceiling de facto becomes a floor. Assuming that dancing is a pleasurable activity, this fourth scenario is a joyous one, one which transcends the conflict inherent in the glass ceiling, not least because dancing is either gender-neutral or at the very least, more strongly associated with women than top managing positions are—and of course, the dancing takes place above the glass ceiling. Here, a grammatical observation may also be made: all three instances of scenario (4) (examples 25, 30, 34) contain future markers (will, going to be); two are questions and the third an indirect question; all three use ever, a marker of a potential future state. In other words, all examples of the (highly positive) fourth scenario are set in

---

\(^{193}\) This opens the door (as it were) to analysis of compounds: glass (i.e. a type of ceiling) can clearly be more easily modified than ceiling, and still retain the sense of being the same ("broken") idiom.

\(^{194}\) For cognitive scenarios, see standard works such as Fauconnier & Turner 2008, Clausner & Croft 1999, Langacker 1987, Lakoff & Johnson 1980.
an unreal world, one that has yet to be realized. This approach can of course quite easily be expanded into a cognitive one.\textsuperscript{195}

At a discourse level, one can then additionally note that this particular idiom is primarily used in the manner of a shell noun (Schmid 2000; cf. Aktas & Cortez 2008, as well as Flowerdew 2003 on ‘signaling nouns’, Hunston & Francis 1999), i.e. summarizing a particular scenario, but with the unusual characteristic of being rooted in a specific social scenario. It consistently holds an important place in these discourses, and thus is different from most idioms, which are not bound to the context in which they are used. In addition, only instances (21) and (18) question the existence of the scenario, and none make fun of, mock or disparage it. (This is not to claim that it cannot be made fun of, merely that our samples do not include any such example.)

This particular set of examples of a single idiom from a single corpus, although not full texts, provides a wealth of opportunities to explore how the idiom works and can interact with the society that fashioned it. It also neatly illustrates how apt John Sinclair’s previously-quoted observation was, about the need to have numerous examples of a word or construction, to see how it works in practice. His rule of thumb, it will be recalled, was 50 instances, and here we had 36 (40, if we include non-idiomatic instances literally about a physical ceiling made of glass), which was barely enough to include an example of how the idiom might be expanded beyond its core social groups. What this use of a rich set of examples can show, however, is many of the different relationships this deceptively simple idiom can enter into. Ideally, we would like to examine each idiom in the same depth, something that is now routinely done with lexical items when compiling dictionaries.

\textbf{1.8 Conclusions and Future Research}

Our first research question, \textit{how the data in a coherent, moderately-sized corpus of English written by young people, based on the English-language online college newspapers then in existence, compares to data from other then-existing corpora} (Papers 1 and 2), can in general be answered in two steps. First, it was—and is—possible to create such specialized corpora and compare them with the major corpora available today. Moreover, even today the \textit{Coll} Corpus remains unique, in that it documents the language of this particular group of middle-class young writers through-

\textsuperscript{195} “It is thus at the level of scenarios, rather than at general domain-level, that attitudinal biases and political preferences of discourse communities become discernible. The main reason for this seems to be that scenarios provide a sufficiently rich conceptual structure to be argumentatively and rhetorically exploitable” (Musolff 2006: 35).
out much of the English-speaking world, and with a size sufficient to allow numerous specific comparisons with other corpora. Although the Coll corpus is largely composed of university-related news and sports reporting, there are other strands in it, including a fairly large chunk of creative writing—and all of it was intended to be read by their peers.

The Coll corpus has made available material that is easily searchable via e.g. WordSmith, even if it has not yet been carried through to the elegant types of user interfaces found in e.g. COCA or COHA. This is again not to challenge the professionally created linguistic corpora (corpora I use almost daily), but will hopefully serve as a useful reminder that individual students and scholars with limited technical expertise can very well mine the Internet to create their own specialized corpora for various investigative purposes—a form of empowerment that is not to be underestimated, particularly in areas such as detailed studies of specific, narrow genres, where the major corpora are largely irrelevant, except as a reference against which to see your new data. The Coll corpus was not particularly onerous to create, even using the relatively primitive search and retrieval tools used then, and a Coll 2015 or the like can certainly be created more easily, quickly and automatically in the future.

This type of individual research into newspaper texts may in the future assume even more importance, as research into journalistic English in general, which experienced a brief golden era of access via annual newspaper CDs, is now only possible in a historical context, since newspapers no longer publish back-issue CD collections. Instead, their assumption is that researchers will search for information via keywords in the newspaper data base, which is of course monumentally inefficient for linguistic searches. Henceforth, it appears likely that only researchers who can pry loose entire multi-year chunks of text for inclusion in corpora like COCA or databases like the Bank of English (presumably primarily for dictionary-making purposes) will be able to tap into these large reservoirs of publicly-read text.

To some extent, this loss can be balanced by modern corpora compiled directly from the Internet; they are both rapid and huge, and not particularly hindered by copyright problems. In addition, these databases (pioneered by the SketchEngine team under Adam Kilgarriff) will bring us a much richer access to English, in terms of geography and domains. And of course the WebCorp approach allows us to efficiently search the Web for examples (but, as noted above, not frequencies).

Turning to research that might continue the Coll work, the first and most obvious project is a COLL2. This year marks the 15th anniversary of the original data collection, or a half-way point to the 30-year interval that marked the Brown-Frown/LOB-FLOB group. Given that there now are at least triple the number of online college newspapers available, it would be exciting to create a new, much larger version, and see what changes (and stabilities) can be seen so far. This new version should of course be properly
entered into a retrieval system that is easily and publicly available. At the very least, issues should be included from the entire academic year, and the metadata should keep track of this. At the same time, the original COLL corpus could be fed into the same system, so as to facilitate comparisons. Copyright issues no longer seem the insurmountable barrier they were in the pre-Facebook era, so that both Colls can be made available through e.g. the University of Oxford Text Archive.

The original Coll and its idioms can certainly be compared to other available corpora, particularly as seen in Paper 2. In a broader sense, given that we now have a number of large corpora, covering the entire English-speaking world, and a time span of several centuries, much of the data found in Coll can be re-examined and placed in this much larger geographical and temporal context. Ongoing changes in e.g. the English modals are one obvious example, as is the continuing expansion of snuck at the expense of sneaked, where these younger speakers appear to be in the vanguard of change.

However, such comparisons, as we have seen, are subject to major caveats about their composition. It is all too easy to be misled by their labels into assuming that surface equivalences are true equivalences, and in at least one sense Coll has no equivalents: there is no other corpus that focuses on unforced online college student writing. At the same time, this is not a problem for Coll alone, but one which is typical for small specific corpora. Beyond that, there is the larger problem endemic to corpus work as a whole, that of representativity and sampling techniques, including how to include the question of impact on readers (Leech 2007: 144–45 speaks of “atomic communicative events” that can be summed); clearly some such measure is important for newspaper and magazine corpora. These solutions must be for the field as a whole, using more sophisticated statistical methods and further corpus data, but also further reflection on such knotty issues as representativity.

This also entails screening out duplications, as numerous articles appear to run for several issues.

Note, however, the sobering details of the COCA copyright disclaimer (available via the help screens at http://corpus2.byu.edu/coha.

The list of key corpora begins of course with the BNC, and probably should include the Time Corpus (BYU), but most importantly includes the COCA and COHA corpora (BYU) for American English, and the brand-new GloWbE and Google corpora (also from BYU). For those who have sufficient funding, the Bank of English (Collins), the OEC (Oxford) and (especially for other languages) the SketchEngine web-based corpora are invaluable compliments.

In terms of corpus impact, the BNC takes pride of place, with Brown/LOB probably being the next most important, as they have affected several generations of both theoretical and applied linguists, not to mention lexicographers. Limited access has prevented the Bank of English from having an equivalent impact.

The preceding should probably be read as a political statement, but probably remains true, for all that. If it is correct, one direct consequence is that the next generation will be heavily
The second and third research questions, to examine idiom frequencies in various corpora, and whether certain subsets are more common in different parts of the corpora, led to Papers 2–4. Paper 2 reexamined the Coll corpus, to see which idioms appeared in college student newspaper writing, and in which configurations. One possibility was that young users use fewer idioms than writers in general, but this did not seem to be the case. The frequency figures arrived at do not seem out of line with those found in general (Moon 1998) or for e.g. newspapers (Minugh 1997). In terms of geographical distribution, the Coll corpus (unlike other then-available corpora) was dominated by North America, to roughly the same extent as demographics indicate, and indeed the items labeled NAm in CCDI were well represented. The Coll corpus indicates that there is a wealth of data to be looked at about how young people use idioms—and this can certainly also be complemented via blogs, computer game chat sites, and other potential sources on the net. One specific question that may help answer this is whether “inventive substitutions” (as easy as torte, quaking in your Doc Martins) are more likely to be found in humorous texts with other verbal fireworks, or are tightly linked to specific cultural settings where these hyponyms have symbolic value; this remains to be investigated.

Paper 3 faced a dilemma often found in corpus linguistics: when looking for a relatively low-frequency phenomenon, the need to amalgamate data from different corpora makes it difficult to answer the question of how the phenomenon works in the individual sub-corpora or subsections of the corpus. Here, 150 of the 300 M words are derived from newspapers, another 50 from radio transcripts, and the final 100 from the BNC (10% of which is comprised of newspaper text). Dropping to the level of the individual idiom, in these 300 million words, there are often less than 10 instances of even the most common “anchorings”, which is not sufficient for detailed analyses of the idiom. Recent research by Steen on deliberate metaphors (e.g. Steen 2013) may also be useful in this respect, as contextual anchorings and other, similar modifications of idioms clearly have links to deliberate metaphors, in that there must be an in some sense deliberate authorial choice to add a textual anchoring or link, as the idiom itself does not call for such an action. Steen’s point is that a metaphor introduces a domain which is not otherwise part of the text, in order to (deliberately) comment on the text; with anchorings, it is

influenced by the freely-accessible corpora being produced by Mark Davies at Brigham Young University: COCA, COHA and the Google corpora.

200 College newspapers reflect the written norms of an only partially coherent social group (most of whom do not contribute to these papers), whereas blogs, chat sites and gaming sites are all likely to be more cohesive in their authors and members, but not necessarily in their written styles (except for a few socially significant in-group phrases and writing codes, as in uses of capital letters, smileys, or words like LOL, w8, U). As argued above, newspapers reach a level of sustained argument seen only in very sophisticated blogs.
the reverse: the anchoring instead introduces the contextual world into the idiom’s microworld (scenario), which usually is not at all related to the text. As yet, it is not clear what triggers this (relatively rare) link. In Study 3, the additional examples of anchoring provided via Webcorp trawls of the Internet suggest that there are many more possibilities than people realize, but that it takes a whole Net to generate them.201

Underlying Paper 4, the Time corpus is extremely uniform as to genre (assuming that Time in fact is the same animal over time), but diachronically may display changing frequencies. The choice to compile the data by decades was of course facilitated (steered?) by the software, which did not easily permit other choices, but this choice needs to be questioned.202 In particular, Figure 2 in Paper 4 strongly suggests that some form of shift happened at the end of the 60s and beginning of the 70s—the question being whether it was one of style or house policy, or reflects a change in the actual language (i.e. the users of American English) that had now reached the writers and editors of Time.203

Turning to our fourth research question, how these idioms vary in their realizations and textual uses, the findings indicate there did not exist a discourse pattern of heavy use of idioms in Coll’s headlines, openings or endings of these articles (Paper 2). These results, although clear, were quite unexpected, and need to be further examined, preferably in a Coll2, as well. There was, however, a clear preference for their use in genres expressing opinions (primarily editorials, “profiles” [e.g. interviews] and letters to the Editor), while idioms were far less common in above all news.204 It could be well worth examining the data in Papers 3 and 4 to see if this same pattern shows up in general newspapers and in Time.

Looking ahead to future research on idioms, several important strands immediately suggest themselves. They arise not merely through the need for further analysis, but also because of how the world has changed. To start

201 The Birmingham-based Webcorp software provides an excellent linguistic tool for mining the net for linguistic examples and patterns, but because it does not draw on a closed corpus, it cannot provide information about frequency. See http://wse1.webcorp.org.uk/home/.
202 Since all data in the Time corpus is identified by date, it is possible, although cumbersome, to extract the idioms and regroup them into other periods. Cf. the statistics discussion in e.g. Hilpert & Gries 2009.
203 The fact that it was the “old-fashioned” idioms which dropped sharply in frequency, but not the other categories investigated, suggests that it was the idioms themselves that were going out of style, rather than the use of idioms per se. This of course agrees with the labeling in CCDI.
204 For more detailed analyses of the function of individual instances in their text, the background sketched in Chapters 8 and 9 in Moon 1998 should prove invaluable. I am particularly intrigued by the suggestion that indirect cohesive devices such as puns and alliteration are sufficient to allow canonical idioms to be used as stand-alone headlines (Moon 1998: 291–93). Since this did not seem to be a major pattern in college online newspapers, it may well be more of a feature in British newspapers (tabloids in particular?).
with, a Coll2 would enable us to see if there has been a shift in ranking among the idioms used by young people, something particularly interesting for “old-fashioned” idioms—are they really disappearing?

For the question of “anchoring”, we may note an article (Stathi 2007) published the same year as Article 3. Using ten specific idioms, Stathi took a major step towards formalizing the patterns of adjectival modification in German VP idioms. The major thrust of her article is to provide a theoretical basis for optional adjectival use as a whole in idioms. Paper 3’s “anchorings” are the direct equivalent of her category “external modification” and partially of her “intermediate level modification”. Her detailed analysis is based on ten German idioms with a verb+noun structure. As with our Paper 3 data, the frequencies of her modified idioms were not even roughly the same, as they ranged from 1.5% (Wind) to 13% (Fettnäpfchen) of the total occurrences of the idioms, which is actually higher than that seen in Paper 3. Linking together this type of analysis with the data found in Paper 3 should prove fruitful.

Another area worth investigating consists of those idioms that contain strong elements of alliteration, rhyme and/or rhythm, such as by fair means or foul, the bee’s knees, the gift of (the) gab. Do such structures inhibit or even block various forms of modification? (For that matter, do they all behave as a group in this respect?)

The diachronic investigation detailed in Paper 4 can now be duplicated and improved by using the magazine section of COHA for the same years. The overlap between Time and COHA is large, but not impossibly so: articles from Time comprise 6.7% of COHA, or 27.6% of its magazine section, and therefore can be filtered out, producing a “Time-free” magazine subcorpus of COHA. It will then be possible to compare the non-Time COHA data with the Time data from the same years. Here, the material from the Leech-Millar discussion (Leech 2003, Millar 2009, Leech 2012) on modals

205 The clear adjective-noun mismatches in the former category are detected as “semantic clashes”, a more explicitly formalistic presentation than in Steen 2013 (2008) or Paper 3; the latter are adjectives which are “not totally semantically incompatible with the noun” (Stathi 2007: 97). Formally, she appears to be correct, but the motivation for choosing these adjectives often appears to come from the surrounding context (which is unrelated to the scenario of the idiom itself), so this is at most a question of degree, or of punning.

206 E.g. jmd./etw. nimmt jmdm./etw. den Wind aus den Segeln (lit. 'sb./sth. takes away from sb./sth. the wind out of the sails' = 'sb./sth. weakens sb./sth.' (trans. Stathi) or jmd. tritt ins Fettnäpfchen' lit. 'sb. steps into the grease pot' = sb. commits a gaffe/faux pas, sb. misbehaves' (trans. Stathi; more idiomatically, 'sb. puts their foot in it').

207 To repeat, this is duplication, not in the sense of the reanalysis of the same data, but that of asking the same questions of a parallel set of data (even if a broad range of American magazines cannot be literally said to be the exact equivalent of Time). The other two major currently available corpora, COHA and the BNC, are synchronic and thus not relevant here.

208 Particular thanks to Mark Davies of BYU for instantly providing me with the files for this calculation.
in the Brown family and Time will provide a valuable background, specifically as to the extent to which these two corpora can be compared in idiom research, as well as their actual data. All too often, corpus work cannot be duplicated with parallel data, so this is a high-priority project.

Another aspect worth considering is to look in detail in both Time and COHA at the uses of the originally religious idioms: can they be divided up into uses in a context that activates a religious context and those that do not? The idiom an olive branch seems a particularly good candidate for this type of review, as the Biblical link is only partial, at best, with the sense of ‘peace offering’ having now become its primary use: of its 42 occurrences in the pre-60s period in Time, the olive branch was associated with Noah only once, the other 41 instances involving the symbolism of peace. The follow-up question, of course, is: does this distribution change over time? (For a discussion on how intertextual awareness about the idioms a sea change and the iron curtain has developed, see Minugh 2007b.)

Looking back on the idiom investigations hitherto, it seems that we have now finally reached the threshold for detailed investigations of how individual idioms behave. For lexicographers, the challenge lies primarily in creating a modern idioms dictionary that better mirrors current geographical distributions and the new idioms from the 21st century. The most difficult work yet lies ahead, in the attempt to re-examine our various idioms in the way John Sinclair mentioned, and which was sketched in Section 1.6 and partially worked out for the glass ceiling in Section 1.7: take at least 50 examples of the idiom from different domain, see how it behaves, where and how it is used, how it can be varied, and whether this variation can be linked to its context and discourse functions. This is no little project, of course, but it would certainly help coax that nine-lived idiomatic cat further out of the corpus bag.

---

209 For an excellent brief presentation of how more advanced statistical analysis can be used on such data, see Hilpert & Gries 2009.

210 To be more specific: this is not a question of running duplicate searches on identical data with different search engines, but of using data from the same decades and genres, but different magazines. Neither search will prove the other “wrong”; discrepancies indicate differences in their use of the language.
References


—. 2004–. *BYU-BNC*. (Based on the British National Corpus from Oxford University Press). Available online at http://corpus.byu.edu/bnc/.

*Corpus Linguistics 25 Years On*. Amsterdam: Rodopi.


Emerson, Ralph Waldo. 1841 [1837]. ‘The American Scholar’.


112


Keats, John. 1816. ‘On First Looking into Chapman’s Homer’.


*Urban Dictionary*. Available at [www.urbandictionary.com](http://www.urbandictionary.com).


**Newspaper CDs (30–50 M words):**


Corpora (with approximate size):
Archer Corpus, 3.3 M words. http://www.manchester.ac.uk/archer/
Brown Corpus, 1 M words. ICAME CD.
COCA Corpus. The Corpus of Contemporary American English, 450 M words and growing. See Davies, Mark, 2008–.
COHA Corpus. The Corpus of Historical American English, 400 M words. See Davies, Mark, 2010–.
Coll Corpus. 3.7 M words. Contact david.minugh@english.su.se.
COLT Corpus, 0.5 M words. Now part of the BNC. http://torvald.aksis.uib.no/colt/.
FLOB Corpus, 1 M words. ICAME CD.
Frown Corpus, 1 M words. ICAME CD.
Google-GB, 34 B words. See Davies, Mark. 2011–.
Google-US, 155 B words. See Davies, Mark. 2011–.
ICE (International Corpus of English). http://ice-corpora.net/ICE/INDEX.HTM
LOB Corpus, 1 M words. ICAME CD.
London-Lund Corpus, 0.5 M words. Compiled by Jan Svartvik. ICAME CD, Oxford Text Archive (http://www.ota.ox.ac.uk/).
MYCASE Corpus, 1.8 M words. http://quod.lib.umich.edu/m/micase/.
Time Corpus, 100 M words, 1923–2006. See Davies, Mark. 2007–.
ukWaK (United Kingdom Web as Corpus). 1.3 B words. Via SketchEngine, http://www.sketchengine.co.uk/ (license required).

Software:
WebCorp (Live). http://wse1.webcorp.org.uk/home/
2 Papers

These studies are reprints from the various original sources. A few minor errors have been silently emended, and one duplicate graph in the original volume has been dropped. Otherwise, they are as originally printed.
2.1 Study 1: The Coll Corpus: Towards a corpus of web-based college student newspapers

0. Introduction. The successful completion of (for their time) very large corpus projects, such as the Brown Corpus and its British successor the LOB Corpus, or, more recently, the COBUILD Bank of English and the BNC, inevitably entails one major drawback: precisely because such corpora are so very useful, there is a very real danger that research at the time becomes concentrated on those particular corpora; in a sense, they come indirectly to define what a corpus “should” be. The title of the 1998 ICAME volume, Corpora Galore, thus serves not merely to indicate its wide-ranging contents, but also as a salutary reminder of the need for many different types of corpora and corpus investigations.

One of the more important potential sources of major “accidental” corpora consists of the annual CDs released by various newspapers. In particular, “quality” newspapers have always formed a major component of large corpora, but only recently has research begun into their potential use as stand-alone corpora.³

Their role as daily reading in middle-class households makes them particularly interesting, although they are clearly somewhat restricted in scope and subject.³

The corpus to be presented in this paper continues this tradition of seeking ways of finding additional sources of large-scale corpora. It is somewhat similar to standard newspaper corpora, but innovative on several accounts, as specified below.

1. The choice of corpus. Although detailed descriptions of English have traditionally been centered upon the standard of written educated British English, increasing emphasis from the 60s onwards has been placed upon opening up this descriptive world to more of the many varieties of English that actually exist. This in turn makes increasingly difficult the Chomskyan demand for reliance on native-speaker intuition.⁴ After all, few indeed are

---

¹ I would distinguish among designed corpora created by linguists (e.g. Brown, BNC), accidental corpora (e.g. newspaper CDs with search engines designed to retrieve news content) and opportunistic corpora (e.g. the Nixon Watergate tapes).
² See the author’s paper at the 1994 TALC conference (Minugh 1997: 67–82). Cf. the brief comments by e.g. Svartvik (Thomas & Short 1996: 9) or the more extensive discussion in Sinclair (Melchers & Warren 1995: 17–33).
³ The Sunday issues and suburban supplements are considerably more varied in this respect.
⁴ Chomsky’s celebrated lack of interest in corpus linguistics, and his preference for the data of native-speaker intuition alone, have of course been noted and objected to by corpus linguists;
the speakers of English who can provide us with reliable intuitive data about e.g. both British and American English.  

By its very nature, the collective enterprise called the English language thus seems to cry out for a corpus linguistics approach. But how are the appropriate corpora to be found or created, and how much are they to be shaped? The first large-scale answers, such as Brown and LOB, were designed and sampled corpora with a clearly defined geographical base, and their relevance as reference corpora lives on, not least thanks to FROWN and FLOB.

Our generation’s megacorpora, however, have broadened the base, perhaps most notably with the BNC’s inclusion of the COLT subcorpus of spoken teenage British speech. Nevertheless, the available multimillion-word corpora, as well as all available newspaper CDs, clearly consist of relatively formal written English, to an overwhelming extent written by established reporters and columnists, which unavoidably leads to the conclusion that these corpora decisively reflect the language of educated middle-aged and older writers.

It should, however, also be imperative to seek out further ways of investigating the language of the coming generation, as well. If we go back to Chomsky’s claims about intuition, it would seem that few academic linguists qualify as native speakers in the under-25 bracket. Let us therefore instead begin with the question of how corpus linguistics may be used to investigate this rather different major segment of English, and do so throughout the English-speaking world.

It seems clear that the written production of under-25s now takes place primarily in two contexts: that of the Internet and that of the educational system. It is probable that the home pages and chat rooms of the Net may be one source that bears more than a passing resemblance to some of the COLT exchanges (although perhaps with a rather larger male dominance), but the present paper will not pursue this more casual side of the Internet.

In the second context, that of the educational system, the only texts that are both abundant in number and aimed at the writers’ peers (rather than their teachers) are student newspapers. Since secondary schools are frequently expected to act in loco parentis, they are normally held accountable for the contents of their student publications, which places certain limits on their freedom. This by no means totally disqualifies them from our interest, since many of the phenomena to be discussed in this paper could also be examined there, such as the use of modals, gender-related language choices and much

---

see e.g. Stubbs 1996: 23 ff. or Ch. 1 of McEnery and Wilson 1999, and references therein. A more general discussion may be found in e.g. Chapter 1 of Van Valin and LaPolla 1998.

5 Obviously, speakers of local dialects may also be proficient users of standard varieties, and there are considerable numbers of e.g. BrE speakers who are also well acquainted with the English of, for example, the Indian subcontinent, the Caribbean or various areas of Africa.
more. A more serious objection is that adult editors (primarily the English teachers at the school in question) may also intervene when it comes to the language itself, and demand a more “correct” (i.e. more formal) style than their students.

College and university editors, on the other hand, have almost complete freedom in all such respects, and the Coll Corpus is therefore based on student newspapers from tertiary educational institutions. A typical claim to uncensored freedom may be seen in the proud disclaimer of the student online newspaper from Auburn, Alabama, the Aumnibus:

[1] No content on this page or following pages has been reviewed or approved by Auburn University Montgomery, and the Aumnibus is solely responsible for its content. … The Aumnibus is the student newspaper of Auburn University Montgomery and is produced entirely by students with no censorship from the faculty advisor or the administration.

A few caveats: Particularly for colleges financed and run by religious communities, this freedom may be somewhat more limited, but as regards the key question of the freedom of the press, it has been assumed that this freedom is fundamentally not at issue in the Coll Corpus catchment areas. As regards the age of the contributors, it is possible that a few of those involved are older students, but given the standard pattern of Western education, it can safely be assumed that the vast majority of contributors are under 25. (Possible exceptions include stories from the “Police Blotter” and, more pertinently, letters from professors and statements by the university or college administration.) There are a fair number of letters to the Editor and freelance articles, but it is clear that a large amount of this material is written by students of journalism, and especially in the smaller papers, only a few writers may be represented. On the other hand, since literally hundreds of newspapers are involved, no individual writers or publishing-house norms can dominate.

Since many corpus questions require large or very large amounts of data, it is highly desirable that these texts already be easily available in electronic form. That condition is fulfilled by college and university student newspaper publications on the Internet. The Coll Corpus is thus designed to explore the possibility that online university and college student newspapers may provide us with a large enough corpus to study the evolving language of the next generation, as it seeks to find a voice that speaks to its own, untrammeled by the restraints of publishing houses or copy-editors, and containing a wide variety of writing. Certain universities, particularly in the U.S., have a strong religious orientation or political commitments, and their tightly-

---

6 Australia, Canada, New Zealand, the Republic of Ireland, South Africa, the United Kingdom and the United States. Censorship is of course always a potential issue, although the very fact that the Coll material is available on the Net mitigates against it.
reasoned theological or economic discussions may be expected to balance some of the more scurrilous parodies and hedonistic aspects of other schools (names withheld by request).

One major drawback of choosing a corpus of this kind is that it is not immediately comparable with designed corpora, even their “journalistic” sections, or with the larger corpora such as the BNC (an objection that also holds for commercial newspaper CDs). The Coll Corpus will in addition almost certainly contain an unusually high component of parody and language playfulness, since its newspapers are considerably less stereotyped than commercial newspapers, and there is a significant component of student creative writing, both poems and stories. Nor is it likely that a full regional balance can be easily maintained, as will be discussed below. Nevertheless, the advantages of having large-scale access to young people’s written language makes the project well worth doing.

2. Collecting the corpus. The technical criterion for inclusion was that the student newspaper already be online, i.e. freely available on the Internet. The initial search for such papers quickly turned up several important search paths, including a relatively accurate listing of over 200 American college newspapers, complete with Web addresses; unfortunately, it has not been updated since May 1999. Canada was also relatively easy, but finding sites in other countries turned out to be a time-consuming task, not least because such compilations do not yet seem to exist. Yahoo’s layered searches sometimes help, but more often than not contain no information about the student newspapers that actually do exist on any given campus. Moreover, many official university websites either do not acknowledge their own student publications or bury the links to them in a page several levels down, under “/campus life/student union/activities” or the like. In the end, the Coll list nevertheless embraced some 300 Websites available in spring 1999, and there is every likelihood that this list will grow fairly rapidly over the next few years.

The method of collection used for the first version of Coll was embarrassingly primitive, yet relatively effective: with the Web and Microsoft Office open at the same time, a cut-and-paste technique sufficed in most cases. In

---

7 See also the discussions in e.g. Ch. 2 of Kennedy 1998.
9 Technical details: The multiple column layout of many web pages caused considerably difficulties, but in nearly all cases could be overcome by copying small segments one at a time. Even pages that appeared non-copyable could usually be copied by working with the coding-page version of the display. Only PDF files from Adobe Acrobat were unobtainable by this method.
this first version, the entire *Coll Corpus* of some 3.9 M words was collected on a Pentium or a Pentium II PC, with a number of texts being downloaded via modem. The entire corpus in its current form takes up only 55 megabytes and thus fits easily onto a Zip disk, and can be effortlessly handled by WordSmith running on a Pentium. Thus, *no sophisticated machines, software or technical expertise are actually required to assemble or explore such a corpus*. This is a point that deserves reiteration: it is fully possible for individual researchers and students working with modest means to collect their own net-based corpora. This is not to claim that access to such expertise or machines would not improve the resultant corpus—on the contrary, one of the major goals of this paper is to arouse interest among those better qualified to work with storing and organizing data of this sort, so as to produce a faster, more user-friendly environment for queries. Nevertheless, it should be emphasized that even without expertise, it appears fully possible to create thoroughly useable multimillion-word corpora within only a month or two!

As indicated above, all publications had to be *student-run* to qualify for inclusion, and thus not be official organs of or derived from outside the college or university. Therefore, all articles specifically labelled as being provided by wirepress services or reprinted from commercial newspapers were excluded; on the other hand, supplements containing, e.g. arts sections consisting of student fiction, were included, as were university documents published in these student newspapers.

The time factor turned out in fact to be extremely difficult to control. Some student newspapers are published daily, others weekly or monthly, some sporadically, and yet others simply leave on their website the last issue posted by a Web-oriented editor, even though he may have graduated three years ago! Many of the online papers that publish more sporadically retain material from issue to issue, giving priority to updating the news and sports sections. One specific drawback of the present material is that it was gathered during the late spring and early summer, a time when many papers post only partial editions, or the (somewhat special) graduation issue from June.

As Knut Hofland (Bergen) was kind enough to point out, a recent discussion on the Linguist List noted several programs that can more fully automatize this Collection process: see his link via http://www.hit.uib.no/corpora/2000-1/0358.html.

10 Technical details: All texts are stored by Word as .txt files, one per newspaper, grouped by state and then region. For search purposes, the regions (US, UK, etc.) are represented individually, as well as in one duplicate directory including all the files. Each file can also be accessed separately. Searches were conducted via Word or WordSmith, but can of course be carried out with other programs, such as WordCruncher. File sizes were all calculated via Word’s “Word Count” function.

11 The word “he” is used advisedly, most webmasters being male; there are a great many female reporters and editors, however, so that the corpus itself is probably considerably less gender-skewed than commercial newspapers.
As regards the past, some papers offer back issues from as far back as 1995 or 1996, which can imply 50–100 issues for a weekly, while others offer only an archive search for the past, while yet others apparently post only the current issue; a few have been withdrawn from the Net, and others that are claimed to be there never quite reached the online stage.

Looking to the future, there are various “Watch this site” signs, occasionally with comments such as that the site was last updated in January 1998. In addition, there are literally thousands of colleges whose student papers are not yet online, and many may be expected to add net versions as time goes on. Where issues have disappeared, it may prove be possible to contact the newspaper directly and obtain back files that way. A very few student papers have closed sites accessible to students only, but the politics of alumni financial support to education, at least in the United States, virtually guarantee that most sites will remain open. Finally, there is always the threat of technological change that could close this window of textual opportunity, such as conversion to purely graphic files.

As with all corpus research, a knottier problem is that of permission to use the material. In its current state, the Coll corpus has simply been collected, and thus is not yet available to the research community at large. However, a first E-mail request to 150 newspapers for permission to include it in a publicly-available corpus resulted in a dozen granting permission at once and one more after a follow-up letter; none refused permission. It may thus be assumed that gaining permission may be an onerous task, but should on the whole not be impossible.

In this initial version of the Coll corpus, each publication is represented by one issue only. The entire issue is included, but links backwards are not followed, even where the reader is invited to via hotlinks. It will subsequently be possible to greatly expand the corpus by relaxing this restriction (admittedly at the cost of loss of representativity). One potential solution is to establish both a balanced monitor corpus and a much larger, geographically-skewed ‘kitchen-sink’ corpus.

The average issue contains about 10–15,000 words, with the largest being just over 30,000; the shortest is the terse statement from the Univ of Southwestern Louisiana:

[2] Due to budget problems, The Vermilion will no longer be placed online.

Geographically speaking, no area of the United States failed to be represented, but that is not true of the entire English-speaking world (see below).

---

12 This feature is otherwise a major advantage of online publishing, as ongoing controversies can be presented from the original article onward, with hotlinks to each editorial, article or letter to the Editor.
In all, as Table 1 indicates, 288 newspapers are represented, heavily skewed in favor of the U.S (where the figures are presented by U.S. Census regions). This distribution is nevertheless not strikingly out of line with the population figures for the traditional, relatively monolingual English-speaking world.

<table>
<thead>
<tr>
<th>Country</th>
<th>Total wds</th>
<th>% wds</th>
<th>Target %</th>
<th>Pop. (M)</th>
<th>Papers</th>
<th>Avg. size</th>
</tr>
</thead>
<tbody>
<tr>
<td>United Kingdom</td>
<td>368,803</td>
<td>10 %</td>
<td>20 %</td>
<td>80</td>
<td>26</td>
<td>14,185</td>
</tr>
<tr>
<td>Australia</td>
<td>48,965</td>
<td>1 %</td>
<td>4 %</td>
<td>18</td>
<td>4</td>
<td>12,241</td>
</tr>
<tr>
<td>New Zealand</td>
<td>23,671</td>
<td>1 %</td>
<td>1 %</td>
<td>3</td>
<td>2</td>
<td>11,836</td>
</tr>
<tr>
<td>Canada</td>
<td>401,614</td>
<td>10 %</td>
<td>6 %</td>
<td>25</td>
<td>35</td>
<td>11,475</td>
</tr>
<tr>
<td>United States</td>
<td>2,980,289</td>
<td>77 %</td>
<td>67 %</td>
<td>270</td>
<td>215</td>
<td>13,862</td>
</tr>
<tr>
<td>Ireland/Eire</td>
<td>28,552</td>
<td>1 %</td>
<td>1 %</td>
<td>4</td>
<td>3</td>
<td>9,517</td>
</tr>
<tr>
<td>South Africa</td>
<td>25,963</td>
<td>1 %</td>
<td>0 %</td>
<td>2</td>
<td>3</td>
<td>8,654</td>
</tr>
<tr>
<td><strong>Totals:</strong></td>
<td><strong>3,877,857</strong></td>
<td></td>
<td></td>
<td><strong>402</strong></td>
<td><strong>288</strong></td>
<td><strong>11,681</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>US breakdown:</th>
<th>Total wds</th>
<th>% wds</th>
<th>Target %</th>
<th>Pop. (M)</th>
<th>Papers</th>
<th>Avg. size</th>
</tr>
</thead>
<tbody>
<tr>
<td>New England</td>
<td>363,799</td>
<td>12 %</td>
<td>5 %</td>
<td>13</td>
<td>23</td>
<td>15,817</td>
</tr>
<tr>
<td>Middle Atlantic</td>
<td>573,085</td>
<td>19 %</td>
<td>14 %</td>
<td>38</td>
<td>22</td>
<td>26,049</td>
</tr>
<tr>
<td>E North Central</td>
<td>547,370</td>
<td>18 %</td>
<td>16 %</td>
<td>44</td>
<td>42</td>
<td>13,033</td>
</tr>
<tr>
<td>W North Central</td>
<td>270,415</td>
<td>9 %</td>
<td>7 %</td>
<td>19</td>
<td>28</td>
<td>9,658</td>
</tr>
<tr>
<td>South Atlantic</td>
<td>450,238</td>
<td>15 %</td>
<td>18 %</td>
<td>49</td>
<td>32</td>
<td>14,070</td>
</tr>
<tr>
<td>E South Central</td>
<td>111,969</td>
<td>4 %</td>
<td>6 %</td>
<td>17</td>
<td>10</td>
<td>11,197</td>
</tr>
<tr>
<td>W South Central</td>
<td>165,891</td>
<td>6 %</td>
<td>11 %</td>
<td>30</td>
<td>19</td>
<td>8,731</td>
</tr>
<tr>
<td>Mountain</td>
<td>199,744</td>
<td>7 %</td>
<td>6 %</td>
<td>17</td>
<td>13</td>
<td>15,365</td>
</tr>
<tr>
<td>Pacific</td>
<td>297,778</td>
<td>10 %</td>
<td>16 %</td>
<td>43</td>
<td>26</td>
<td>11,453</td>
</tr>
<tr>
<td><strong>US Totals</strong></td>
<td><strong>2,980,289</strong></td>
<td>100 %</td>
<td>100 %</td>
<td><strong>270</strong></td>
<td><strong>215</strong></td>
<td><strong>13,862</strong></td>
</tr>
</tbody>
</table>

Table 1. The geographical distribution of online college student newspapers.

In the geographical sense, this is a paper about a future corpus, rather than the present-day possibilities: for many types of comparison that demand large corpora, such as low-frequency lexical items or constructions and most idioms, numerous areas of the English-speaking world are not yet sufficiently represented by student online newspapers, although that situation will inevitably change.

3. The beast itself. Nearly all student papers offer five major sections:
(i) news, often divided into campus news and general news
(ii) editorials, op-eds and Letters to the Editor
(iii) sports (with a major component of women’s sports)
(iv) features (including feature pieces, columnists and the odd horoscope)
(v) arts and entertainment (including film, CD, theater reviews)

Other areas that may be represented include:

(vi) university policy documents, decisions, visions and speeches
(vii) community news, including the “Police Blotter”
(viii) “Greek news”, i.e. news from the world of fraternities and sororities
(ix) classified advertisements
(x) publication policies, presentation of staff and E-mail response slots

Given the relative uniformity of writer and reader age and social status, it is clear that this corpus cannot approach the breadth of the BNC or COBUILD corpora, but that should not affect many of the comparisons that it will permit. It is the latter that will now be discussed.

4. Analyzing the material. Since the corpus merely consists of text files, the primary method of querying the corpus has been WordSmith searches. To enable comparisons, the results have been normalized to instances per million, which also facilitates comparison with the corresponding 1M word corpora from the 1960s (Brown, LOB) and 1990s (Frown, FLOB). In consequence of the small size of the Irish, Australian, New Zealand and South African components (taken together, they comprise less than 4% of the corpus material), they have normally been included with the UK component to form the category “non-North American”. In addition, there are potential comparisons to be made with the BNC and COBUILD corpora.

---

13 It should continually be borne in mind that these four corpora of written English cover a wider range of text types than the present corpus (although not as wide as either COBUILD or the BNC), so that the comparisons presented here must be treated with a certain reserve.
4.1. Modal verbs. The figures for the modals (per million words) are presented in Table 2. As might be expected, the frequencies for the major modals continue to develop along the lines indicated by the 60s to 90s Brown/LOB to Frown/Flob shifts, with can, will and the contracted ‘ll all gaining, while could, would and particularly must and might are even less frequent. The increasing reluctance to use may shown in Frown is only partially supported here, and the gain by will is much larger than anything indicated by the standard corpora.

<table>
<thead>
<tr>
<th>Item:</th>
<th>Brown</th>
<th>Frown</th>
<th>LOB</th>
<th>FLOB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokens</td>
<td>1M</td>
<td>1M</td>
<td>1M</td>
<td>1M</td>
</tr>
<tr>
<td>Conversion</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>CAN</td>
<td>1,913</td>
<td>1,997</td>
<td>1,801</td>
<td>1,999</td>
</tr>
<tr>
<td>can</td>
<td>1,751</td>
<td>1,737</td>
<td>1,655</td>
<td>1,772</td>
</tr>
<tr>
<td>cannot</td>
<td>252</td>
<td>193</td>
<td>318</td>
<td>240</td>
</tr>
<tr>
<td>can’t</td>
<td>162</td>
<td>260</td>
<td>146</td>
<td>227</td>
</tr>
<tr>
<td>COULD</td>
<td>1,762</td>
<td>1,655</td>
<td>1,729</td>
<td>1,766</td>
</tr>
<tr>
<td>could</td>
<td>1,589</td>
<td>1,472</td>
<td>1,604</td>
<td>1,564</td>
</tr>
<tr>
<td>couldn’t</td>
<td>172</td>
<td>180</td>
<td>122</td>
<td>198</td>
</tr>
<tr>
<td>could’ve</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>MAY</td>
<td>1,392</td>
<td>940</td>
<td>1,422</td>
<td>1,208</td>
</tr>
<tr>
<td>MIGHT</td>
<td>671</td>
<td>639</td>
<td>781</td>
<td>642</td>
</tr>
<tr>
<td>might</td>
<td>671</td>
<td>639</td>
<td>776</td>
<td>641</td>
</tr>
<tr>
<td>mightn’t</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Item:</td>
<td>Brown</td>
<td>Frown</td>
<td>LOB</td>
<td>FLOB</td>
</tr>
<tr>
<td>might’ve</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>MUST</td>
<td>1,016</td>
<td>670</td>
<td>1,114</td>
<td>816</td>
</tr>
<tr>
<td>must</td>
<td>1,009</td>
<td>663</td>
<td>1,094</td>
<td>803</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COLL</th>
<th>US</th>
<th>Non-US</th>
<th>Non-NAm</th>
<th>Canada</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAN</td>
<td>2,763,781</td>
<td>956,891</td>
<td>500,336</td>
<td>456,555</td>
</tr>
<tr>
<td>can</td>
<td>0.36</td>
<td>1.05</td>
<td>2.00</td>
<td>2.19</td>
</tr>
<tr>
<td>could</td>
<td>2,242</td>
<td>2,767</td>
<td>2,784</td>
<td>2,749</td>
</tr>
<tr>
<td>couldn’t</td>
<td>1,936</td>
<td>2,357</td>
<td>2,380</td>
<td>2,330</td>
</tr>
<tr>
<td>could’ve</td>
<td>135</td>
<td>134</td>
<td>104</td>
<td>166</td>
</tr>
<tr>
<td>may</td>
<td>306</td>
<td>411</td>
<td>404</td>
<td>418</td>
</tr>
<tr>
<td>COULD</td>
<td>1,084</td>
<td>1,161</td>
<td>1,243</td>
<td>1,071</td>
</tr>
<tr>
<td>could</td>
<td>969</td>
<td>1,056</td>
<td>1,129</td>
<td>975</td>
</tr>
<tr>
<td>couldn’t</td>
<td>111</td>
<td>103</td>
<td>110</td>
<td>96</td>
</tr>
<tr>
<td>could’ve</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>MAY</td>
<td>1,135</td>
<td>1,067</td>
<td>1,043</td>
<td>1,093</td>
</tr>
<tr>
<td>MIGHT</td>
<td>353</td>
<td>414</td>
<td>470</td>
<td>353</td>
</tr>
<tr>
<td>might</td>
<td>353</td>
<td>413</td>
<td>468</td>
<td>353</td>
</tr>
<tr>
<td>mightn’t</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Item:</td>
<td>Brown</td>
<td>Frown</td>
<td>LOB</td>
<td>FLOB</td>
</tr>
<tr>
<td>might’ve</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>MUST</td>
<td>484</td>
<td>439</td>
<td>452</td>
<td>425</td>
</tr>
<tr>
<td>must</td>
<td>483</td>
<td>436</td>
<td>446</td>
<td>425</td>
</tr>
</tbody>
</table>

14 Untagged versions of the four reference corpora were used, as the Coll Corpus is merely raw texts. This means that items such as can are given in toto, rather than merely instances of the verb, and are thus exaggerated, compared to items such as should or would, which have no noun homonyms.

15 One possible explanation is that much of the Coll Corpus was gathered during the part of the year around May.
These changes, together with the effective vanishing of ought and shall, and the literal disappearance ofshan’t, point to a familiar major problem (Mair 1998: 153–54): is there an actual change in the language, or merely a wider acceptance of less formal registers? More precisely: since relatively informal texts are to a greater extent being published today than a generation ago, we may merely be noting a shift in taste as to what gets published, rather than a real change in the norms of the printed language (formal texts would thus still be formal in the same way—but there would be fewer such texts). In the present case, this is complicated by the possibility that these young speakers have not yet mastered the more formal English used by their elders, but will so in the coming decades. It nonetheless seems likely that after several years of college or university education, these writers in their early 20s are not likely to change their writing habits, with the obvious exception of those that enter fields with archaizing demands, such as the law. Yet, if informal styles are indeed expanding into areas previously reserved for formal styles, a point which the much higher ratio of can’t to cannot would

---

16 This problem is similar to the apparent-time question of language change, as treated in e.g. Chapter 3 of Labov 1994.
17 Cf. e.g. Croft 2000: 72–73 on the distinction between competence (of the individual) and convention (of the speech community).
seem to indicate, how are we to interpret the surprisingly low number of instances of ‘ll, a form that would seem to be an obvious feature of a less formal style? This question will be discussed at length in section 5, below.

4.2. Lexical comparisons across the varieties. Table 3 provides statistics, normalized to per 1 M words, for a few of the traditional British/American vocabulary distinctions. The first group is that of the -st and -s endings in amid etc. While Americans (and Canadians) continue to completely reject whilst and only occasionally use amongst, it seems as if young Americans are accepting towards and actually prefer amidst—since this data is in sharp contrast to both Brown and Frown, it is a point worth investigating further.

Traditional BrE/AmE distinctions such as windscreen/windshield, lorry/truck and nappy/diaper provide few surprises, especially as the latter pair is not normally part of college life! As for other, less well-known lexical distinctions, the limitations of a “small” corpus become apparent: 4 M words are insufficient for such lexical items as gazump, blindside, or schlep(p); here, corpora of at least an order of magnitude larger are necessary, not least in order to check distribution between varieties. This claim would appear to be contradicted by the extraordinary number of American instances of the British slang term shag; they derive, however, almost entirely from reviews of a film called The Spy Who Shagged Me, featuring such forgettable characters as Felicity Shagwell and Ivana Humpalot.

Unlike the data seen for the established corpora, in the Coll corpus, all areas of English appear to be abandoning the mutated vowels in the comparative elder/older and superlative eldest/oldest. Only derivatives such as village elders appear to be surviving.

<table>
<thead>
<tr>
<th></th>
<th>Brown</th>
<th>Frown</th>
<th>LOB</th>
<th>FLOB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv. factor</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Br/Am:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>amid</td>
<td>14</td>
<td>17</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>amidst</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>among</td>
<td>369</td>
<td>451</td>
<td>304</td>
<td>249</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>US</th>
<th>Canada</th>
<th>Non-NAm</th>
<th>Non-US</th>
</tr>
</thead>
<tbody>
<tr>
<td>amid</td>
<td>6</td>
<td>7</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>amidst</td>
<td>11</td>
<td>4</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>among</td>
<td>247</td>
<td>197</td>
<td>92</td>
<td>142</td>
</tr>
</tbody>
</table>

18 Of the three, only blindside (from American football) is not in the Collins or Macquarie Dictionaries; gazump is not in Webster’s 10th ed. Idioms were not investigated here, but their frequency range appears to be 1–10 per 10M (Cobuild Dictionary of Idioms 1995:v,xvii, confirmed in Minugh 1999), which also indicates that the present corpus is as yet far too small for such areas.

19 The specifically American sense of shag, ‘to practice catching fly balls’, a baseball term, occurs twice in the American material, and the ‘shag rug’ sense once.
Table 3. Comparisons of Coll and the established corpora: Vocabulary I.
In the area of gender and lexis, it is hardly surprising that the ratios from the spoken part of the BNC are relatively consistently among the most gender-neutral ones, since it was balanced for speaker gender.\textsuperscript{20} To a noticeable extent the Coll corpus also appears to have more material about women than normal newspapers, as seen in Table 4. In Coll, spokesperson actually out-numbers spokesman, perhaps heralding a permanent change in our language. On the other hand, the terms are by a full order of magnitude more common in regular newspapers, indicating that this is a term firmly entrenched in the larger world of (male-dominated) politics. Certain other omissions are more startling, such as the extraordinarily low search results for acquaintance/date rape (4 and 11 instances, respectively, for the entire Coll corpus).

4.3. Lexis and syntax. As for items touching on grammar (Table 4), snuck continues to gain ground on sneaked, although here too a larger corpus would be desirable. There is, however, enough material to continue an earlier investigation (Mair 1998: 150) on the British willingness to accept prevent + V-ing without the preposition from:

\[3\] It is, however, a skill worth acquiring; the more fixations you have, the more there is to prevent you from being bored. Which is good. [Leicester University, Ripple]

\[4\] This should be done twice as the double lining prevents the custard leaking through and sticking to the tin. [University of Manchester, Grip]

Interestingly enough, the BrE college data is not nearly as conclusive as is the shift between LOB and FLOB would seem to indicate. Canadians, like Americans, seem not to be taking part in this shift. Since this change is not

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|c|c|c|}
\hline
Term & Coll & BNC & NYT 96 & IND 99 & LAT 96 \\
\hline
he/she & 2.14 & 1.79 & 2.41 & 3.01 & 2.28 \\
his/her/hers & 1.79 & 1.32 & 2.26 & 2.75 & 2.20 \\
spokesman/-woman & 2.24 & 6.43 & 2.26 & 3.75 & 1.95 \\
spokesman/-person & 0.77 & 15.66 & 138.14 & 24.00 & 36.97 \\
man/woman & 2.42 & 2.68 & 1.87 & 2.92 & 2.05 \\
men/women & 0.48 & 0.97 & 1.05 & 1.32 & 0.95 \\
\hline
\end{tabular}
\caption{Gender ratios in Coll and other current corpora. (Items in \textbf{boldface} indicate the most gender-neutral ratios for each category.)}
\end{table}

\textsuperscript{20} Aston & Bernard 1998: 33. This is apparently not the case for the much larger, written section (Aston & Bernard 1998: 28–31).
of American origin, it should be of particular interest to monitor the possible spread of this new structure to other areas of the English-speaking world; the present type of corpus offers one future way to do so.\textsuperscript{21}

\begin{table}[h]
\begin{center}
\begin{tabular}{lrrrr}
 & \textit{Brown} & \textit{Frown} & \textit{LOB} & \textit{FLOB} \\\n\hline
Conv. factor & 1.00 & 1.00 & 1.00 & 1.00 \\
sneaked & 6 & 2 & 1 & 1 \\
snuck & 1 & 0 & 0 & 0 \\
\hline
\textit{prevent + from} & 50 & 36 & 37 & 22 \\
\textit{prevent -from} & 0 & 2 & 7 & 20 \\
if...would have & 0 & 1 & 1 & 0 \\
\hline
mand subj & & & & \\
mand active & 45 & 64 & 2 & 22 \\
passive & 59 & 36 & 10 & 22 \\
should active & 8 & 5 & 50 & 38 \\
passive & 6 & 4 & 31 & 29 \\
\hline
US & 0.36 & 2.19 & 2.00 & 1.05 \\
Canada & 0.00 & 0.00 & 0.00 & 0.00 \\
Non- NAm & 2.2 & 2.2 & 2.2 & 2.2 \\
Non-US & & & & \\
\hline
US & & & & \\
Other & 54 & 55 & 41 & 22 \\
UK & 22 & 16 & 31 & 2 \\
Can & 4 & 0 & 5 & 4 \\
\hline
\end{tabular}
\end{center}
\caption{Comparisons of \textit{Coll} and the established corpora: Grammar I.}
\end{table}

A shift in the hypothetical past \textit{if}-clause that is now quite common in spoken American involves the double marking via \textit{would}:

\begin{itemize}
\item [5] Sprint coach and drill sergeant, Coco James, who is faster than a two-legged cow, summed up the meet with these words: “The team performed admirably under the conditions, and if we would have had preliminaries we would have done better.” [University of Minnesota (Morris), \textit{The University Register}]
\end{itemize}

This pattern was noted more than a decade ago by Quirk et al. (1985: 1011n.c), and has consistently been stigmatized in e.g. American college entrance exams.\textsuperscript{22} Despite its nearly ubiquitous presence in the spoken form, it is strikingly rare in these college texts, with only 7 US and 2 Canadian instances (normalized to 3 and 4, respectively).

\textsuperscript{21} In this perspective, the current relative paucity of data from many English-speaking areas should be weighed against the fact that only a very few of the \textit{Coll} student newspapers online in the United States seem to go back further than 1997.

\textsuperscript{22} In one such handbook, Brownstein & Weiner (1978: 336,338) provide a practice exercise where \textit{If he would have listened to me, he would not have applied for that position} is corrected as follows: “Sequence of tenses requires that the verb in an \textit{if} clause be in the past perfect tense. Correction: \textit{If he had listened to me...}.”
The mandative subjunctive \( \textit{v} \) the \textit{should} forms was a point dealt with in some detail by Marianne Hundt (1998: 159–75), building inter alia on Johansson and Norheim’s (1988: 27–36) earlier work with the \textit{Brown} and \textit{LOB} corpora. Examples [6] and [7] represent the two choices under consideration:²³

[6] Those two candidates, who lost big at the straw poll, can demand that Bush come clean all they want. [Iowa State University, \textit{Iowa State Daily}]

[7] Buhrow recommends that engineers who aspire to become brewmasters should take classes in biology, biochemistry, cereal chemistry, microbiology and fluids while earning their undergraduate degrees. [University of Iowa, \textit{Hawkeye Engineer}]

With their data as a jumping-off point, it may be noted that the present student data, while not as striking as the data from those corpora, clearly indicates that the mandative subjunctive is strongly preferred by Americans and Canadians over the \textit{should} construction. But it should also be noted that even in BrE college texts, the mandative form is far more common than the \textit{should} form, in sharp contrast to \textit{LOB} and even \textit{FLOB}. Moreover, a breakdown into active \( v \) passive forms indicates that the active now dominates, unlike the data from the previous corpora, which suggested that the (presumably more formal) passive was the more likely of the two.

<table>
<thead>
<tr>
<th>Source</th>
<th>Corpus size</th>
<th>Verb totals</th>
<th>Verb types</th>
<th>Hapax legomena</th>
<th>Hapax %</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{BNC}</td>
<td>100 M</td>
<td>6129</td>
<td>566</td>
<td>321</td>
<td>57 %</td>
</tr>
<tr>
<td>\textit{BN}</td>
<td>44 M</td>
<td>1549</td>
<td>205</td>
<td>129</td>
<td>63 %</td>
</tr>
<tr>
<td>\textit{Coll}</td>
<td>4 M</td>
<td>231</td>
<td>85</td>
<td>61</td>
<td>72 %</td>
</tr>
</tbody>
</table>

\textbf{Table 6.} The \textit{way} construction.

The \textit{way} construction, with various verbs that can fit in the V slot, e.g. \textit{She wisecracked/clawed/slithered/smiled her way through the crowd}, is an example of a construction that can be investigated in a corpus of this magnitude.²⁴ As Table 6 indicates, the pattern is the same as that found in larger corpora, such as the \textit{BNC}, with the expected higher percentages of single instances in smaller corpora. With no less than 85 different verbs to choose from in \textit{Coll}, there is nevertheless ample evidence of how variable the construction may be, including such creative surprises as

²³ Other choices also exist; see Hundt’s MAVEN article and the references therein.
²⁴ Cf. especially Goldberg 1995 on this construction (in the Fillmorean sense).
5. A question of formality. The above are only a small sample of the types of searches that may be run on this material. Another aspect worth examining is that of formality. This is a question that has repeatedly been raised, not least by Mair and Biber in numerous publications.\textsuperscript{25} For the Coll corpus, this is particularly acute, since student publications may exhibit a larger component of informal styles than many others. Some of the various types of evidence are presented in Table 7.

\begin{table}[h]
\centering
\begin{tabular}{|l|rrrr|rrrr|}
\hline
\hline
Conv. factor & 1.00 & 1.00 & 1.00 & 1.00 & 0.36 & 2.19 & 2.00 & 1.05 \\
\hline
\multicolumn{2}{|l|}{Formality:} & & & & & & & \\
\hline
albeit & 2 & 10 & 4 & 16 & 9 & 9 & 22 & 16 \\
lest & 17 & 10 & 3 & 8 & 3 & 0 & 4 & 2 \\
whereas & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
behove & 0 & 0 & 0 & 1 & 0 & 0 & 2 & 1 \\
behoove & 1 & 2 & 0 & 0 & 1 & 0 & 0 & 0 \\
granted [that] & 8 & 7 & 1 & 4 & 13 & 9 & 4 & 6 \\
whether it be & 3 & 1 & 7 & 0 & 9 & 4 & 6 & 5 \\
\hline
on & 6,719 & 6,898 & 6,912 & 7,122 & 7,189 & 7,272 & 7,469 & 7,375 \\
upon & 493 & 196 & 407 & 243 & 168 & 177 & 180 & 179 \\
\hline
\end{tabular}
\caption{Formality in Coll and the established corpora.}
\end{table}

Beginning with a series of formal items, such as \textit{lest} and \textit{albeit}, the present data matches that from \textit{FLOB}, confirming that \textit{albeit} is now well-established in BrE, while \textit{lest} is nearly absent from the college corpus. Of the 9 instances of the latter, only 5 may be said to have been used in a serious context, that of a Biblical or political quote (examples [9] and [10]), while the others range from the dubiously serious to the clearly playful (examples [11] and [12]):

\begin{itemize}
\item [9] Perhaps owing to the unfortunate Christian dictate of “judge not, lest ye be judged,” we have managed to bind our moral development to the shallowest of circles with this fatuous standard. [Carnegie Mellon University: \textit{The Tartan Online}]
\end{itemize}

\textsuperscript{25} See Biber, Conrad & Reppen 1998 for an overview.
When the Government breezed into office 12 months ago, they did so with the largest majority of any Labour government, lest we should forget. [The Edinburgh Student, Edinburgh University]

Parents avert your eyes lest something within this column sends you into a conniption fit. [Virginia Tech: Collegiate Times]

For example, one of my personal faves . . . er, favorites, is “jook.” It’s Bahamian for “stab.” So let this be a warning: beware lest ye happen to commit a verbal offense in my presence, for I may, perforce, be overcome by the insatiable need to jook you. [Princeton University: The Daily Princetonian]

Of the 40 instances of albeit, on the other hand, only [13] and [14] appear to be at all playful. Examples [15] to [17] illustrate its use (in a wide range of subject areas) as an alternative to although:

Living in Clinton for a year, albeit in some sadistically altered former barn whose stability has always been in question, has provided me with a sense of connection and familiarity with this place that campus life never could have supplied. [Hamilton College: SpecOnline]

The soldier in me finally became too much, and I, the Black Hand, changed the course of history with one simple, albeit strategic, pinch. “Fire in the hole!” I bellowed. [University of Wisconsin (Milwaukee): The Post Online]

I’m as giddy as a little school girl, albeit a school girl with a self inflicted SLAYER tattoo. The real kicker on this one is that they’re playing with the gnarly NOMADS, the triple-guitared QUADRAJETS and Missoula natives gone-west-coast-big-time, FIREBALLS OF FREEDOM. [University of Montana: Montana Kaimin]

Western Australian students still go on bar crawls; they still have clubs and societies; and they still have student newspapers. (albeit full of corporate advertising. Eds) [ University of Newcastle: Opus]

If a bum shoulder wasn’t going her keep her out though, a new position wouldn’t either. The new position, albeit less stressful than catcher, was nonetheless quite taxing on Slone’s injury. [University of Chicago: Chicago Maroon]

Thus, of the two, lest may well function as a sign of parody or playfulness, unless in a direct quote, whereas albeit is clearly still part of the normal range of the English in the Coll Corpus.26

One example which deserves closer consideration is the uses to which thou/thee/thy/thine are now put. Since this pronoun is increasingly no longer a natural part of our life, except in prayers to God, its use is becoming con-

---

26 The ratio is 40/1532 for albeit/although. For the Brown/Frown/LOB/FLOB (at 4 million words), the ratio is 32/1540.
fined to quotes, parody or otherwise ‘exuberant’ language; the high value for ‘pseudo-archaic’ (as in [20]) is a direct indication of this latter pattern.

<table>
<thead>
<tr>
<th>Abbrev</th>
<th>Type</th>
<th>per M:</th>
<th>Coll</th>
<th>BNC</th>
<th>94 Ind</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q</td>
<td>Quote</td>
<td>1.0</td>
<td></td>
<td></td>
<td>3.85</td>
</tr>
<tr>
<td>PQ</td>
<td>Pseudo-quote</td>
<td>1.25</td>
<td></td>
<td></td>
<td>.35</td>
</tr>
<tr>
<td>PI</td>
<td>Pseudo-idiom</td>
<td>0.00</td>
<td></td>
<td></td>
<td>.58</td>
</tr>
<tr>
<td>I</td>
<td>Idiom</td>
<td>.75</td>
<td></td>
<td></td>
<td>.48</td>
</tr>
<tr>
<td>A</td>
<td>Archaic</td>
<td>.75</td>
<td></td>
<td></td>
<td>.68</td>
</tr>
<tr>
<td>PA</td>
<td>Pseudo-archaic</td>
<td>1.25</td>
<td></td>
<td></td>
<td>.18</td>
</tr>
<tr>
<td>D</td>
<td>Dialect</td>
<td>0.00</td>
<td></td>
<td></td>
<td>.48</td>
</tr>
<tr>
<td>S</td>
<td>Song</td>
<td>0.00</td>
<td></td>
<td></td>
<td>.40</td>
</tr>
<tr>
<td>M</td>
<td>Meta-comment</td>
<td>0.00</td>
<td></td>
<td></td>
<td>.10</td>
</tr>
<tr>
<td><strong>Totals:</strong></td>
<td></td>
<td>5.00</td>
<td>20.09</td>
<td>7.08</td>
<td></td>
</tr>
</tbody>
</table>

Table 8. Uses of *thou* and related forms, per 1 million words.

[18] Compiling the “Best of the Blotter” for State Of The Gate, I saw that a map of Africa had been stolen from Alumni, as well as some combination door handles. Come on, people. I seem to remember a neat quote from my Western Traditions reading (which I actually did) - something about “Thou shalt not steal.” Somebody important said that -hmm -could it have been God?!?! [Colgate University *The Colgate-Maroon*]

[19] Are we forgetting that The United States of America kicks some serious ass? We are already on the most bitchin’ continent, so why not explore it and discover just why this country tis of thee is such an oh so sweet land of liberty. [California State Univ, Chico, *The Orion Online*]

[20] Tune in next year to find out: Will I ever get a piece of ass? Will someone eventually kick my ass? Is it wrong to covet thy neighbors ass? And, of course, will I ever get my sammiches, ass? Ya know, if you say “ass” enough times, it really starts to sound stupid. [University of North Dakota, *The Digital Student*]

Among other indications of formality, the continuing decrease in use of *upon* is also clearly seen in this corpus. The formal structure, *granted [that]*, does, however, seem to be holding its own. Another point of comparison is provided by the subject-operator inversion with an introductory adverbial such as *hardly* or *no sooner*. It seems clear that the *Coll* use of this type of construction is within the normal range of the comparison corpora.
<table>
<thead>
<tr>
<th>per 4 M wds</th>
<th>Coll</th>
<th>Ind 93(^{27})</th>
<th>BROWN</th>
<th>FROWN</th>
<th>LOB</th>
<th>FLOB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardly</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Never</td>
<td>29</td>
<td>21</td>
<td>44</td>
<td>8</td>
<td>16</td>
<td>36</td>
</tr>
<tr>
<td>Scarcely</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Seldom</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No longer</td>
<td>10</td>
<td>1</td>
<td>8</td>
<td>12</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>No sooner</td>
<td>2</td>
<td>5</td>
<td>20</td>
<td>4</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>Not until</td>
<td>3</td>
<td>0</td>
<td>16</td>
<td>12</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td><strong>Totals:</strong></td>
<td>46</td>
<td>32</td>
<td>96</td>
<td>44</td>
<td>40</td>
<td>60</td>
</tr>
</tbody>
</table>

Table 9. Introductory negative adverbials with subject-operator inversion, per 4 M.

<table>
<thead>
<tr>
<th></th>
<th><em>Brown</em></th>
<th><em>Frown</em></th>
<th><em>LOB</em></th>
<th><em>FLOB</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Conv. factor</strong></td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td><strong>Informality:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>its (correct)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>its (error)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>error %</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>it’s (correct)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>it’s (error)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>error %</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>ain’t</em></td>
<td>44</td>
<td>57</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td><em>bloody hell!</em></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><em>damn!</em></td>
<td>28</td>
<td>40</td>
<td>21</td>
<td>24</td>
</tr>
<tr>
<td><em>fuck!</em></td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td><em>shit!</em></td>
<td>0</td>
<td>17</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 10. Comparisons of *Coll* and the established corpus: Vocabulary III.

\(^{27}\) January 2 – February 6, which is roughly 10% of the nearly 40 M words in that year’s CD.
The most striking instances of how any increase in informality can nevertheless maintain barriers against specific socially stigmatized forms are to be seen in the cases of *its/it’s* and *ain’t*. A comparison of over 6,000 instances of *its/it’s* resulted in an error rate of only 2% for the American writers; the British students are clearly less concerned about this error, as their contribution produced an 11% error rate for *its* (instead of the correct *it’s*), although insertion of a gratuitous apostrophe was at a much lower rate, 4%. This particular pair of mistakes is of special interest, as it will not be picked up by computer spelling checkers. Finally, students are considerably more reluctant to use *ain’t* than a variety of swear words, with only *fuck* still clearly being on at least some North American taboo lists.

6. The future. The examples examined above raise a number of problems, not least the general issue of (in)formality, a matter that needs considerable attention, as people become increasingly empowered to reach others via electronic media such as the Internet.

This preliminary version of the student online newspaper corpus has a number of flaws, the most serious being that copyright permission has not yet been obtained, which precludes release in its present form. The next step is therefore to see to it that the corpus becomes official, publicly available, improved and augmented, in preparation for further synchronic and diachronic studies over the next few years, particularly as more university and college students from the other major varieties of English begin to go electronic on the scale now seen in the United States. Our young people are providing us with a wealth of living data, even as they reshape the language we are handing over to them—it is merely up to us to learn to listen to them.

References:

Hundt, Marianne (1998). ‘It is important that this study (should) be based on the analysis of parallel corpora: On the use of the mandative subjunctive in four major varieties of English’, in Lindquist et al., 159–175.
LOB Corpus. ICAME Collection of English Language Corpora, 2nd ed. CD. Bergen: HIT Centre.
2.2 Study 2: The College Idiom: Idioms in the COLL Corpus

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 2002a) 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 Idi-
oms (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 2002a), 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, 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% 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.
Table 1: Geographical distribution of online college student newspapers in COLL

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

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

<table>
<thead>
<tr>
<th>Lowest</th>
<th>1st quartile</th>
<th>Median</th>
<th>Average</th>
<th>3rd quartile</th>
<th>Highest</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>6,749</td>
<td>10,860</td>
<td>12,982</td>
<td>16,185</td>
<td>53,370</td>
</tr>
</tbody>
</table>

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).

4 The figures in Table 1 are slightly different than those cited in Minugh 2002a: 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.
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. 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

---

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: 273–78, 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.
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 inch” (Craccum, New Zealand), rather than the canonical go the extra mile.

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):

(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,

---

7 For an example of detailed examination of such permutations, see Gustawsson 2006.
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

<table>
<thead>
<tr>
<th>Tot tokens of headwords&lt;sup&gt;8&lt;/sup&gt;</th>
<th>Id types in <em>CCDI</em></th>
<th>Id types in COLL</th>
<th>Id tokens in COLL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Totals</td>
<td>407,639</td>
<td>3,485</td>
<td>1,422</td>
</tr>
<tr>
<td>Standard idioms</td>
<td></td>
<td></td>
<td>4,701</td>
</tr>
<tr>
<td>Variant idioms</td>
<td></td>
<td></td>
<td>250</td>
</tr>
<tr>
<td>Other idioms</td>
<td></td>
<td></td>
<td>488</td>
</tr>
</tbody>
</table>

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*, 1<sup>st</sup> ed. (1995: xvii).<sup>9</sup>

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

<sup>8</sup> 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.

<sup>9</sup> 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).
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. Second, 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)

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

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

<table>
<thead>
<tr>
<th>Norm</th>
<th>Actual</th>
<th>Newspaper</th>
<th>City</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.7</td>
<td>5</td>
<td><em>The Linfield Review</em></td>
<td>Linfield C, Oreg</td>
<td>10,707</td>
</tr>
<tr>
<td>4.7</td>
<td>3</td>
<td><em>The DePauw Online</em></td>
<td>DePauw U, Ind</td>
<td>6,450</td>
</tr>
<tr>
<td>4.6</td>
<td>4</td>
<td><em>UCF Knight Wire</em></td>
<td>U of Cent Florida</td>
<td>8,727</td>
</tr>
<tr>
<td>4.2</td>
<td>3</td>
<td><em>E Carol Tech Onlin</em></td>
<td>North Carolina</td>
<td>7,091</td>
</tr>
<tr>
<td>4.2</td>
<td>1</td>
<td><em>The Voice</em></td>
<td>Langara C, Br Col</td>
<td>2,405</td>
</tr>
<tr>
<td>4.0</td>
<td>8</td>
<td><em>Tulane Hullabaloo</em></td>
<td>Tulane, La</td>
<td>20,027</td>
</tr>
<tr>
<td>3.8</td>
<td>1</td>
<td><em>The Independent</em></td>
<td>Clark Coll, Wash</td>
<td>2,616</td>
</tr>
</tbody>
</table>
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. 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

<table>
<thead>
<tr>
<th>Idioms labels</th>
<th>Neutra</th>
<th>BrI</th>
<th>mainly BrI</th>
<th>NAm</th>
<th>Austr</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idioms found</td>
<td>565</td>
<td>109</td>
<td>32</td>
<td>36</td>
<td>3</td>
<td>745</td>
</tr>
<tr>
<td>Label correct</td>
<td>21</td>
<td>10</td>
<td>49</td>
<td>47</td>
<td>1</td>
<td>745</td>
</tr>
<tr>
<td>Label ‘incorrect’</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td></td>
<td>745</td>
</tr>
</tbody>
</table>

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

---

10 CCDI has a vanishingly small proportion of idioms labeled [Scottish] or [Irish]. For example, *set the heather on fire* is labeled [mainly Scottish].
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)

<table>
<thead>
<tr>
<th>Idiom</th>
<th>COLL (per m wds)</th>
<th>OHPC (Hector)</th>
<th>BNC (100 m)</th>
<th>Bank of English (211 m)</th>
<th>Oxf Engl Corp (1 b)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>spill the beans</em></td>
<td>1.08</td>
<td>0.39</td>
<td>0.39</td>
<td>0.63</td>
<td>0.54</td>
</tr>
<tr>
<td><em>beg the question</em></td>
<td>2.17</td>
<td>2.5</td>
<td>1.16</td>
<td>1.40</td>
<td>1.81</td>
</tr>
<tr>
<td><em>call the shots</em></td>
<td>0.54</td>
<td>0.94</td>
<td>0.48</td>
<td>1.22</td>
<td>0.89</td>
</tr>
</tbody>
</table>

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 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).
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

<table>
<thead>
<tr>
<th>Word</th>
<th>Tokens</th>
<th>Senses</th>
<th>Word</th>
<th>Tokens</th>
<th>Senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>eye</td>
<td>103</td>
<td>18</td>
<td>home</td>
<td>43</td>
<td>9</td>
</tr>
<tr>
<td>line</td>
<td>102</td>
<td>16</td>
<td>name</td>
<td>43</td>
<td>6</td>
</tr>
<tr>
<td>hand</td>
<td>101</td>
<td>21</td>
<td>heart</td>
<td>40</td>
<td>8</td>
</tr>
<tr>
<td>mind</td>
<td>98</td>
<td>9</td>
<td>profile</td>
<td>39</td>
<td>2</td>
</tr>
<tr>
<td>door</td>
<td>87</td>
<td>11</td>
<td>there</td>
<td>39</td>
<td>1</td>
</tr>
<tr>
<td>face</td>
<td>78</td>
<td>14</td>
<td>day</td>
<td>37</td>
<td>10</td>
</tr>
<tr>
<td>kick</td>
<td>71</td>
<td>7</td>
<td>corner</td>
<td>36</td>
<td>5</td>
</tr>
<tr>
<td>edge</td>
<td>59</td>
<td>7</td>
<td>time</td>
<td>35</td>
<td>3</td>
</tr>
<tr>
<td>track</td>
<td>56</td>
<td>7</td>
<td>ball</td>
<td>33</td>
<td>8</td>
</tr>
<tr>
<td>work</td>
<td>54</td>
<td>4</td>
<td>foot</td>
<td>33</td>
<td>20</td>
</tr>
<tr>
<td>finger</td>
<td>53</td>
<td>12</td>
<td>business</td>
<td>32</td>
<td>6</td>
</tr>
<tr>
<td>head</td>
<td>53</td>
<td>20</td>
<td>hell</td>
<td>32</td>
<td>12</td>
</tr>
<tr>
<td>ground</td>
<td>52</td>
<td>11</td>
<td>air</td>
<td>29</td>
<td>5</td>
</tr>
<tr>
<td>up</td>
<td>48</td>
<td>3</td>
<td>blood</td>
<td>28</td>
<td>12</td>
</tr>
<tr>
<td>mark</td>
<td>47</td>
<td>8</td>
<td>dead</td>
<td>28</td>
<td>5</td>
</tr>
<tr>
<td>wake</td>
<td>46</td>
<td>2</td>
<td>sight</td>
<td>28</td>
<td>4</td>
</tr>
<tr>
<td>boy</td>
<td>44</td>
<td>8</td>
<td>way</td>
<td>28</td>
<td>4</td>
</tr>
<tr>
<td>board</td>
<td>43</td>
<td>5</td>
<td>fire</td>
<td>26</td>
<td>6</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Idiom</th>
<th>Count</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>mind</td>
<td>61</td>
<td>keep/bear sth in mind</td>
</tr>
<tr>
<td>door</td>
<td>48</td>
<td>open/close doors</td>
</tr>
<tr>
<td>work</td>
<td>43</td>
<td>in the works</td>
</tr>
<tr>
<td>wake</td>
<td>42</td>
<td>in the wake of sth</td>
</tr>
<tr>
<td>face</td>
<td>39</td>
<td>face to face</td>
</tr>
<tr>
<td>kick</td>
<td>39</td>
<td>kick ass/butt</td>
</tr>
<tr>
<td>there</td>
<td>39</td>
<td>be there for sb</td>
</tr>
<tr>
<td>edge</td>
<td>34</td>
<td>the cutting edge (= advanced)</td>
</tr>
<tr>
<td>profile</td>
<td>33</td>
<td>a high profile, high-profile</td>
</tr>
<tr>
<td>hand</td>
<td>31</td>
<td>in your hands, in the hands of sb</td>
</tr>
<tr>
<td>finger</td>
<td>28</td>
<td>point the finger at, finger-pointing</td>
</tr>
<tr>
<td>shoe</td>
<td>25</td>
<td>step into/fill sb’s shoes</td>
</tr>
<tr>
<td>up</td>
<td>25</td>
<td>up~ and coming</td>
</tr>
<tr>
<td>eye</td>
<td>24</td>
<td>open your eyes</td>
</tr>
<tr>
<td>school</td>
<td>24</td>
<td>old school</td>
</tr>
<tr>
<td>hold</td>
<td>23</td>
<td>put on hold</td>
</tr>
<tr>
<td>line</td>
<td>23</td>
<td>the bottom line</td>
</tr>
<tr>
<td>corner</td>
<td>22</td>
<td>just around the corner</td>
</tr>
<tr>
<td>line</td>
<td>22</td>
<td>a fine line</td>
</tr>
<tr>
<td>scene</td>
<td>22</td>
<td>behind the scenes</td>
</tr>
<tr>
<td>time</td>
<td>22</td>
<td>(hit) the big time</td>
</tr>
<tr>
<td>name</td>
<td>21</td>
<td>call sb names, name-calling</td>
</tr>
<tr>
<td>up</td>
<td>21</td>
<td>up and running</td>
</tr>
<tr>
<td>belt</td>
<td>20</td>
<td>under your belt</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>Papers 12</th>
<th>Size</th>
<th>Category</th>
<th>Idioms</th>
<th>I/10M</th>
</tr>
</thead>
<tbody>
<tr>
<td>(of 50)</td>
<td>(1,000 wds)</td>
<td>all items</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>17.6</td>
<td>Editorial</td>
<td>50</td>
<td>8.14</td>
</tr>
<tr>
<td>21</td>
<td>43.3</td>
<td>Profile</td>
<td>95</td>
<td>6.29</td>
</tr>
<tr>
<td>37</td>
<td>117.8</td>
<td>Arts</td>
<td>241</td>
<td>5.87</td>
</tr>
<tr>
<td>35</td>
<td>87.3</td>
<td>Sports</td>
<td>138</td>
<td>4.54</td>
</tr>
<tr>
<td>49</td>
<td>288.8</td>
<td>Feature</td>
<td>435</td>
<td>4.32</td>
</tr>
<tr>
<td>19</td>
<td>21.4</td>
<td>Letters</td>
<td>28</td>
<td>3.76</td>
</tr>
<tr>
<td>29</td>
<td>43.1</td>
<td>News</td>
<td>45</td>
<td>3.00</td>
</tr>
<tr>
<td>9</td>
<td>9.0</td>
<td>Ads</td>
<td>4</td>
<td>1.28</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Papers</th>
<th>Size</th>
<th>Category</th>
<th>Idioms</th>
<th>I/10M</th>
</tr>
</thead>
<tbody>
<tr>
<td>(of 50)</td>
<td>(1,000 wds)</td>
<td>no ‘other’</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>17.6</td>
<td>Editorial</td>
<td>45</td>
<td>7.33</td>
</tr>
<tr>
<td>21</td>
<td>43.3</td>
<td>Profile</td>
<td>84</td>
<td>5.56</td>
</tr>
<tr>
<td>37</td>
<td>117.8</td>
<td>Arts</td>
<td>227</td>
<td>5.53</td>
</tr>
<tr>
<td>35</td>
<td>87.3</td>
<td>Sports</td>
<td>126</td>
<td>4.14</td>
</tr>
<tr>
<td>49</td>
<td>288.8</td>
<td>Feature</td>
<td>405</td>
<td>4.02</td>
</tr>
<tr>
<td>19</td>
<td>21.4</td>
<td>Letters</td>
<td>23</td>
<td>3.09</td>
</tr>
<tr>
<td>29</td>
<td>43.1</td>
<td>News</td>
<td>44</td>
<td>2.93</td>
</tr>
<tr>
<td>9</td>
<td>9.0</td>
<td>Ads</td>
<td>3</td>
<td>0.96</td>
</tr>
</tbody>
</table>

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.

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.
• 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

<table>
<thead>
<tr>
<th>Size (1000s)</th>
<th>Category</th>
<th>Idioms found</th>
<th>Percent found</th>
<th>Idioms expected</th>
<th>Percent expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>117.8</td>
<td>Arts</td>
<td>18</td>
<td>12.3%</td>
<td>30</td>
<td>20.2%</td>
</tr>
<tr>
<td>288.8</td>
<td>Feature</td>
<td>59</td>
<td>40.4%</td>
<td>72</td>
<td>49.6%</td>
</tr>
<tr>
<td>43.1</td>
<td>News</td>
<td>16</td>
<td>11.0%</td>
<td>11</td>
<td>7.7%</td>
</tr>
<tr>
<td>43.3</td>
<td>Profile</td>
<td>25</td>
<td>17.1%</td>
<td>11</td>
<td>7.4%</td>
</tr>
<tr>
<td>87.3</td>
<td>Sports</td>
<td>28</td>
<td>19.2%</td>
<td>22</td>
<td>15.0%</td>
</tr>
<tr>
<td>582.2</td>
<td>Total</td>
<td>146</td>
<td>100%</td>
<td>146</td>
<td>100%</td>
</tr>
</tbody>
</table>

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)

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

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

<table>
<thead>
<tr>
<th>Area</th>
<th>Idioms</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Zealand</td>
<td>5</td>
<td>32.5%</td>
</tr>
<tr>
<td>Great Britain</td>
<td>21</td>
<td>21.0%</td>
</tr>
<tr>
<td>US</td>
<td>47</td>
<td>18.8%</td>
</tr>
<tr>
<td>Australia</td>
<td>2</td>
<td>14.1%</td>
</tr>
<tr>
<td>Canada</td>
<td>4</td>
<td>13.6%</td>
</tr>
</tbody>
</table>
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 ‘new’ idioms (i.e., not in CCDI)

<table>
<thead>
<tr>
<th>word</th>
<th>token</th>
<th>idiom</th>
</tr>
</thead>
<tbody>
<tr>
<td>door*</td>
<td>48</td>
<td>open/close ~s</td>
</tr>
<tr>
<td>boy</td>
<td>15</td>
<td>good old ~s</td>
</tr>
<tr>
<td>bang</td>
<td>13</td>
<td>finish with a ~</td>
</tr>
<tr>
<td>bridge*</td>
<td>13</td>
<td>~ the gap</td>
</tr>
<tr>
<td>business*</td>
<td>13</td>
<td>none of sb’s ~</td>
</tr>
<tr>
<td>hand</td>
<td>13</td>
<td>a helping ~</td>
</tr>
<tr>
<td>kick*</td>
<td>13</td>
<td>~ back (= relax)</td>
</tr>
<tr>
<td>fun*</td>
<td>10</td>
<td>~ and games</td>
</tr>
<tr>
<td>boy</td>
<td>9</td>
<td>bad ~</td>
</tr>
<tr>
<td>boy</td>
<td>9</td>
<td>wonder ~</td>
</tr>
<tr>
<td>ground*</td>
<td>9</td>
<td>stand your ~</td>
</tr>
<tr>
<td>board</td>
<td>8</td>
<td>get on the ~ (=first goal)</td>
</tr>
<tr>
<td>board</td>
<td>8</td>
<td>get/come on ~ (= join)</td>
</tr>
<tr>
<td>rein</td>
<td>8</td>
<td>take the ~s</td>
</tr>
<tr>
<td>door*</td>
<td>7</td>
<td>foot in the ~</td>
</tr>
<tr>
<td>melting*</td>
<td>7</td>
<td>the ~ pot (=assimilation)</td>
</tr>
<tr>
<td>back*</td>
<td>5</td>
<td>~ to basics</td>
</tr>
<tr>
<td>bomb</td>
<td>5</td>
<td>drop a ~</td>
</tr>
</tbody>
</table>

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)

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)

<table>
<thead>
<tr>
<th>Type</th>
<th>Nr</th>
<th>%</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Films</td>
<td>79</td>
<td>31%</td>
<td>Wonder Boys, The Full Monty, Drop Dead Gorgeous</td>
</tr>
<tr>
<td>Albums</td>
<td>32</td>
<td>12%</td>
<td>Salad Days, Bury the Hatchet, Heads Are Gonna Roll</td>
</tr>
<tr>
<td>Coinages</td>
<td>30</td>
<td>12%</td>
<td>Out of the Cold (housing), Glass Ceiling (report)</td>
</tr>
<tr>
<td>Bands</td>
<td>26</td>
<td>10%</td>
<td>Ten Foot Pole, Eagle-Eye Cherry, Top Dog</td>
</tr>
<tr>
<td>Songs</td>
<td>17</td>
<td>7%</td>
<td>My Name is Mud, Another Brick in the Wall</td>
</tr>
<tr>
<td>Organization</td>
<td>15</td>
<td>6%</td>
<td>Hotter Than Hell (Harley Days), the Old Boys Union</td>
</tr>
<tr>
<td>Bar/Restaurant</td>
<td>13</td>
<td>5%</td>
<td>The Fat Cat, The Dark Horse Tavern, Mick E Fynn’s</td>
</tr>
<tr>
<td>TV shows</td>
<td>9</td>
<td>3%</td>
<td>Short and Curlies, Get a Life, Seventh Heaven</td>
</tr>
<tr>
<td>Sports</td>
<td>8</td>
<td>3%</td>
<td>Dream Team, The Trail Blazers, Wild Card</td>
</tr>
</tbody>
</table>

For example, there were no examples of ahead of/behind the curve in the BNC.
<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Percentage</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stores</td>
<td>7</td>
<td>3%</td>
<td>The Paper Trail (stationary), The Common Ground</td>
</tr>
<tr>
<td>Plays</td>
<td>5</td>
<td>2%</td>
<td>The Other Side of the Closet</td>
</tr>
<tr>
<td>Books</td>
<td>4</td>
<td>2%</td>
<td>Be My Guest (Hilton), Daniel in the Lion’s Den</td>
</tr>
<tr>
<td>Magazines</td>
<td>4</td>
<td>2%</td>
<td>Raw Deal</td>
</tr>
<tr>
<td>Theaters</td>
<td>4</td>
<td>2%</td>
<td>The Black Box, Off Limits (exhibition)</td>
</tr>
<tr>
<td>Commerce</td>
<td>2</td>
<td>1%</td>
<td>Helping Hands (nanny agency), Spic ’n Span</td>
</tr>
<tr>
<td>Computers</td>
<td>2</td>
<td>1%</td>
<td>Cloak and Dagger (game)</td>
</tr>
<tr>
<td>History</td>
<td>2</td>
<td>1%</td>
<td>the Reign of Terror</td>
</tr>
</tbody>
</table>

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 metalevel 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

<table>
<thead>
<tr>
<th>Type of change</th>
<th>Tokens</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>substitution</td>
<td>116</td>
<td>a juggling act, alarm bells go off, your cup of juice, easy as torte, a blot on our history</td>
</tr>
<tr>
<td>grammar/roles</td>
<td>31</td>
<td>come &gt; bring full circle, draw the curtain &gt; is drawn, in pole position &gt; a pole-sitter</td>
</tr>
<tr>
<td>elaboration or specification</td>
<td>30</td>
<td>much ado about the choreography, over the hump &gt; Hump Day, all dressed up and nowhere to sing, a purchasing wedge</td>
</tr>
<tr>
<td>shift</td>
<td>18</td>
<td>that’s all he wrote, in &gt; find the groove, to jump off the bandwagon</td>
</tr>
<tr>
<td>reversal of usual meaning</td>
<td>15</td>
<td>miss the bandwagon, the buck starts here, all that glitters is gold, time off your hands</td>
</tr>
<tr>
<td>puns</td>
<td>12</td>
<td>crash and learn, the “Cookie” crumbles (person’s nickname), to Kerry the torch</td>
</tr>
<tr>
<td>blend of 2 idioms</td>
<td>9</td>
<td>a whole other ball of wax, the other side of the closet, blaze a trail of glory</td>
</tr>
<tr>
<td>twists</td>
<td>9</td>
<td>wake up and smell the bird seed, at the crack of noon, come hell or high budget overruns</td>
</tr>
<tr>
<td>inverted WO</td>
<td>5</td>
<td>to add injury to insult, silver clouds have a dark lining, more dollar for the bang</td>
</tr>
<tr>
<td>errors</td>
<td>5</td>
<td>score the brunt of their runs, put on the (baeck) burner, hand/tongue and cheek</td>
</tr>
</tbody>
</table>

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

---

\(^{14}\) Both Langlotz and Gustawsson provide systematically defined categories for variation types; the labels in Table 16 are not intended to be as carefully worked out.
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 inter-national 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)

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):

---

\(^{15}\) 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.
(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.

References:

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:


2.3 Study 3: The Filling in the Sandwich: Internal modification of idioms

Abstract

Idiomatic expressions – defined as (relatively) fixed and semantically opaque units such as a one-horse town or buy the farm ‘die’ – are basically self-contained, but can be ‘anchored’ in the discourse at hand via e.g. post-modification: A great many people thought that the pendulum of permissiveness had swung too far. But internal expansion is also possible: These dangers are being swept under the risk-factor rug. Using the BNC and newspaper CDs as corpora of sufficient size (approximately 300 million words in all), the patterns and frequency of such anchoring internal expansions in contemporary English are investigated, and compared with those for alternative formulations and the simplex form. Anchoring internal expansion is found to be generally possible, and occasionally inventive, but usually infrequent (with exceptions such as not have a leg to stand on); anchoring the idiom via exemplification in a following clause is a primary discourse alternative.

1. Introduction

Idiomatic and formulaic phrases have long been part of language. Only a stone’s throw from where I live in Stockholm, a huge block on the edge of a field in the village of Ed still proclaims:¹

[1] rahnualtr lit rista runar efR fastui moþur sina onems totR
Ragnvald let carve runes after Fastvi, mother his, Onäm’s daughter

to i aipi kupro hialbi ant hena
(who) died in Ed, God help soul her!

The words kupro hialbi ant hena (God help her soul!) are a traditional phrase found on many rune-stones, and parallel to the Wife of Bath’s

[2] God have hir soule! (Wife of Bath’s Tale, l. 536)

or our modern equivalent

[3] God save her soul!

¹ Swedish runestone U 112. The runes are continuous, with single spacers between words; the larger spaces have been inserted here to make it easier to follow via the translation.
all of which are both formulaic and a prayer, and thus idiomatic expressions anchored firmly in the discourse of religion. Note also that even if they appear fixed, they often have alternate, more expanded forms, such as

[3a] God have mercy on her benighted soul

New fixed expressions (in the sense used in Moon 1998) can of course also appear at any time. One such came into English about two decades ago, when a hamburger chain called Wendy’s began running a series of advertisements centered around the question

[4] Where’s the beef?

implying that their competitors failed to have much more content than the seeded rolls you see on the outside. The theme was picked up by Walter Mondale in the 1984 elections, who in turn implied that his political opponent, Ronald Reagan, packaged and sold a similar product, all exterior and no content. Whether the beef idiom will be ephemeral or last remains, of course, to be seen.

In this paper, however we shall restrict our choice to a series of relatively well-known idioms, all of which are found in the Collins COBUILD Dictionary of Idioms (CDI) (2002) such as:

[5] to kick the bucket
[6] to die with your boots on
[7] to bring coals to Newcastle

and examine the extent to which these prepackaged chunks of language can be internally expanded so as to link them into the discourse within which they are used. Thus, idiom (7) actually turns up in the British National Corpus (BNC) in the following form:

[7a] In their protest they sought to restore some political coals to Newcastle: the Declaration of Arbroath in 1320 committed the lay and ecclesiastical nobles of Scotland to support Robert Bruce, who stood against proud Edward’s army, and sent him homeward to think again. (BNC: G1Y13)

Here, it is intuitively obvious both that coals are not normally political, and that the sentence nevertheless makes perfectly good sense.

---

2 A fuller version: “The movie rights [to A Confederacy of Dunces] have been sold more than Uncle Tom, but now they’re in the hands of Drew Barrymore. God have mercy on her benighted soul – she’s going to appear herself as an expanded version of the stripper, Darlene, but today she committed further blasphemy and announced in a French interview that Ignatius will be played by Philip Seymour Hoffman.” (http://www.worldwiderant.com/archives/001138.html, accessed on April 10, 2003)
2 Analytic framework

A first starting point may be found in the semantic analysis provided by Thomas Ernst in an article from 1980 called ‘Grist for the linguistic mill: idioms and “extra” adjectives’. There, he distinguished among three types of adjectives that can modify idioms (Ernst 1980: 52–53):

- **external modification**, such as *Carter doesn’t have an economic leg to stand on*. Here, the adjective refers to the surrounding (external) text.
- **internal modification**, such as *When will you get it through your small head that this isn’t the way to do it?* The adjective modifies an element of the idiom, but does not change its reference. (Cf. example (3a), above.)
- **conjunction modification**, such as *Malvolio deserves almost everything he gets, but... there is that little stab of shame we feel at the end for having had such fun pulling his cross-gartered leg for so long*. Here, the underlined adjective reactivates the literal sense of the idiom, conjoining it to the figurative interpretation.

Ernst’s analysis is concerned with the interpretation of the semantics of the idiomatic phrase, and he thus merely states in passing that the ‘external’ adjectives affect the entire idiom. The idiom is, however, more than merely modified: it is in fact hooked into or anchored in the discourse. It is this process that will be examined in the present article. Along the way, we shall see how common such anchorings are, and whether it merely is adjectives that can provide the hooks.

Moon notes that these anchoring adjectives “provide ‘external modification’, where the inserted adjective contextualizes the whole FEI [*Fixed Expression or Idiom*] in much the same way that a disjunct or sentence adverbial might” (Moon 1998: 174–75). Her corpus provides her with numerous examples, such as the following:

[8] At our house, as ever, we are a little late **getting our Christmas act together**. *(OHPC: journalism)* (Moon 1998: 176)

which might easily be given the alternate form

[8a] At our house, as ever, **as Christmas approaches**, we are a little late **getting our act together**.

This pattern is clearly more restricted than post-modification, which is particularly useful for allowing major expansion of an idiom, as in

[9] (Jourdain) headed south and **bore the brunt** of a ferocious storm which tore his mainsail and forced him to limp to the island of Madeira for a pitstop. *(The Independent 2000)*
At the same time, inserting the ‘hook’ or ‘anchor’ into the middle of the idiom can make it seem less shopworn and give it more punch, as when the British soap character Zak Dingle notes:

[10] Hey, nobody rains on a Dingle parade! (Emmerdale Farm)

as opposed to a potential alternative, post-modification:

[10a] Hey, nobody rains on a parade by the Dingle family!

Moon calls this type of extension ‘exploitations’, “the stylistic manipulation of the lexis (and semantics) of FEIs: perhaps to provide some sort of defamiliarization, and typically providing humour. [...] It is more marked and prevalent with metaphorical FEIs than any other type, since they contain the images which are most easily exploited.” (Moon 1998: 170).

In a relatively early corpus context, Ernst’s discussion of idiom modification was followed up by Nicolas (1995), who focused on adjectives inserted into the NP of V-NP idioms, introducing an ‘adverbial-based modifier classification’ with eight different potential (and five realised) types of such adjectives, basing his argument on introspection, complemented by data from the 1991 Guardian and 1987–89 Wall Street Journal.3 His discussion is primarily a detailed analysis of the various types of modification and the structures of the NPs, together with their implications for the semantics of modification, with relatively little overall corpus data and no frequency data for the individual items.

Another useful distinction was made by Naciscione (2001), who works with ‘phraseological units’ in discourse, distinguishing at the discourse level between the ‘core use’, which is the simplex form used without further elaboration, and ‘instantial use’, which is when the idiom is used in a more elaborate form that “results in significant changes in its form and meaning determined by the context” (Naciscione 2001: 28). She notes that this is a ‘stylistic’ concept, and instantial uses are thus not copied verbatim by other writers. Harkening back to Geoffrey Chaucer once more, she states that “none of the 1164 instantial forms (in his works) became part of (the) phraseological stock of the English language” (Naciscione 2001: 31). Her discussion includes a number of examples of the instantial uses relevant to this paper.

3 The corpus material

Nicolas remarks (1995: 248): “[...] although the research reported here was carried out before I read Ernst’s paper, it can reasonably be seen as an extension of his work [...]” Ironically, the present author had read Ernst’s paper, but not Nicolas’, before doing the research reported here, and the same caveat can thus be said to apply to the present paper.
To obtain a broad sample from various types of English, the following corpora were examined:

- The *BNC*
- The *Los Angeles Times* (CD from 1990)
- *Broadcast News* (CD, radio transcripts from 1995)
- The *New York Times* (CD from 1996)
- The *Independent* (CD from 2002)

This provides material of about 300 million words, none of it overlapping. The various CDs have their own internal search engines, which made some searches difficult (they work best with idioms with highly distinctive components, such as *Hoist by/with one’s own petard*). The technical difficulties should not, however, have any major bearing on the results.

After an initial exploration of a number of idioms from the *CDI* in the *BNC*, 55 such expressions were then selected and searched for (Table 1):

Table 1: The idioms investigated.

<table>
<thead>
<tr>
<th>ace in the hole</th>
<th>new/fresh blood</th>
<th>fly the coop</th>
</tr>
</thead>
<tbody>
<tr>
<td>hold all the aces</td>
<td>blot on the escutcheon</td>
<td>couch potato</td>
</tr>
<tr>
<td>acid test</td>
<td>soften the blow</td>
<td>paper over the cracks</td>
</tr>
<tr>
<td>balancing act</td>
<td>bone of contention</td>
<td>cut and dried</td>
</tr>
<tr>
<td>hidden agenda</td>
<td>die with your boots on</td>
<td>raw deal</td>
</tr>
<tr>
<td>whet your appetite</td>
<td>bear the brunt</td>
<td>just deserts</td>
</tr>
<tr>
<td>upset the applecart</td>
<td>drop in the bucket</td>
<td>a dime a dozen</td>
</tr>
<tr>
<td>out with the ark</td>
<td>bull in a china shop</td>
<td>back to the drawing board</td>
</tr>
<tr>
<td>babe in the wood(s)</td>
<td>bite the bullet</td>
<td>easier said than done</td>
</tr>
<tr>
<td>the whole ball of wax</td>
<td>carry the can</td>
<td>at a low ebb</td>
</tr>
<tr>
<td>baptism by/of fire</td>
<td>put the cart before the horse</td>
<td>catch her eye</td>
</tr>
<tr>
<td>bats in the belfry</td>
<td>cat and mouse</td>
<td>tempt fate/providence</td>
</tr>
<tr>
<td>amount to a hill of beans</td>
<td>skin a cat</td>
<td>at fever pitch</td>
</tr>
<tr>
<td>spill the beans</td>
<td>poisoned chalice</td>
<td>a finger in every pie</td>
</tr>
<tr>
<td>fit the bill</td>
<td>cheek by jowl</td>
<td>grey area</td>
</tr>
<tr>
<td>birds of a feather</td>
<td>chip on his shoulder</td>
<td>a drop in the ocean</td>
</tr>
</tbody>
</table>

---

4 For the purposes of this paper, it is assumed that newspaper CDs can function as corpora, a point that has been repeatedly argued elsewhere.

5 Although earlier newspaper CDs often are in machine-readable form (usually ASCII), more recent years tend to be in compressed files, so that only the search engine provided with the CD can be used to access the data. Additionally, titles and summaries are often in separate files, and only accessible when the relevant articles are displayed by the search engine.
In particular, it should be noted that while the majority of these idioms have a pattern that easily allows the insertion of a modifier, others, such as *full circle*, *low ebb*, *cut and dried* or *couch potato*, are not as easily extended. This broad selection of different types of idiom patterns was deliberate, since it was likely to be of interest to see whether a variety of idioms could be anchored in the text in various fashions. Moreover, some of them can clearly be seen as more or less limited phrases, so that the citation form for e.g. *blood out of a stone/turnip* could also be considered to be *squeeze/get blood out of a stone/turnip*.

In the search, a fair amount of variation was permitted as regards form: if the idiom turned up via a search of the salient elements, it was normally included, even if it was rather different than the canonical form. An example:

> [11]More important is the fact that nobody can really tell whether Germany has a successful economy because of its voting system (though, between you and me, this seems as unlikely as a cart pulling a horse) or whether it seems to have satisfactory constitutional arrangements because its economy has been working so well for so long. (BNC: AKR 888)

Since the corpora were not examined *in toto*, some of the more recherché examples may well have been missed.

4 Categorization of the data

The primary distinction at stake here is whether the idiomatic expression has been modified such that it is anchored in the surrounding text or not. In its canonical form, an idiomatic expression is self-contained and sufficient unto itself, primarily serving as a means of summarizing or commenting upon a situation, rather than providing more concrete information about it. It is this prototypically independent status that makes any overt linking to the text so noticeable.

The first category will thus be the ‘simplex’ or canonical form, with no extension beyond the basic syntactic shifts between positive and negative, active or passive and so forth. Even this category is not completely straightforward, however, since a number of idioms contain e.g. possessives, so that *to catch X’s eye* will normally include a component determined by the discourse:
Although it was the dazzling play in the backs that caught the eye that day, Britt was perhaps the most influential player on the field. (Ind 2002)

Colin Healey, 22, who caught McCarthy’s eye after joining Coventry City on loan from Celtic, […] will be particularly disappointed. (Ind 2002)

It is not, you would think, the ideal way to catch the eye of the great Britain coach David Waite for this autumn’s series against New Zealand. (Ind 2002)

Here, the examples move from the abstract form the eye to a specific s-genitive form, to the (potentially much longer) postpositioned of-genitive. This anchoring in the text, however, is a relatively straightforward reference-tracking device, and no more surprising than Ernst’s type 2 adjective expansion, as in:

“Yet we cannot get visas and we are such a small drop in the ocean compared with the businessman.” (BNC: HJG 1067)

We munched our way through an average 18 pasta meals per head last year, a mere drop in the ocean compared to the Italians, who managed to swallow a massive 300 meals each. (BNC: K37 23)

Here, we can also add time adverbials, intensifiers and quantifiers, whether modifying the noun or the verb. Thus,

But as far as I can tell, people are still chomping at the bit to build here. (LAT 1990)

What is more problematic is to decide how to handle some of the various types of additional information that appear in connection with these idioms. Consider:

Having been at the end of a queue of Germans and Austrians all chomping at the bit to go heliskiing, it was our turn. (BNC: G2W 471)

It’s countdown time to Croke Park and Derry footballers are chomping at the bit in their bid to make sure that “Sam” remains in Ulster for yet another year. (BNC: K32 1994)

US carriers and Lufthansa, among others, are chomping at the bit for new Berlin and East German air routes now that the air corridors, carved out in 1948, are becoming obsolete. (BNC: A94 196)
[21] These dates have now become a **bone of contention in Hong Kong** as rugby administrators in the British territory have belatedly realized… (BNC: CB2 1750)

[22] Under the old two-leg system, Spartak would not have **had a leg to stand on** after their 5–0 drubbing in the first meeting. (Ind 2002)

[23] That achievement **comes full circle** as young people who have never been on stage before perform in front of thousands at the Kennedy Center. (BN 1995)

It seems reasonable to assume that (18)–(20) modify the idiom, and are merely forms of ‘modification’, whereas *in Hong Kong* in (21) is clearly peripheral to the idiom, although not to the discussion, and can easily be moved, e.g. to the front of the sentence:

[21a] **In Hong Kong**, these dates have now become a **bone of contention** as rugby administrators in the British territory have belatedly realized…

Similar reasoning applies for the temporal clause in (22), and arguably even for the *as* clause in (23). The latter, however, is fairly close to the following example, where the second clause exemplifies the first:

[24] My father taught me the **acid test** of a gentleman: respect for those who can be of no possible service to him. (Ind 2002)

[25] Critics of Gorbachev ask why he does not **bite the bullet** and move directly to a **market economy** – as Poland has. (LAT 1990)

These seem to be tightly enough linked to the idiom to be worth keeping track of as a separate group, marked by being independent clauses, usually connected via *and* or a Colon. We may call them ‘exemplifications’, as they amplify the idiom by offering a concrete example of its meaning.

Another group that should be separated out concerns cases where the idiom is in fact the name of a book, play, movie, TV show, song or the like. Names are relatively unproblematic in this context, although worth noting, in that they help keep the idiom alive and in circulation.

The result is a five-category system: simplex, modified (whether via pre-modification or post-modification), names, exemplifications and the subject of our primary interest, ‘anchorings’. (This latter category is what Nicolas (1995: 242) designates as the ‘viewpoint’ category.) These categories are here treated as mutually exclusive, although anchorings clearly are a sub-group of modification, distinguished only by their discourse semantics.

5 Results
The basic results are seen in Table 2.

Table 2: Distributive percentages for the idioms investigated.

<table>
<thead>
<tr>
<th>Canonical form</th>
<th>Tokens</th>
<th>Simplex forms %</th>
<th>Modifiers %</th>
<th>Anchoring %s</th>
<th>Names %</th>
<th>Exemplifications %</th>
</tr>
</thead>
<tbody>
<tr>
<td>ace</td>
<td>65</td>
<td>33.8</td>
<td>35.4</td>
<td>0.0</td>
<td>18.5</td>
<td>12.3</td>
</tr>
<tr>
<td>aces</td>
<td>9</td>
<td>66.7</td>
<td>33.3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>acid</td>
<td>118</td>
<td>30.5</td>
<td>53.4</td>
<td>2.5</td>
<td>8.5</td>
<td>5.1</td>
</tr>
<tr>
<td>act</td>
<td>360</td>
<td>20.0</td>
<td>66.4</td>
<td>6.9</td>
<td>1.7</td>
<td>5.0</td>
</tr>
<tr>
<td>agenda</td>
<td>297</td>
<td>48.1</td>
<td>35.3</td>
<td>3.4</td>
<td>10.8</td>
<td>2.4</td>
</tr>
<tr>
<td>appetite</td>
<td>258</td>
<td>22.1</td>
<td>75.2</td>
<td>2.3</td>
<td>0.0</td>
<td>0.4</td>
</tr>
<tr>
<td>applecart</td>
<td>55</td>
<td>61.8</td>
<td>20.0</td>
<td>7.3</td>
<td>5.5</td>
<td>5.5</td>
</tr>
<tr>
<td>ark</td>
<td>5</td>
<td>80.0</td>
<td>20.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>babe</td>
<td>18</td>
<td>38.9</td>
<td>33.3</td>
<td>16.7</td>
<td>11.1</td>
<td>0.0</td>
</tr>
<tr>
<td>ball</td>
<td>13</td>
<td>84.6</td>
<td>7.7</td>
<td>7.7</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>baptism</td>
<td>87</td>
<td>41.4</td>
<td>44.8</td>
<td>11.5</td>
<td>0.0</td>
<td>2.3</td>
</tr>
<tr>
<td>bats</td>
<td>9</td>
<td>88.9</td>
<td>11.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>beans</td>
<td>37</td>
<td>62.2</td>
<td>37.8</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>beans</td>
<td>100</td>
<td>47.0</td>
<td>47.0</td>
<td>3.0</td>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>bill</td>
<td>290</td>
<td>58.6</td>
<td>41.0</td>
<td>0.3</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>birds</td>
<td>53</td>
<td>39.6</td>
<td>5.7</td>
<td>7.5</td>
<td>45.3</td>
<td>1.9</td>
</tr>
<tr>
<td>bit</td>
<td>72</td>
<td>33.3</td>
<td>63.9</td>
<td>0.0</td>
<td>0.0</td>
<td>2.8</td>
</tr>
<tr>
<td>blind</td>
<td>13</td>
<td>61.5</td>
<td>30.8</td>
<td>7.7</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>blood</td>
<td>24</td>
<td>87.5</td>
<td>8.3</td>
<td>4.2</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>blood</td>
<td>266</td>
<td>51.9</td>
<td>41.0</td>
<td>3.4</td>
<td>1.9</td>
<td>1.9</td>
</tr>
<tr>
<td>blot</td>
<td>5</td>
<td>20.0</td>
<td>20.0</td>
<td>40.0</td>
<td>0.0</td>
<td>20.0</td>
</tr>
<tr>
<td>blow</td>
<td>209</td>
<td>33.0</td>
<td>59.3</td>
<td>7.7</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>bone</td>
<td>133</td>
<td>25.6</td>
<td>69.9</td>
<td>3.0</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>boots</td>
<td>20</td>
<td>65.0</td>
<td>5.0</td>
<td>0.0</td>
<td>25.0</td>
<td>5.0</td>
</tr>
<tr>
<td>brunt</td>
<td>706</td>
<td>3.0</td>
<td>96.3</td>
<td>0.7</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>bucket</td>
<td>132</td>
<td>48.5</td>
<td>45.5</td>
<td>4.5</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>bull</td>
<td>26</td>
<td>80.8</td>
<td>11.5</td>
<td>7.7</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>bullet</td>
<td>171</td>
<td>42.7</td>
<td>14.0</td>
<td>3.5</td>
<td>3.5</td>
<td>36.3</td>
</tr>
<tr>
<td>can</td>
<td>63</td>
<td>50.8</td>
<td>41.3</td>
<td>3.2</td>
<td>1.6</td>
<td>3.2</td>
</tr>
</tbody>
</table>

To conserve space, only percentages are given, but the actual numbers are of course recoverable from the percentages and the number of tokens.
Perhaps the most striking overall result is that the filling in the sandwich is so meagre, weighing in at only 2.7% anchorings, exactly the same level as for names. In consequence, removing the names will have relatively little effect on the anchorings, normally <0.2% for individual items, and 0.1% for the overall figures. As the relative figures for the other categories can nevertheless show considerably larger changes, Table 3 is also provided, giving the figures for the idioms, but with the names removed:
Table 3: Distributive percentages for the idioms investigated (names removed).

<table>
<thead>
<tr>
<th>Reference word</th>
<th>Canonical form</th>
<th>Tokens</th>
<th>Simplex forms %</th>
<th>Modifiers %</th>
<th>Anchorings %</th>
<th>Exemplifications %</th>
</tr>
</thead>
<tbody>
<tr>
<td>ace</td>
<td>ace in the hole</td>
<td>53</td>
<td>41.5</td>
<td>43.4</td>
<td>0.0</td>
<td>15.1</td>
</tr>
<tr>
<td>aces</td>
<td>hold all the aces</td>
<td>9</td>
<td>66.7</td>
<td>33.3</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>acid</td>
<td>acid test</td>
<td>108</td>
<td>33.3</td>
<td>58.3</td>
<td>2.8</td>
<td>5.6</td>
</tr>
<tr>
<td>act</td>
<td>balancing act</td>
<td>354</td>
<td>20.3</td>
<td>67.5</td>
<td>7.1</td>
<td>5.1</td>
</tr>
<tr>
<td>agenda</td>
<td>hidden agenda</td>
<td>265</td>
<td>54.0</td>
<td>39.6</td>
<td>3.8</td>
<td>2.6</td>
</tr>
<tr>
<td>appetite</td>
<td>whet your appetite</td>
<td>258</td>
<td>22.1</td>
<td>75.2</td>
<td>2.3</td>
<td>0.4</td>
</tr>
<tr>
<td>applecart</td>
<td>upset the applecart</td>
<td>52</td>
<td>65.4</td>
<td>21.2</td>
<td>7.7</td>
<td>5.8</td>
</tr>
<tr>
<td>ark</td>
<td>out with the ark</td>
<td>5</td>
<td>80.0</td>
<td>20.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>babe</td>
<td>babe in the wood(s)</td>
<td>16</td>
<td>43.8</td>
<td>37.5</td>
<td>18.8</td>
<td>0.0</td>
</tr>
<tr>
<td>ball</td>
<td>whole ball of wax</td>
<td>13</td>
<td>84.6</td>
<td>7.7</td>
<td>7.7</td>
<td>0.0</td>
</tr>
<tr>
<td>baptism</td>
<td>baptism by/of fire</td>
<td>87</td>
<td>41.4</td>
<td>44.8</td>
<td>11.5</td>
<td>2.3</td>
</tr>
<tr>
<td>bats</td>
<td>bats in the belfry</td>
<td>9</td>
<td>88.9</td>
<td>11.1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>beans</td>
<td>amount to a hill of beans</td>
<td>37</td>
<td>62.2</td>
<td>37.8</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>beans</td>
<td>spill the beans</td>
<td>98</td>
<td>48.0</td>
<td>48.0</td>
<td>3.1</td>
<td>1.0</td>
</tr>
<tr>
<td>bill</td>
<td>fit the bill</td>
<td>290</td>
<td>58.6</td>
<td>41.0</td>
<td>0.3</td>
<td>0.0</td>
</tr>
<tr>
<td>birds</td>
<td>birds of a feather</td>
<td>29</td>
<td>72.4</td>
<td>10.3</td>
<td>13.8</td>
<td>3.4</td>
</tr>
<tr>
<td>bit</td>
<td>champ at the bit</td>
<td>72</td>
<td>33.3</td>
<td>63.9</td>
<td>0.0</td>
<td>2.8</td>
</tr>
<tr>
<td>blind</td>
<td>blind as a bat/c</td>
<td>13</td>
<td>61.5</td>
<td>30.8</td>
<td>7.7</td>
<td>0.0</td>
</tr>
<tr>
<td>blood</td>
<td>blood out of a stone/turnip</td>
<td>24</td>
<td>87.5</td>
<td>8.3</td>
<td>4.2</td>
<td>0.0</td>
</tr>
<tr>
<td>blood</td>
<td>new/fresh blood</td>
<td>261</td>
<td>52.9</td>
<td>41.8</td>
<td>3.4</td>
<td>1.9</td>
</tr>
<tr>
<td>blot</td>
<td>blot on the escutcheon</td>
<td>5</td>
<td>20.0</td>
<td>20.0</td>
<td>40.0</td>
<td>20.0</td>
</tr>
<tr>
<td>blow</td>
<td>soften the blow</td>
<td>209</td>
<td>33.0</td>
<td>59.3</td>
<td>7.7</td>
<td>0.0</td>
</tr>
<tr>
<td>bone</td>
<td>bone of contention</td>
<td>132</td>
<td>25.8</td>
<td>70.5</td>
<td>3.0</td>
<td>0.8</td>
</tr>
<tr>
<td>boots</td>
<td>die with your boots on</td>
<td>15</td>
<td>86.7</td>
<td>6.7</td>
<td>0.0</td>
<td>6.7</td>
</tr>
<tr>
<td>brunt</td>
<td>bear the brunt</td>
<td>706</td>
<td>3.0</td>
<td>96.3</td>
<td>0.7</td>
<td>0.0</td>
</tr>
<tr>
<td>bucket</td>
<td>drop in the bucket</td>
<td>131</td>
<td>48.9</td>
<td>45.8</td>
<td>4.6</td>
<td>0.8</td>
</tr>
<tr>
<td>bull</td>
<td>bull in a china shop</td>
<td>26</td>
<td>80.8</td>
<td>11.5</td>
<td>7.7</td>
<td>0.0</td>
</tr>
<tr>
<td>bullet</td>
<td>bite the bullet</td>
<td>165</td>
<td>44.2</td>
<td>14.5</td>
<td>3.6</td>
<td>37.6</td>
</tr>
<tr>
<td>can</td>
<td>carry the can</td>
<td>62</td>
<td>51.6</td>
<td>41.9</td>
<td>3.2</td>
<td>3.2</td>
</tr>
<tr>
<td>cart</td>
<td>put the cart before the horse</td>
<td>85</td>
<td>80.0</td>
<td>10.6</td>
<td>4.7</td>
<td>4.7</td>
</tr>
<tr>
<td>cat</td>
<td>cat and mouse</td>
<td>211</td>
<td>68.2</td>
<td>28.0</td>
<td>3.3</td>
<td>0.5</td>
</tr>
<tr>
<td>cat</td>
<td>skin a cat</td>
<td>22</td>
<td>27.3</td>
<td>59.1</td>
<td>4.5</td>
<td>9.1</td>
</tr>
<tr>
<td>Idiom</td>
<td>Nr</td>
<td>% anchorings</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------------</td>
<td>------</td>
<td>--------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>have a leg to stand on</td>
<td>39</td>
<td>23.1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rain on your parade</td>
<td>34</td>
<td>23.1%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>babe in the wood(s)</td>
<td>18</td>
<td>18.8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>birds of a feather</td>
<td>53</td>
<td>13.8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>paper over the cracks</td>
<td>47</td>
<td>12.8%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>baptism by/of fire</td>
<td>87</td>
<td>11.5%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chip on his shoulder</td>
<td>125</td>
<td>9.6%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Although the overall figures are quite low, a few idioms stand out for both number and percentage of anchorings. Here are the top 10, again with the names removed.  

Table 4: The 10 highest-frequency anchorings.
But even among these items, the most frequent amount to less than one in four instances, whereas a dozen idioms have (other forms of) modification in more than 50% of their occurrences. In addition, these frequencies suggest that some of the idioms are truly self-contained, hardly admitting of modification at all (16 are more than 2/3 simplex, with another 17 over 50% simplex), while others demand modification or exemplification. The highest frequencies are as follows:

Table 5: The 10 highest-frequency simplex idioms.

<table>
<thead>
<tr>
<th>Idiom</th>
<th>Nr</th>
<th>% simplex</th>
</tr>
</thead>
<tbody>
<tr>
<td>a dime a dozen</td>
<td>45</td>
<td>97.8</td>
</tr>
<tr>
<td>easier said than done</td>
<td>214</td>
<td>90.2</td>
</tr>
<tr>
<td>bats in the belfry</td>
<td>9</td>
<td>88.9</td>
</tr>
<tr>
<td>blood out of a stone/tournip</td>
<td>24</td>
<td>87.5</td>
</tr>
<tr>
<td>die with your boots on</td>
<td>15</td>
<td>86.7</td>
</tr>
<tr>
<td>just deserts</td>
<td>44</td>
<td>86.4</td>
</tr>
<tr>
<td>whole ball of wax</td>
<td>13</td>
<td>84.6</td>
</tr>
<tr>
<td>bull in a china shop</td>
<td>26</td>
<td>80.8</td>
</tr>
<tr>
<td>out with the ark</td>
<td>5</td>
<td>80.0</td>
</tr>
<tr>
<td>put the cart before the horse</td>
<td>85</td>
<td>80.0</td>
</tr>
</tbody>
</table>

Table 6: The 10 highest-frequency idioms with modifiers.

<table>
<thead>
<tr>
<th>Idiom</th>
<th>Nr</th>
<th>% modifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>bear the brunt</td>
<td>706</td>
<td>96.3</td>
</tr>
<tr>
<td>whet your appetite</td>
<td>258</td>
<td>75.2</td>
</tr>
<tr>
<td>bone of contention</td>
<td>132</td>
<td>70.5</td>
</tr>
<tr>
<td>balancing act</td>
<td>354</td>
<td>67.5</td>
</tr>
<tr>
<td>champ at the bit</td>
<td>72</td>
<td>63.9</td>
</tr>
<tr>
<td>cheek by jowl</td>
<td>68</td>
<td>61.8</td>
</tr>
<tr>
<td>soften the blow</td>
<td>209</td>
<td>59.3</td>
</tr>
<tr>
<td>skin a cat</td>
<td>22</td>
<td>59.1</td>
</tr>
<tr>
<td>at a low ebb</td>
<td>195</td>
<td>59.0</td>
</tr>
<tr>
<td>acid test</td>
<td>108</td>
<td>58.3</td>
</tr>
</tbody>
</table>
Table 7: The 10 highest-frequency idioms with exemplification.

<table>
<thead>
<tr>
<th>Idiom</th>
<th>Nr</th>
<th>% exemplification</th>
</tr>
</thead>
<tbody>
<tr>
<td>bite the bullet</td>
<td>165</td>
<td>37.6%</td>
</tr>
<tr>
<td>carry coals to Newcastle</td>
<td>14</td>
<td>21.4%</td>
</tr>
<tr>
<td>blot on the escutcheon</td>
<td>5</td>
<td>20.0%</td>
</tr>
<tr>
<td>ace in the hole</td>
<td>53</td>
<td>15.1%</td>
</tr>
<tr>
<td>back to the drawing board</td>
<td>234</td>
<td>15.0%</td>
</tr>
<tr>
<td>preach to the converted</td>
<td>69</td>
<td>11.6%</td>
</tr>
<tr>
<td>skin a cat</td>
<td>22</td>
<td>9.1%</td>
</tr>
<tr>
<td>full circle</td>
<td>398</td>
<td>8.0%</td>
</tr>
<tr>
<td>die with your boots on</td>
<td>15</td>
<td>6.7%</td>
</tr>
<tr>
<td>upset the applecart</td>
<td>52</td>
<td>5.8%</td>
</tr>
</tbody>
</table>

Not surprisingly, these lists are nearly mutually exclusive, with only 4 of 36 items occurring twice (although in different constellations): 

- bull in a china shop (simplex, anchoring),
- die with your boots on (simplex, exemplification),
- skin a cat (modifier, exemplification),
- soften the blow (modifier, anchoring).\(^8\)

Other factors must also be considered, as well. One structure that actively discourages further modification is when the idiom is inserted into an NP as a premodifier, as in

[26] MILES O’BRIEN, Anchor: You don’t have to run a marathon or pump up like Arnold Schwarzenegger to get a good workout, but you do have to leave your couch potato days behind. (BN 1995)

Here, the difficulties of layered processing would appear to be sufficient explanation for the difficulty of further modification, exemplification or anchoring.

Bite the bullet and carry coals to Newcastle are particularly prone to exemplification, i.e. having an additional clause instantiating what was done or must be done. And in a sense, they point the way forward to what is probably a more fruitful way of analyzing this data, i.e. going beyond the sentence level, and looking at the relation between the idiom and its total embedding in the text. As an example, consider (27), where the text splits up what is clearly a coherent statement by Mr. Gort:

\(^8\) In addition, a few, such as out with the ark and blot on the escutcheon (with only 5 tokens each) are so infrequent that they must be treated with caution.
Many of the idioms examined in this study are simply dropped into the text, like a stone into a well (such as a dime a dozen), but that is clearly not the entire story. As Naciscione notes, “Instantial use may be realized in one sentence or it may exceed sentence boundaries. It may also be carried over to the subsequent paragraphs, chapters, parts of a book or even cover a whole book” (Naciscione 2001: 30).

6. *Webcorp and the Web*

An attempt was also made to test various of these idioms on the Internet, using the *Webcorp* search engine (although this data was not included in the main body of the paper). Since there is no way of knowing how many Net hits one might maximally get, the data is primarily useful for confirming trends seen in the regular corpora. Here are some examples of collocations found via *Webcorp*:

- **Ace in the** \{Albany, battlefield, collective, digital, economic, empty, energy, hair, headline, IT, Labour, money, plot, snook, video\} **hole.**
- **Babe in the** \{acting, advertising, Astral, baseball, cruciverbalist, Diplomacy, financing, historiography, Linux, NFL, ping, political, senatorial\} **wood(s).**
- **Baptism of/by** \{budgetary, bureaucratic, cannon, chemical, fiscal, judicial, legal, lyrical, media, political, proxy, tactical\} **fire.**
- **Drop in the** \{budget, casino, celestial, conceptual, financial, fiscal, funding, income, linguistic, Microsoft, ratings, sexual, virtual\} **bucket.**

Clearly, although a number are rather startling or obscure, the majority are relatively predictable, as well as tending to be one-word anchorings (as would be predicted by the right-branching nature of NPs in English). None recurred with any noticeable frequency, however. Perhaps the most startling result was that *Webcorp* rarely found the maximum number of hits it is set to report on, although that may have been caused in part by search problems with wild cards (we are after all not looking for the canonical form alone).

7 **Conclusions**

It seems clear that it is possible to split idioms and insert anchoring material, but that this is done rather more sparingly than was proposed (in pre-corpus days) by Ernst, who remarked “This sort of stylistic device is fairly common in newspaper and magazine articles” (Ernst 1980), and by Nicolas, who claims that “contrary to received views, at least 90% of V-NP idioms, in-
cluding many usually regarded as completely frozen, appear to allow some form of (syntactically) internal modification” (Nicolas 1995: 233).

As the present corpus-based figures indicate, ‘allowing’ a construction is not at all the same as ‘encouraging’ or ‘mandating’ it. At less than 3% overall, the anchorings remain a minor option, rather than a major pattern, and one that would also seem to be rather less frequent than Moon suggests, with her wealth of genuine examples. The frequencies found here clearly do not suggest that most of these idioms gleefully and repeatedly lend themselves to the verbal fireworks quoted in her Chapter 6; perhaps the BNC and such are merely duller. (But see the examples in the Appendix.)

Although the present study is basically synchronic (the data ranging from the early BNC citations to the 2002 Independent), a more diachronic study might well reveal shifts in the distribution of least some of these idiomatic expressions, particularly if they experience a surge in popularity such as the initial uses of couch potato (starting in the 1970s). Here, however, we lack a fundamental prerequisite, i.e. large amounts of easily accessible older corpus data. One might only note in passing that the data provided by the Collins COBUILD Dictionary of Idioms indicate that of the 55 idioms examined here, only two are labelled as “old-fashioned [BRITISH]”, and five more as “mainly British” or “mainly American”, while 29 are listed as relatively frequent (starred “key idioms”).

However, before dismissing this pattern of embedding as merely predictable and dull, let us examine a much more intriguing example from the American poet and writer Richard Brautigan’s Trout Fishing in America (1967). Throughout the chapter called ‘Trout Fishing in America Peace’, he uses the title as an anchoring phrase of his own in a startling manner. First consider extracts from the opening of this chapter

In San Francisco around Easter time last year, they had a trout fishing in America peace parade. They had thousands of red stickers printed and they pasted them on their small foreign cars, and on means of national communication like telephone poles. The stickers had WITNESS FOR TROUT FISHING IN AMERICA PEACE printed on them. […] They carried with them Communist trout fishing in America peace propaganda posters. (Brautigan 1967: 98):

Such a text has – to put it mildly – a rather strange pattern, one that is not found anywhere else in the work. To begin with, we may note that trout fishing in America is in itself a literary anchoring device that links this chapter’s discourse with the many other uses of the phrase in the work. Next, remove

---

9 At the end of his study, Nicolas concludes that “Almost all V-NP idioms can have syntactic NP modification. This is very clear, so that it is now apparently unmodifiable idiom NPs such as make way and go halves that stand in need of explanation” (1995: 249).
trout fishing in America from all NPs where it occurs, such as WITNESS FOR TROUT FISHING IN AMERICA PEACE. If you do so, a text that could have been written by the John Birch Society steps forth, attacking the peace movement of the 1960s:

In San Francisco around Easter time last year, they had a […] peace parade. They had thousands of red stickers printed and they pasted them on their small foreign cars, and on means of national communication like telephone poles.
The stickers had WITNESS FOR […] PEACE printed on them. […]
They carried with them Communist […] peace propaganda posters.

If you then remove all the terms that are negatively evaluative of these people’s activities, you find a simple report of what various people were doing in California at that time.

In San Francisco around Easter time last year, people had a […] peace parade. They had thousands of […] stickers printed and they pasted them on their […] cars, and on […] like telephone poles.
The stickers had WITNESS FOR […] PEACE printed on them. […]
They carried with them […] peace posters.

TROUT FISHING IN AMERICA can thus be read as being inserted into the WITNESS FOR PEACE phrase to teach us to unhook ourselves from the propaganda hooks that ensnare the text.

In conclusion, you may well ask, Where’s the beef? As far as our idioms go, it seems that they are best seen as not being the main attraction of the text they are embedded in, any more than a hamburger patty is the be-all and end-all of every sandwich. And a good thing, too, as we otherwise really would have a beef!

References

Corpora


Appendix: examples for further analysis

Since this paper has turned up a large number of interesting examples, it seems appropriate to provide some further data, both of obvious cases of anchoring, and of some borderline ones, all linked by idioms.

[28] An energy economist, Charles M. Studness of Manhasset, said the secrecy surrounding the plans was troubling. ‘Are people trying to do something in the public interest, or is there a hidden political agenda?’ Mr. Studness asked. (NYT 1996)

[29] The head of the Irvine Values Coalition, the group that pushed Measure N, said Plummer’s comments were proof that opponents would push what he called “the hidden homosexual agenda”. (LAT 1990)

[30] In a city where running the schools has always been a delicate political balancing act, questions of bureaucratic authority also always concern political influence. (NYT 1996)


[32] Linger ing is an Italian art, and it can be studied at Cova. Those at the bar manage a particularly Italian balancing act in which, standing, they are able to eat cake, drink espresso, gesture while speaking and keep their shawls and purses perched on their slender shoulders. (NYT 1996)

[33] The pitcher says to the umpire: “You’re blind as a baseball bat, you eggplant, and your mother wears go-go boots”. (LAT 1990)

[34] Bones of Contention: The Archaeopteryx scandals. (Ind 2002)

[35] But Mr. Lebed clearly relishes his reputation as a bull in the Kremlin china shop. […] ‘They think that they have lassoed me and that I have to obey and play by the rules,’ he said. (NYT 1996)

[36] But with growing scientific sophistication, it is possible to learn much more about exactly what happens when carcinogenic chemicals rampage like molecular bulls through the body’s cellular china shop. (NYT 1996)
“rewriting” is probably too strong a phrase – the plan was transformed last week into the favoured infant of the Lord Chancellor who has, in his historic ceremonial capacity, been made to carry the Great Can of State. (Ind 2002)

DANA KENNEDY: I just blanked out for a minute. And Robert DeNiro who is a crack armed robber. And this three-hour story is basically a cat-and-mouse game, a mano-a-mano game, between the two of these guys in which Al Pacino basically tries to track down Robert DeNiro. (BN 1995)

KEITH BYARS has one big chip on his shoulder pads concerning JIMMY JOHNSON. (NYT 1996)

The flip side of Robin Leachifying the rich is that, given the basic human need to despise somebody, we focus on the poor. In a country with a chocolate chip on its shoulder, that sentiment has often gotten translated into racial terms (“Negroes are happier not reading,” old-time Southerners used to proclaim). (LAT 1990)

Such grungeplay is probably the strongest link to Velocity Girl’s new deal with Sub Pop: other than that, they hardly fit the Bud-swilling, neck-thrusting bill, and the world is probably a better place for it. (BNC: CK6 2483)

Instead, the man who passed his first coaching licence at the age of 22 when he was a young player at Celtic, has taken what many would consider to be a managerial poisoned chalice. (Ind 2002)

Ms. Schiffren, like Mr. Dole and others who contend that Uncle Sam is strangling family life, puts the tax-collcting cart before the income-generating horse. Soccer moms and dads feel economic pressure these days not because the Government is taking too much out of their paychecks, but because employers aren’t putting enough in. (NYT 1996)

Title: PUTTING A PICKWICK CART BEFORE THE HORSE […] Thirty years ago in medical school we studied the Pickwickian syndrome, named for the portly Dickens character, whose sufferers had poor sleep habits and daytime drowsiness. (NYT 1996)

Title: ON SKINNING SCHRODINGER’S CAT

THOSE who follow physics the way other people follow baseball quietly cheered at the news last week that scientists had finally skinned Schrodinger’s cat. (NYT 1996)

Since the announcement of the layoffs, government officials and labor leaders have sought to soften the economic blow of the company’s departure. (LAT 1990)

I’m happy to stick to the original. And I’m determined to feed it to the undiscerning teenaged louts who hang around the family manse. Last time I tried, I decided to soften the tannic blow by mixing it with an equal part of orange juice. Reaction: mixed. (Ind 2002)

Wolves’ vociferous streak has a negative side, both Butler and Rae tempting providence and the referee after being cautioned. (Ind 2002)

picture and sound quality. Then Murdoch weighed in. ‘Why not subdivide each channel into four or five channels,’ he asked. It’s called multi-plexing and it upset the digital apple cart. (BN 1995)

MILES O’BRIEN: Telescope dealer Martin Cohen says binoculars are a good first choice because if you’re not impressed by the stars,
binoculars can easily be used for other things. But, if they whet your astronomical appetite, this might be the next logical step. (BN 1995)
2.4 Study 4: Is Time A’Changin’?: A diachronic investigation of the idioms used in Time

Abstract. A newly-available net-based corpus of 105 million words of written American English (Time Magazine, 1923–2006, at http://corpus.byu.edu/time) was investigated for the occurrence and diachronic distribution of various types of ‘pure’ idioms such as be raining cats and dogs. Idioms from the Collins COBUILD Dictionary of Idioms (2002 (1995)) were selected for four types of variation and change. Group 1, the 46 idioms labeled ‘old-fashioned’, proved to be noticeably more common before 1970. Group 2, several constructions of the type as scarce as X, exhibited considerably more variation than in more diversified corpora such as the British National Corpus. Group 3, Biblically-derived idioms, were generally less common after 1960, but with the lowest frequencies in the 1930s. The frequencies for the final group, 32 idioms focusing on deception, were relatively constant from the 1950s on, with an interesting dip in the 1970s. Changes in editorial policies may possibly have influenced these results. While not of sufficient magnitude for detailed studies of individual items over time, the Time corpus clearly is sufficient to provide us with a great deal of data and numerous valuable insights into the use of these idioms.

Keywords: idiom, corpus, language change, variation, American English.

1 Introduction

By now a spry octogenarian, Time first hit the newsstands in March, 1923, and (together with Newsweek, which was launched a decade later) has remained a weekly icon of American households ever since. The recent release of the Time corpus (http://corpus.byu.edu/time) by Mark Davies marks a milestone in diachronic corpus linguistics, as it has now become possible quickly and easily to tap into this linguistic record of middle-class American writing. The present paper explores one dimension of this usage: Time’s idioms, and some of the ways this usage may have changed.

2 Idioms

In the present paper, idioms are defined in the narrower sense of (relatively fixed or frozen) idiomatic expressions (Langlotz 2006, Nunberg et al. 1994;
cf. also Alm-Arvius 2006, Cruse 2006, Gustawsson 2006, Makkai 1972. They include such well-known phrases as:

(1) a one-horse town
(2) to catch forty winks
(3) It ain’t over till the fat lady sings

These expressions form a rather amorphous subset of the more general class of formulaic language (Wray 2002), for which further terms such as fixed expressions and idioms (Moon 1998) abound.

Idioms in this narrower sense (what Moon (1998: 23) calls opaque metaphors or pure idioms) will here be taken to exclude a number of related phenomena, such as ad hoc coinages not in general use, foreign phrases (idée fixe, caveat emptor), single-morpheme items (cran-, as in cranberry), compounds (fire escape, dance hall) and intertextual quotations (O brave new world!, No man is an island, Let them eat cake!). However, the borders leak, the guards are not always vigilant, and so we find that the erstwhile intertextual phrase a sea change is now generally used as simply an idiom meaning ‘a major change’, with no discernible links to The Tempest.

As the more salient characteristics of idioms in this narrower sense, Nunberg et al. (1994: 492–93) present:

- Conventioanality: Idioms are conventionalized: their meaning or use can’t be predicted, or at least entirely predicted, on the basis of a knowledge of the independent conventions that determine the use of their constituents when they appear in isolation from one another.
- Inflexibility: Idioms typically appear only in a limited number of syntactic frames or constructions, unlike freely composed expressions (e.g. *the breeze was shot, *the breeze is hard to shoot, etc.).
- Figuration: Idioms typically involve metaphors (take the bull by the horns, metonymies (lend a hand, count heads), hyperboles (not worth the paper it’s printed on), or other kinds of figuration.
- Proverbiality: Idioms are typically used to describe—and implicitly, to explain—a recurrent situation of particular social interest (becoming restless, talking informally, divulging a secret. or whatever) in virtue of its resemblance or relation to a scenario involving homey, concrete things and relations—climbing walls, chewing fat, spilling beans.
- Informality: Like other proverbial expressions, idioms are typically associated with relatively informal or colloquial registers and with popular speech and oral culture.
- Affect: Idioms are typically used to imply a certain evaluation or affective stance toward the things they denote.¹

¹ This passage is abridged from the original, which also includes further material in footnotes.
Of these “more-or-less orthogonal properties”, the preference for proverbi-
ality and informality are the ones that might be most expected to have a neg-
ative impact on the use of idioms in *Time*, a magazine which is part of a re-
latively formal written genre. As we shall subsequently see, however, *Time*
has long had a house predilection for playful language, and contains numer-
ous feature articles that fall within less formal spheres.

In order to have an operational definition of idioms, *CCDI*, the *Collins*
*COBUILD Dictionary of Idioms* (2002), has been chosen as the reference
point or selectional base.² It includes a total of 1,390 headwords (with singu-
lar and plural forms combined) and 3,485 different idioms, a number of them
with two or more main variants (which are not distinguished when counting
for the present paper). *CCDI* is solidly based on the corpus evidence pro-
vided by the Collins COBUILD *Bank of English* (*BoE*) database, which at the
time of *CCDI*’s first edition (1995) consisted of just over 200 M (i.e. 200
million) words. Its major drawback for the present investigation is that the
American component of the *BoE* was much smaller than a geographically
balanced corpus of English would require, a point that needs to be reme-
bered in a study of *Time* data. This is most clearly seen in its tendency to
overlabel items as BrE (Minugh, in press).

It has repeatedly been shown that although individual idioms often are
experienced as salient in their contexts (particularly as headlines), they are as
a whole strikingly infrequent in corpora. As *CCDI* (1995: v) notes,

… idioms are comparatively infrequent […] Nearly one third of the idioms
in this dictionary occur less often than once per 10 million words of the
corpus. The idioms in the highest frequency band occur in our data at least
once per two million words of English. […] only a few of these occur as
frequently as any of the words we have marked for frequency in *The
COBUILD English Dictionary*.

These figures of less than 1 per 1 M words have been repeatedly confirmed
(Moon 1998, Minugh 1999: 65, Minugh, in press) and have never been seri-
ously questioned. As a consequence, it has until quite recently not been pos-
sible to conduct major empirical studies based on corpus evidence of the
kind envisioned by Sinclair (1991: 102):

So if we need, say, fifty occurrences of a sense of a word in order to de-
scribe it thoroughly, then the corpus has to be large enough to yield fifty in-
stances of the least common sense.

² This selection is not completely unproblematic. A comparison with the Oxford equivalent,
*The Oxford Dictionary of English Idioms* (1993), indicated a surprisingly large number of
items found in only the one dictionary or the other: almost 30% for both works.
Sinclair’s benchmark, amply motivated in his 1991 study, would then imply that for idioms occurring once per 5 M words, and having only one sense (a typical situation), a satisfactory investigative corpus should consist of at least 250 M words. Moreover, an investigation wishing to examine change over time should then have successive subcorpora, each of at least that order of magnitude.

3 The *Time* corpus

By such a yardstick, the *Time* corpus is clearly inadequate for a definitive investigation. Nevertheless, at roughly 100 M words, it is the first large-scale historical corpus of (an important genre in) American English, and therefore well worth investigating further. With that caveat in mind, let us now consider the make-up of the *Time* corpus.

The *Time* corpus is essentially a search engine that examines the entire range of articles in *Time* from its first publication in 1923 onward. This database contains about 105 M words in over 275,000 articles and is text-only (no advertisements, pictures, or the like). It permits reasonably sophisticated corpus searches for strings that allow us to find not merely the canonical forms of idioms, but also variants, such as (1), here repeated for convenience’ sake, and (1a):

(1) a one-horse town
(1a) a one-mule town

The genre of the *Time* corpus is clearly journalistic, normally a prose that is collectively written and edited, rather than primarily bylined from one author. The range of areas written about is not unlike that of a major newspaper. As regards time constraints, these are more relaxed than in a daily newspaper, although breaking stories clearly have a much tighter deadline than many of their feature articles. The language is almost exclusively American English, although in more recent years a greater number of (trans-
lated) articles or partial contributions have appeared from abroad.\(^5\) As with newspapers, a certain minimal duplication can occur in e.g. weekly lists of the most popular movies or books.

One way in which a corpus can be examined for representativeness is to look at the distribution of some common words, and compare that to the total number of words. This was carried out with the words \{and, he, they, she, man, house, state, money\}, comparing entire decades only, i.e. 1930s to 1990s. There is a quite satisfactory match between the general distribution of text and these words, as is shown in Figure 1 (p. 6). The shape of the curve suggests that we first see a rise in the size of issues as people’s buying power and the magazine’s circulation are on the increase, followed by a shift to a more picture-oriented format, probably fueled by a drop in the costs and complications of color printing. Even if the motivations for such changes lie outside the scope of the present paper, it is clear that the words in these texts have not undergone any startling relative changes in frequency.

![Figure 1. Distribution of words (1930s–1990s): articles per decade v. total number of words per decade.](image)

### 4 Idioms: variation and change

\(^5\) A complicating factor for the magazine is that there now are three additional regional versions (Europe, Asia, and South Pacific) plus *TIME for Kids*; this variation is apparently not reflected in the database, which presents the core American version.
4.1 Variation

Previous discussions of idioms have not indicated any patterns in which the individual idioms exhibit changes as idioms over time, although they naturally may be expected to participate in any change that occurs in the pronunciation and forms of their component words. Given that the idioms themselves often are anchored in given images or schemata, this is only to be expected. Thus, we would not expect social changes to effect a change in the featured part of idiom (1), producing form (1′):

(1) a one-horse town
(1′) a one-car town

or in some Flash Gordon future, even

(1″) a one-rocket town

Note that this is not to claim that such idioms cannot possibly be found, merely that they would be extremely rare, precisely because they violate the pattern of \{an animal to ride on or pull a wagon\}.

A query of the 2B word net-based OEC database produced one-{oxen, mongrel} town, plus 85 instances of one-horse town, but none indicating a shift to a non-animate source of locomotion. Instead, the variation that does exist, and which can be considerable, instead varies the type of object for which there is only one instance in the town. Thus, we find instances of one-{industry, paper, pub, stop, team, typesetter} town, plus 15 instances of a locomotion-oriented variant: one-street town(s). Perhaps the most spectacular variant was the following:

(4) Croydon has just changed from being a one Starbucks town into being a two Starbucks town (OEC weblog 20.03.0009.002)

Looking at the Time data, we find a similar pattern: one-{newspaper (10), industry (4), paper (3), company (2), party (2), crop, doctor, factory, saloon, square, stoplight, street, tart, team} town, with the figures in parentheses indicating the number of instances. The phrase one-crop town is particularly interesting, as it is on the fuzzy border of idiomaticity, being quite close to a typical (non-idiomatic) phrase from the domain of economics, a one-crop economy.

6 Both hyphenated and non-hyphenated variants included.
7 The element town can of course also be varied, but the variation is sharply limited: \{town, hamlet, village, ?city\}. A one-orchestra city could be a potential variant of this idiom, but a one-party country is probably merely referential (Mackenzie & Mel’čuk 1986: 101).
Another point worth noting is that although the *Time* corpus is only a twentieth the size of the *OEC*, it actually contains more variants of this particular idiom, which is in keeping with the impression mentioned earlier, i.e. that *Time* has a relatively strong predilection for playfulness of language.

**4.2 Change**

One type of change may be described as ‘loss’, whereby an idiom becomes regarded as so ‘old-fashioned’ that it ultimately drops out of use. There are thus two dimensions involved here: how the idiom is perceived by speakers, and frequency-based corpus evidence of its non-use/disappearance. The two can to some extent be conflated, since dictionaries and style guides provide (corpus-based?) labels such as ‘old-fashioned’, a normative judgment that may hasten an idiom’s demise—at least, to the extent that such guides have an effect on users of English. This latter point is not irrelevant for a corpus based on a single editing staff and house style, as it is presumably more sensitive to such judgments.

To some extent, idioms, like any other lexical items, can disappear merely as a consequence of falling out of fashion, a phenomenon primarily seen in (overused) slang. But another pressure on lexical items can be the fact that their anchor in everyday life disappears, through changes in the world people experience and know. As the horses in example (1) cease to be regarded as part of most people’s everyday life, this scenario (owning a horse > doing okay) becomes less and less relevant, although this particular scenario may have been saved by a transformation: increasing numbers of people have horses as a form of pet or hobby. Other such areas might include terms from sailing ships or the sphere of agriculture.

As more drastic examples, consider (5)–(7):

(5) a blot on your escutcheon
(6) a mess of pottage
(7) it ain’t over until the fat lady sings

Ever since blotting paper began to disappear with the post-WWII development of the ballpoint pen and disappearance of ink-wells, *blot* has been losing ground, from a high of 9.6 tokens/M of *blot* in the 1940s to 1.9 in the current decade. Of course, escutcheons have never been in great favor in the peerless U.S., and the last *Time* reference to an actual escutcheon was in 1930. Thus, it is hardly strange that (5) is opaque to Americans, who lack the domain of a titled upper class—it may never have been all that relevant for non-heraldic Americans, and certainly is not so today.
A larger-scale form of this same kind of change is when an entire major domain undergoes a sea-change, such as the Western shift from an overtly religious Christian society, with Biblical texts continually present and invoked, to its present state, where such texts and their metaphors are much less frequently invoked. If an idiom such as (6) is no longer associated with the Biblical story of Jacob and Esau (Gen. 25:29–34), it become almost unintelligible, since a mess of has long since become primarily associated with untidiness or even excrement, rather than a dish of lentils, and pottage is also obscure.\(^8\)

A second form of change would be when an individual idiom becomes more popular, spreading beyond its original domain(s) and/or geographical regions. Thus, (7), an American idiom documented since at least the 1970s, first surfaces in the *Time* corpus in 1984:

(8) In the old joke, the opera is not over until the fat lady sings. (*Time*, June 18, 1984)

However, it really took off and spread when George G.W. Bush used the expression on the campaign trail in New Hampshire in 1988. Political cartoons and reporting helped spread an awareness of the idiom to Great Britain, and by 1991 there even was a Dublin band called *The Fat Lady Sings*. The idiom is listed in *CCDI* (1995: 233) without a regional label, and is also to be found in the *BNC*, and is currently rampant on Internet discussions of the 2008 U.S. elections.

Other idioms emerge full-blown and Venus-like from the murky waters of politics, as the Nixonian *smoking gun* (labeled an Americanism by *CCDI* [1995: 174–75]), or Churchill’s 1940 recasting of the older *blood, sweat and tears* as *blood, sweat, toil and tears* (*CCDI* [1995: 35]).\(^9\) The high social status of the first user and the high stakes of its first use guarantee that it becomes widely disseminated, although not necessarily remembered.

5 The *Time* test data

5.1 Old-fashioned idioms

Of its 3,485 entries (some 4,400 if all the variants are counted), the *CCDI*, with its emphasis on “real English,” excludes all obsolete or archaic idioms,

---

\(^8\) For its Biblical use (the exact phrase was not in the *King James Version*), see *OED*, mess n.\(^1\), I, ‘a portion of food’, sense 2a. This is the only subsense not marked ‘obsolete’ or ‘regional’. Neither this use of *mess* nor the word *pottage* are to be found in e.g. the Cambridge *Advanced Learner’s Dictionary*.

\(^9\) For its origins, see e.g. Flavell & Flavell (2006: 43–44).
but retains a total of 46 idioms it labels as ‘old-fashioned’. These ‘old-fashioned’ idioms include:

- not have a bean
- not know how many beans make five
- have a bee in your bonnet
- cut the cackle!
- set your cap at someone
- be the cat’s whiskers
- be raining cats and dogs
- be in the catbird seat
- pull someone’s chestnuts out of the fire
- get your dander up
- cheer someone to the echo
- enough is as good as a feast

The reader is at this point invited to see whether he/she would accept the above items. It is of course entirely possible for a given idiom to be in general use, but not in an individual’s personal repertoire (a point our English Department has repeatedly observed when creating tests for learners, with the test checkers often complaining that they do not recognize a few idioms selected by the test makers). The most famous of these ‘old-fashioned’ idioms is probably be raining cats and dogs, which—as corpus linguists have repeatedly noticed—may hold the distinction of being the most-taught and least-used idiom in English. In the author’s (subjective) judgment, at least, the following are still clearly in circulation:

- be in apple-pie order
- a little bird/birdie told me
- the cat’s meow (not the cat’s whiskers!)
- warm the cockles of your heart
- in two shakes of a lamb’s tail
- snug as a bug in a rug
- catch forty winks

and possibly the following:

- in for a penny
- turn up like a bad penny
- scarce as hen’s teeth
- (not) up to snuff
- sober as a judge
These 46 items were therefore tested against the *Time* corpus. Most idioms, as *CCDI* notes, may be expected to occur between 1 and 10 times per 10 M words, so the ‘old-fashioned’ items may be expected to occur at less than that level in recent decades, but be more frequent previously. The results clearly indicate that a number of these idioms have been in use throughout the 20th century, as Table 1 (below) illustrates.

Table 1. The 10 most frequent ‘old-fashioned’ idioms in *Time* corpus, per 10 M words

<table>
<thead>
<tr>
<th>Idiom</th>
<th>Total Freq</th>
<th>20s–50s</th>
<th>60s–00s</th>
</tr>
</thead>
<tbody>
<tr>
<td>get someone’s dander up</td>
<td>6.4</td>
<td>10.7</td>
<td>2.5</td>
</tr>
<tr>
<td>(not) be up to snuff</td>
<td>3.3</td>
<td>4.7</td>
<td>2.0</td>
</tr>
<tr>
<td>cheer someone to the echo</td>
<td>2.1</td>
<td>3.9</td>
<td>0.4</td>
</tr>
<tr>
<td>be in the catbird seat</td>
<td>2.0</td>
<td>2.2</td>
<td>1.3</td>
</tr>
<tr>
<td>(be in the catbird seat)(^{10})</td>
<td>(0.8)</td>
<td>(0.0)</td>
<td>(2.7)</td>
</tr>
<tr>
<td>a blot on your escutcheon</td>
<td>1.7</td>
<td>2.4</td>
<td>0.7</td>
</tr>
<tr>
<td>warm the cockles of your heart</td>
<td>1.5</td>
<td>2.6</td>
<td>0.4</td>
</tr>
<tr>
<td>have a bee in your bonnet</td>
<td>1.4</td>
<td>1.8</td>
<td>0.4</td>
</tr>
<tr>
<td>cock a snook at someone</td>
<td>1.0</td>
<td>1.8</td>
<td>0.4</td>
</tr>
<tr>
<td>of the first water</td>
<td>1.0</td>
<td>1.4</td>
<td>0.5</td>
</tr>
<tr>
<td>(pushed around) from pillar to post</td>
<td>0.9</td>
<td>1.2</td>
<td>0.5</td>
</tr>
</tbody>
</table>

In addition, at least one such item, *be in the catbird seat*, appears to have come into circulation in the latter part of the century—and here Thurber’s play may have been instrumental in helping it become known. Another six items occurred once in the latter part of the century and not at all in the former part, even if only one occurrence is too slim a reed to base claims upon; only one other item, *turn up like a bad penny*, occurred twice in the latter, but once in the former part of the century.

At the same time, we may note that a total of 11 items occurred only once in the 85 years of the *Time* corpus, and a further 10 did not occur at all. This could be interpreted as indicating that they are more than old-fashioned, and are in fact obsolete or archaic. However, 18 of these 21 were labeled as BrE by *CCDI*, so that they are not to be expected in AmE.\(^{11}\) Those entirely missing are as follows:

---

\(^{10}\) The figures in parentheses are with all overt references to Thurber’s play *The Catbird Seat* removed.

\(^{11}\) It should perhaps be mentioned that the 200+ M then-*Bank of English* had a relatively small AmE contribution, and its component from *Time* would not have been particularly large, so that there is little danger of circular reasoning based on identity between the *BoE* texts and the *Time* corpus.
• know how many pins make five
• pin your ears back (‘listen carefully’)
• enough is as good as a feast
• be up a gum tree
• not my line of country
• (clever as) a cartload of monkeys
• not on your nelly
• sell the pass
• two shakes of a lamb’s tail (‘BrE, old-fashioned AmE’)
• spoil the ship for a ha’porth of tar

Since the absolute numbers of tokens for the individual idioms per decade are relatively small (even the largest, get someone’s dander up, occurs more than 5 times per decade during only the 30s to the 60s), their numbers have been aggregated for the comparison in Figure 2 (cf. also Table 1, above).

Figure 2. The (adjusted) frequencies per 10 M for ‘old-fashioned’ idioms in the Time corpus

What emerges is a fairly strong sense that some sort of change appears to have taken place in the (late?) 60s, and the temptation is almost irresistible to link this linguistic shift to the social changes ushered in by the social turmoil in the U.S. of the late 60s: student unrest, the growing power of the civil rights movement, the resistance to the war in Vietnam, ‘flower power’ and the first stirrings of the feminist movement. If this is correct, we are seeing a re-evaluation of ‘received wisdom’.

5.2 Creative variation
In examining the variations for the idiom *a one-horse town*, we have already claimed that *Time* tends to encourage verbal fireworks, one way of accomplishing this being precisely through variation of (otherwise relatively) fixed idioms. Treating them as constructions with a fairly open slot, we can then examine further instances, such as:

(9) as scarce as X  
(10) as slippery as a Y  
(11) as happy as a Z

The canonical form for (9) is \( X = h e n ' s \text{ teeth} \), which in fact did occur six times.\(^\text{12}\) But there are no less than a further 22 instances which vary this pattern. Here are some of the more spectacular ones, which fall into two categories: the relatively rare cases of sheer inventiveness, as in

- as scarce as heavy-metal bands at Republican rallies
- with good seats at good shows always as scarce as bagels in Mecca

and the more common pattern of context-triggered comparisons, as in

[12] Giggles on the Associated Press service are about as scarce as deadpan reporting in the National Lampoon
[13] signs of a Sino-Soviet thaw are about as scarce as palm trees in Peking and Moscow
[14] In questions of air safety, definitive answers are as scarce as anti-gravity screens
[15] an estimated 30% increase in store traffic has…made parking spaces as scarce as Cabbage Patch dolls
[16] Hollywood is finding that adaptable novels are as scarce as cheap real estate
[17] prominent women economists were almost as scarce as generals in skirts
[18] The three networks alone settled 350 staffers in Des Moines, and hotel rooms were as scarce as subways
[19] good generals were as scarce as good shoes in the Continental Army.

A similar pattern may be seen in (10), *as slippery as Y*, where only two of the ten instances are the canonical *as slippery as an eel*, with a third instance

\(^{12}\) Similar data may be found for *as rare as X*. *As scarce as could be* did not occur, although numerous other adjectives occurred in this construction.
playing on it (slippery as an eel’s hips). The others are either physically related, as in

- a brick path worn slippery as slate
- with a ball as slippery as shaving soap

or inventive:

- Kansas men (= ‘football players’) were slippery as noodles
- Slippery as wrestlers covered with oil, [neutrons] slide through the electric fields
- The stucco and chicken-wire cliffs of Hollywood success… [are] treacherous, lonely and slippery as glass
- as slippery as Wagner’s without Wagner’s soaring sense of continuity
- this principle is as slippery as a wet fox in a rabbit hole.

For (11), the structure happy as a Z, no single canonical form seems to exist; instead, CCDI lists a number of variants: as happy as {a clam, a lark, a pig in muck, a sandboy, Larry}, with clam being AmE, lark neutral, and the others BrE. No particular order of preference is given.

Here, however, the Time corpus outdoes itself. All but sandboy are represented in the corpus, some repeatedly: happy as a clam (at high tide) (4+3), a lark (5), a pig (in muck) (2+1), as well as three proper names (Lark [a car], Larry 2x). There are a further two instances of happy as a grig (‘lively person’) and three of happy as a king, but more noticeably, there are another 46 additional terms, all different. Here are a few:

- as happy as a five-year-old with his curls cut off
- wonderful Florida sunshine that he is apt to sit back, happy as a grapefruit, and soak it up
- as happy as a hayride down the middle aisle of Oklahoma!
- The Hotel Business: ...as happy as a room clerk with a waiting list.
- as happy as a Teletubby on tequila

Again, we see this mixture of context-linked and purely inventive forms. And again, it may be noted that the (various) canonical forms account for only 18 of the 70 instances. This is far more variation than that found in e.g. the BNC (Gustawsson 2006, Minugh 2006), and confirms the claim that Time is more playful in its use of idioms.

5.3 Secularization?
Although modern America is often characterized as more overtly religious than much of today’s Europe, it may nevertheless be plausibly maintained that the secularizing process has not left the U.S. untouched (cf. the discussion in Gilbert 2000: 24–44). If this is the case, there should be a shift towards less awareness of Biblically based idioms, as well.

A search of the idioms in CCDI revealed a total of 33 idioms directly deriving from the Bible (this fact is rarely indicated in CCDI, which provides little or no etymological information for its idioms, so that a few other idioms with Biblical origin may have been missed). All 33 are to be found in the Time corpus, although with frequencies ranging from 10.7 to 0.2 per 10 M. The most frequent idioms often have inflated numbers because they are the name of a contemporary book or play; the second number indicates the frequency when such references are removed:

- an olive branch (10.9, 10.7)
- a thorn in your side/flesh (9.1)
- the day of reckoning (7.4)
- wailing and gnashing of teeth (6.1)
- beat/turn swords into plowshares (5.9, 5.1 without the Plowshare movement)
- in seventh heaven (5.5, 1.7 without the play/film)
- fall by the wayside (5.1)
- fall from grace (4.6, 4.4 without the play)

Of these items, the most frequent, *(hold out) an olive branch*, is one of the idioms that most easily lends itself to more general use, since it functions as a synonym for ‘(offer to) make peace’, ‘(be willing to) negotiate’. This sense does not require a religious framework, unlike e.g. *fall from grace*, which primarily is to be construed via the scenario of Christian theology (particularly through the loaded term *grace*). In general, we may note that many of these terms have accrued specific historical and cultural meanings that may be invoked, such as the Plowshare peace group of the 80s. These specific meanings help keep the idioms alive, at the same time as they remove them from their primary use as transmitters of general human experience and wisdom.

Again, as the numbers are not very large, the aggregate numbers will be considered when investigating the change in the use of these idioms over time in the Time corpus. The results may be seen in Figure 3, below.

Perhaps the most interesting aspect of Figure 3 is the dip that occurs in the Great Depression. That was of course a time of major economic trauma and conflict, and it is possible that a religious scenario that included days of judgment was not what people wanted to hear, as the judgment would apply to them. During the 40s and 50s, a judgmental scenario may well have better fitted the concept of apocalyptic struggles with the Nazis, and then the
Communists, both of which were external enemies of America. Such speculations would, however, require an investigation into the details of the use of these idioms, a procedure beyond the present investigation.

Figure 3. Biblical idioms in the *Time* corpus (per 10 M)

In any case, with the exception of the 30s, we may possibly be seeing a pattern not unlike that of Figure 2, where there is a break with the past in the 60s: from then on, there is a relatively steady, lower level for such idioms, with no sign that the Bush administration has succeeded in convincing *Time* to bring these particular Biblical idioms back into use. A word of caution is in order, however: as these particular idioms are anchored in the Bible, their potential for being brought back into various domains of public use is considerably larger than many of the other idioms based on e.g. agricultural life of the past. In other words, they may at any time arise like Lazarus from the tomb of the forgotten.

5.4 In an age of deception

The semantic field of deception offers an opportunity to examine a relatively static area of society, as it is difficult to believe that deception would suddenly become more salient to human activities. The need for such idioms might be expected to remain relatively constant during the near-century of the *Time* corpus, even if it is by no means given that the gatekeepers at *Time*
(i.e. its editors and policy-makers) maintained exactly the same policy towards articles mentioning such behavior.

To test this question, another feature of the CCDI dictionary was utilized: the second edition of this dictionary lists 32 different semantic areas for its idioms (2002:Index, 1–17): DECISIONS, LOVE, INFORMATION, HELP AND ENCOURAGEMENT, SADNESS, MONEY and so on. The group that will be examined here is DECEPTION, with 41 idioms. All but three of these are relatively easily searchable, and these 38 form the basis for the next investigation. Of these 38, 5 are labeled as currently British, 6 American and 1 Australian. The most common are listed below, again with the second figure indicating that proper names have been removed, and with regional labels in square brackets:

- give/pay lip service to sth (17.5, 17.0)
- go through the motions (13.7)
- cloak and dagger (8.2)
- a shell game (5.5)
- monkey business (4.6)
- made out of whole cloth (4.2) [AmE]
- sell someone down the river (2.7)
- a white lie (3.3, 2.6)

These are rather a mixed bag, with some derived from espionage (cloak and dagger), others from fraudulent tricksters (shell game), yet others from hypocrisy (lip service, go through the motions), and at least one from the American slave trade (sell sb down the river).

As with the religious idioms, the numbers are too small for reliable item-by-item analysis. The top quartile, or eight most common idioms, as listed above, account for no less than 654 of the total of 887 tokens, or 73.7%. Many of the other expressions fail to occur at all for decades at a time. Instead, the aggregate figures will again be used, the results being seen in Figure 4 (below).

Interestingly enough, the Depression of the 30s does not seem to have unleashed a storm of accusations about fraud, at least in Time. Instead, the peak appears to have been in the 50s (a decade traditionally seen as soporific, at least under the Eisenhower administration). Here, the split between the first and second half of the century does not seem to hold; rather, there is an eroding of trust that culminates during the period when fear of Communism is at its height, followed by the distrust in the Vietnam War period, with a return

---

13 As this paper is not technically oriented, the details of the searches have not been presented. For an item such as pull sb’s leg, the problem lies in the noise ratio: normally, this is a question of so many false hits appearing in a simple search that it becomes impractical to investigate the item in question.
to skepticism from the Reagan administration on through today. As noted
above, however, such a large-scale shift need not reflect actual social chang-
es, since it suffices for Time’s gatekeepers to shift policies on what they
write about or how they express it.

Figure 4: Time corpus tokens from the semantic field ‘fraud’, per 10M

Returning to the idioms themselves, we may note that a few of them were
completely missing from the Time corpus:

- do the dirty on someone [BrE]
- sell the pass [BrE]
- lie through your teeth
- be economical with the truth

In at least the latter instance, it is somewhat surprising that be economical
with the truth is listed without a regional label; it originated in an inspired
ministerial moment in the mid-80s, in connection with the Spycatcher trials
in Britain, and appears to be firmly anchored in the British political tradition,
so that if it were to appear in Time (which it failed to do) one would expect it
to be in a British connection.

Among the items that appeared rarely, one in particular stands out: to
speak/talk with forked tongue. This was a standard cliché of Hollywood
films, where the Indian brave or chief (it varied) expressed his disbelief in
the probity of the white man—usually correctly so. It clearly failed to make
it into *Time*, presumably precisely because it was such a cliché, although it appears to be making a strong comeback as a term of abuse on the Internet.\textsuperscript{14}

Again, however, it appears that a much more detailed study of the individual instances would be necessary, before making major claims about the shifts in deception idioms during these eras.

6. Conclusions

The most important conclusion to be drawn from this investigation is that the *Time* corpus comprises a fascinating resource for information about American English of the twentieth century. Its 100 M words contain a wealth of examples for many different types of investigation into linguistic and cultural matters, and the corpus search engine is sufficiently sophisticated to provide access in a relatively non-technical way.

At the same time, it must be remembered that *Time* is a collective enterprise under the control of the fairly stringent editing of one single publishing house, and represents one single genre, albeit with articles on many aspects of human life. News items from the spheres of politics and economics dominate, and of course are primarily drawn from the United States. These caveats having been stated, it nevertheless remains true that the *Time* corpus is the first major diachronic source for studies of American English, and no comparable source yet exists for British (or other varieties of) English.

With regard to general analyses of idioms in the strict sense, the size of the *Time* corpus is clearly still inadequate to achieve even the reasonable goal of 50 instances per meaning or diachronic unit examined, as set forth by John Sinclair in 1991; for most of these idioms, the entire corpus is not even sufficient to produce 50 instances in total. Here, it appears likely that only the Internet-based, automatically-retrieved corpora of the 21\textsuperscript{st} century will be large enough to begin to satisfy that criterion. But for a general analysis of English, that merely begs the question of genre, as well, since Internet writing is not equivalent to all writing—not to mention the additional mode of spoken English.

On a more positive note, the *Time* corpus appears to contain a plethora of examples for studies of how idioms may be varied, particularly through context-based (text-anchored) variation of one of their slots, as in the spectacular variation the *Time* corpus exhibits as regards constructions such as a one-X town. Since the great danger with idioms is their overuse as given clichés, it is indeed nice to know that being happy as a clam at high tide is not the only game in Time’s town.

\textsuperscript{14}The Internet as a source is beyond the scope of the present paper, but it would appear as if its untrammeled nature, as opposed to the relative dignity of *Time*, may mean that numerous ways of attacking and belittling opponents are making a strong comeback via the Internet.
References


Sammanfattning (Summary in Swedish)

Denna sammanläggningsavhandling undersöker idiomatiska uttryck i det moderna engelska språket, med data som framtagits genom korpuslingvistiska metoder. I arbetet ingår fyra tidigare publicerade delstudier, förutom ett inledningskapitel (‘kappa’) som sätter in dessa i ett bredare uppdaterat sammanhang.

Språkligt sett är idiomatiska uttryck ("idioms") kärnan i avhandlingen. Det har visat sig inte vara möjligt att formellt definiera dem som grupp, men de kännetecknas av tre huvuddrag som de allra flesta forskare accepterar idag:

1. **Fasta delar.** Ett idiom är ett (mer eller mindre) fast, flerordigt uttryck, där vissa delar eller hela uttrycket inte kan ändras utan att dess metaforiska betydelse går förlorad.

2. **Institutionalisering.** Ett idioms form och betydelse är bekanta för ett stort antal infödda talare, och finns dokumenterade i ordböcker.

3. **Icke-kompositionalitet (semantisk opacitet).** Ett idioms betydelse kan inte direkt härledas ur de olika delarnas normala betydelser, och dess betydelse är således inte ”bokstavlig”. Betydelsen kan vara mer eller mindre svår att gissa.

Några exempel på engelska: *to take a busman’s holiday* (ung. ‘att ha arbetssemester’, dvs att på fritid utföra samma uppgifter som när man yrkesarbetar), *to have bats in your belfry* (‘att ha tomtar på loftet’) eller *to be a dime a dozen* (‘att vara en dussinvara’) uppfyller dessa kriterier: man kan inte plocka in en annan yrkesgrupp än busmen eller ens byta ut holiday mot day off/vacation; likaså med bats och belfry (även om australiensare hellre än-avänder en annorlunda, men liktydig fras, *to have kangaroos in the upper paddock*); inget annat myntslag eller mängdbegrepp fungerar för a dime a dozen (om man säger att något kostar 5 dollars a dozen, så är det en prisuppgift, inte en anmärkning på kvaliteten). Alla tre idiom finns belagda i ordböcker och förekommer i tidningar and romaner. Slutligen kan man notera att uttrycket inte avslöjar exakt vad som menas, även om man kan ana att gräsärande känguruer i bra hagmark inte är jordbruksmässigt försvarbart.

Idiomatiska uttryck tycks vara en naturlig del av språk, även om man lätt kan konstatera att idiom ofta är språkspecifika: ett uttryck i engelskan kan saknas
på svenska (som i fallet busman’s holiday ovan), eller tvärtom (att elda för kråkorna eller ingen ko på isen måste på engelska uttryckas via omskrivningar); i vissa fall finns ett motsvarande uttryck hämtat ur en annan domän (to have bats in the belfry – att ha tomtar på loftet), men de kan även vara nästan exakt lika (to take the bull by the horns – att ta tjuren vid hornen). Som ytterligare exempel kan man nämna ha tomtar på loftet, ta sin Mats ur skolan, ett måndagsexempler, mota Olle i grind, för att nämna några exempel. Läsaren som vill botanisera i ett stort antal idiom och ordspråk på flera språk hänvisas till Martinsson 2004.

Det har redan anytts att de tre huvudkriterier på idiom alla har sina brister, som man bör känna till:

1. ”Fasta” uttryck kan ibland vara flexibla, t ex genom possessiva pronomen, t ex bats in {his, her, their, my, your, our} belfry. Vidare kan de ofta varieras grammatiskt sett (olika tempus, passivisering osv), t ex let the cat out of the bag ’avslöja en hemlighet’ > No cats were let out of the bag at today’s meeting, thank God! Man kan ibland variera med verb eller substantiv ur samma semantiska fält, t ex to {cut, slash, pare, shave} to the bone ’att göra drastiska nedskärningar’, to be a load of {nonsense, crap, bull, baloney, shit} ’att vara komplett nonsens’, men ibland går det inte alls att variera (as the crow flies ’fägelvägen’).

2. När det gäller hur känt ett uttryck är kan det finnas ålders- och regionrelaterad variation; rent allmänt tar det tid att lära känna idiom, då de är tämligen sparsamt förekommande, vilket är ett återkommande tema i avhandlingen. Ingen ko på isen tycks vara mindre vanlig än ingen fara på taket, till exempel.

3. När det gäller hur idiom är sammansatta, kan man notera att ett idiom kan vara allt ifrån helt obegripligt (ont krut förgås inte så lätt) till tämligen genomsnitslig när man vet mer om dess uppkomst (i detta fall kommer ont krut av tyskans Unkraut ‘ogräs’), till helt begripligt även när man hör det för första gången (om avstånd: fågelvägen). Detta gäller avkodningen, men inkodningen är mycket svårare, även för infödda talare, då man inte i förväg kan lista ut hur en tanke blivit konkretiserad i ett språk: fågelvägen är knepigt nog för ’raka/gena vägen’, men engelskans motsvarighet, as the crow flies, är helt omöjlig att gissa sig till.

Definitionen kan egentligen utökas med ytterligare en dimension: uttrycket för nästan alltid in en ny, konkret situation (scenario) och föremål i texten: när man tar tjuren vid hornen eller lets the cat out of the bag, är det sällan en verklig tjur eller katt är med i berättelsen eller samtalen. Vi tycks framkalla en berättelse där dessa figurer redan ingår (grammatiskt sett syns det i de
bestämda formerna i *tjuren*, *hornen*, *the cat* och *the bag*), och där handlingen är klar. När man missförstår ett idiom är det ofta så att man associerar i en riktning som inte är den som av hävd anses rätt, som i en tolkning av *att göra någon en björntjänst*, där *björn-* tas som tecken på något stort och positivt, medan den traditionella betydelsen är stort och negativt (björnen som av ren god vilja skulle slå ihjäl en fluga på ägarens huvud).


Till skillnad från tidigare forskning om idiom på engelska präglas den moderna diskussionen av att man i mycket stor utsträckning hämtar exemplen från korpusmaterial, dvs texter (skriftliga eller muntliga) som har yttrats eller skrivits i syfte att kommunicera, och sedan systematiskt samlats in. Så är det även här: nästan utan undantag är exemplen hämtade ur korpustexter.

Idiomforskare har naturligtvis alltid samlat på exempel, men i och med att korpuslingvistik på allvar gjorde sitt intåg på 1980-talet har forskare kunnat arbeta systematiskt med större textmassor, och därmed kunnat få fram detaljerad information om t ex frekvens och även om mindre vanliga språkliga fenomen. Idiom tillhör deras mera sällsynta uttryck (var för sig, dock inte som grupp), och Moons 1998 studie *Fixed Expressions and Idioms in English: A corpus-based approach* markerar starten på denna process.

I 1.1.2 presenteras mycket summariskt korpusar och korpusforskning på engelska. Alltsedan 1960-talet har forskning om det engelska språket varit i den gynnsamma situationen att större korpusar och bättre sökmotorer alltid har utvecklats på engelska texter, och först senare på andra språk, vilket på gott och ont har gjort korpusforskning inom engelska till ett pionjärarbete, något som först nu har börjat klinga av.


Kapitel 1.2 lägger fram avhandlingens syfte och frågor. Ett övergripande syfte är att genom användning av korpusmetoder undersöka olika mönster i bruket av idiom på engelska, främst hur dessa kan varieras i frekvens och lexikaliskt sett, men även med hänsyn till geografi (brittiskt/amerikanskt), texttyp och textfunktion.

Detta leder fram till den första frågan: hur ser språket ut i en korpus på några miljoner ord som enbart består av internettexter från tidningar av och för universitetsstudenter? Dessa unga skribenter (18–24 år, nästan alla med engelska som modersmål och utbildningsspråk) skriver för sina gelika, vilket är en texttyp helt olik de olika ”learner corpuses”, vars studentalster huvudsakligen kommer från studenter som inte har engelska som modersmål, och skriver uppsatser för språkkurser. Inte heller har de stora korpusar som nämnts ovan en sådan ”studentskribent”-komponent.

De övriga frågorna gäller idiom i de olika korpusar som användes: hur ofta förekommer idiom i dessa olika texter och texttyper, dvs frågor om frekvens i olika sammanhang? Är vissa grupper idiom mer frekventa i olika korpusar, eller vissa tidsperioder? Hur varierar idiom lexikaliskt sett, och hur varierar de när de används i olika textsorter? Då dessa studier genomfördes när de olika korpusarna var nyutkonna var svaren inte alls givna. Bakom dessa frågor fanns det dessutom ytterligare ett problem: eftersom idiom är lågfre-


Avsnitt 1.3 tar upp metoderna för dessa undersökningar och 1.4 deras resultat. Den första frågan är vad som ska ingå i korpusen. Man kan antingen använda sig av redan existerande korpusar som man inte kan påverka eller skapa egna med det innehåll man vill ha. ”Hemmagjorda” korpusar medför risken att de inte alls är så professionellt gjorda (och med betydligt sämre metadata), men samtidigt kan de fylla en nisch som annars inte blir undersökt. Det finns även en pedagogisk poäng med denna demokratisering av forskning.


Studie 1 handlade om att skapa och undersöka en mindre korpus (storleksordning 3–4 miljoner ord). Materialet fanns på nätet, och var därmed oftast lättillgängligt; tekniskt sett var det för det mesta inte svårt, och behöver inte diskuteras vidare. Däremot fanns det avvägningsfrågor som var betydligt
svårare. Först, geografisk representativitet: man vill gärna att textmängden för varje land motsvarar dess andel befolkningsmässigt, men i det här fallet är det svårt, eftersom USA proportionerligt sett hade fler studenttidningar utlagda på nätet, medan t ex Sydafrika knappast hade några sådana alls. I förlängningen blir problemet ännu svårare, då många f d Commonwealthländer som Indien och Kenya inte alls fanns på nätet, fast universitetsutbildningen bedrivs på engelska.

En tänkbar lösning är att ta med mer text från underrepresenterade länder, men detta i sin tur introducerar en annan form av skevhett, då just dessa texter får oproportionerligt stort inflytande. I stället valdes ut en komplett upplaga från varje tidning (vissa gav ut nytt material dagligen, andra en gång i veckan, och somliga hade bara orkat ge ut en enda utgåva under flera år), så att bredden prioriterades över geografin. Detta till trots var den geografiska skevheten mindre än 10%, med över 300 olika universitet från USA, Kanada, Storbritannien, Irland, Sydafrika, Australien och Nya Zeeland representerade, alla från våren 1999. Då denna korpus var helt ny och okänd, jämfördes sedan ett antal lexikaliska och grammatiska faktorer som kunde förväntas uppvisa variation mellan åldersgrupper (Coll-skribenter är ju mycket yngre än de som representeras i de flesta korpusar) och länder. Vissa trender och förändringar kunde bekräftas, t ex att former som elder, eldest trädde mycket kraftigt tillbaka för mera regelbundna former som older, oldest. Däremot var den amerikanska nybildningen sneak, snuck, snuck ’smyga (på)’ på väg att bli lika vanlig som sneak, sneaked, sneaked.

Studie 2 använde sedan Coll för att i detalj undersöka idiom i dessa yngre skribenters texter: hur frekventa var idiomen, vilka var vanligast, var några särskilt användbara i titlar? Det visade sig att i huvudsak var användningen ungefär som för de större korpusarna, men att frekvensen varierade stort mellan olika idiom. Flera amerikanska idiom fanns med som saknades i CCDI-listan, men även idiom från bägge sidor av Atlanten—ett tecken på att idiomboken behöver baseras på en större, bredare korpus (CCDI hade en klar brittisk slagsida år 2002).

Studie 3 inriktade sig i stället på hur idiom infogas i huvudtexten: ett idiom är normalt sett komplett i sig (som noterats ovan), och har inget direkt att göra med texten det infogas i. Men genom att lägga till ett ord eller en fras som har att göra med texten (och inte idiomet) kan man ”förankra” idiomet i texten, t ex Plummer’s comments were proof that opponents would push what he called “the hidden homosexual agenda”, där homosexual har inget att göra med en hidden agenda som sådan, utan specificerar vilken dagordning som är gömd. Studien undersökte flera stora korpusar (300 miljoner ord, med en stor komponent talat språk, och mycket tidningsstexter, från Storbritannien och USA) för att se hur vanlig denna variant och andra
grammatiska möjligheter, t ex efterställning, är. Överraskande nog visade det sig att förankring via framförställda adjektiv var sällsynt (< 5%), men att vissa idiom tillåt den i mycket högre grad (upp till 20%). Detta bekräftar dels att idiom inte alls behöver anpassas till den omkringliggande texten, dels att man inte bör ”dra alla idiom över en kam”.


Avsnitt 1.5 granskar dessa studiers upplägg, förtjänster och brister, och relaterar dem till nyare forskning. Relativt lite har ändrats, teoretiskt sett, men desto mer när det gäller tillgång till alternativa korpusar och mängd data att tillgå. Diskussioner om korpusdesign och representativitet tyder på att en större statistisk medvetenhet är viktig, t ex när man bestämmer brytpunkter i diakroniska undersökningar. Detta har hittills inte varit lika viktigt för undersökningar om idiom, där huvudproblemet varit att få fram tillräckligt med exempel, men blir allt viktigare för arbete med större korpusar i framtiden.

Respekt för genrer och domänar har vuxit fram på senare år: en korpus som *BNC* är visserligen en enhet (brittiskt språk under 1990-talet), men består

Det som framgår av alla dessa undersökningar är att idiomatiska uttryck har en icke oansenlig mängd varianter som kan vara svårupptäckta, om man bara letar efter standardformerna. Däremot är det rätt sällsynt med texter som bygger in idiom som rubrik eller som en nyckeldel av strukturen.

Detta bekräftas i en specialstudie i avsnitt 1.6, där alla ledartexter i Coll identifieras och undersöks för hur deras idiom fungerar i texten. Ledartexter är den mest idiomrika tidningsdelen över lag (nyheter är den idiomfattigaste delen) och där ser man tydligt att mycket få idiom finns i ledarrubrikerna (jfr. Jaki 2014 om brittiska tidningar, som tycks följa ett annat mönster). Däremot är det mycket vanligt att idiom i ledartexter har en emotionell laddning (negativ oftare än positiv) och mycket vanligt är att de väljs ut för att passa in i textens huvudbudskap: om man kräver att universitetsledningen fattar ett svårt beslut ligger det nära till hands att ett idiom som take the bull by the horns dyker upp. Dock kan man inte förutse att ett visst idiom dyker upp—det är slående hur stor variationen är: endast ett par idiom förekommer mer än två gånger i dessa 112 ledare (med 60,000 ord och 148 förekomster av idiom), och inget idiom förekommer i mer än tre ledare.

I avsnitt 1.7 prövas ett annat sätt att greppa idiom: uttrycket the glass ceiling ’osynlig barriär för främst kvinnor i karriär’ (vars motsvarighet glastaket förekommer på svenska, om än inte i lika stor omfattning) undersöks ur olika synpunkter, på basis av närmare 40 belägg ur amerikanska radioprogram från 1995, då uttrycket var rätt nytt. Man kan t ex konstatera att nästan alla verb som är länkade med the glass ceiling uttrycker smärta och/eller våld: {bump into, bang, hit, smash, break through}. Den kanske intressantaste aspekten är grammatisk: i tre fall handlar det om dancing on the glass ceiling, vilket är ett glädjerikt sätt att helt komma förbi hindret, och i alla tre är det antingen en fråga eller en önskan om en bättre framtid (When will we...), inte ett konstaterande av något som är. Sammanlagt kan dessa 36 exempel nästan uppnå Sinclairs krav på 50 belägg, dock gällde hans krav 50 per betydelse (och i en mängd olika domäner).
I 1.8 presenteras en sammanfattning och förslag till framtida forskningsuppgifter. Snart har det gått 15 år sedan Coll insamlades, och 15 och 30 år har inom korpuslingvistik blivit standardintervaller för diakroniska hållpunkter, så tiden borde vara mogen att skapa en Coll2015. Det finns betydligt fler online-tidningar från universitet nu än då, och fler länder är representerade.

Del 2 består av de enskilda tidigare publicerade uppsatserna, som här publiceras i oförändrat skick.

Sammanfattningsvis kan man enligt författarens mening hävda att det viktigaste när det gäller framtida idiomstudier är att konsekvent börja utnyttja alla nya storkorpusar för att kunna undersöka idiom var för sig i detalj. Är man intresserad av en viss domän kan man dessutom skapa mindre korpusar som tillåter detaljstudier om just denna domän. Allt detta bör göras innan man försöker dra allmänna slutsatser om hur idiom fungerar i stort. Som man säger på engelska, The devil is in the details, men det är även där som man hittar sanningen.