コーパスとしての電子辞書の数量的考察*

中郷　慶

Electronic Dictionaries as Corpora: a quantitative analysis

Kay Nakago

1　はじめに

コンピュータ本体や周辺機器の開発と普及にともなって、さまざまな学問分野において新しい研究方法が模索されている。言語学でもそれは同様で、研究を推進するさまざまなソフトウェアが開発されてきた。コンピュータを利用した研究の何よりの利点は、膨大なデータをほんの一瞬で処理できるということである。従来の手作業によるデータ収集と分析では、何年もの歳月をかけてやっと可能であった研究や、そもそも不可能であった調査が、コンピュータを利用すればごく短時間で、しかも「正確」な計量が可能となる。例えば、Stubbs (1997)はある単語が、しばしば、女性(female)と軽蔑的(pejorative)な意味合いの両方の含意を持つという仮説を、2億語のコーパスを利用して立証した。Stubbs は、little ともっともよく共起するのは girl であり、man と共起する場合は、ridiculous little man のような句であるということを示した。little についてのこの事実は、考えてみれば明白なことであるが、little の実際の振る舞いの様子は、大規模コーパスを調査してのみ立証可能である。この例の他にも、コンピュータの助けなくしては全くなしえなかった研究がたくさんある。しかし、コンピュータを用いて電子テキストを分析する場合に忘れてはならないのは、研究対象とする電子テキストの規模や内容が分析に耐えられるものであるのかについての配慮である。また、分析の方法論が正しいかについても常に考察が必要である。なぜなら、コンピュータは与えられた素材を計算するだけで、出力データが全く信頼できない可能性があるからである。

この論文では、コーパス研究の具体的な方法を提示し、電子辞書がコーパスとして利用可能であるかを検証する。読者によるデータの検証、再現を容易にするために、データ算出の具体的な方法もできるだけ明示する。本論文の構成は次の通りである。2節では、コーパス研究に際して留意すべき事柄を簡単に述べる。3節では、電子化された辞書、つまり、辞書の CD-

＊　本研究は1999年度愛知淑徳短期大学学術研究助成研究奨励費（個人）の助成を受けて行われた。

ROM 版をコーパスと見なした場合に得られるデータの妥当性を、大規模コーパスから得られるデータとも比較しながら検討する。4節はまとめである。

2　コーパスとは何か

「コーパスとは機械可読なテキストの集合である」とするのが、コーパスについての一般的な理解である。しかし、コーパスは必ずしもコンピュータ化されている必要はない。また、その内容は書き言葉であっても話し言葉であってもよい。ただし、気を付けなければならないのは、機械可読なテキストの集合がすなわちコーパスではないということである。中郷(1999)では、この点について議論し、コーパスを分類するのに、広義のコーパスと狭義のコーパス、また、サンプルコーパスとモニターコーパスを分けて考えることが重要だと述べた。ここで、その内容を簡単に振り返っておこう。広義のコーパスとは、「コンピュータを利用した言語研究のために、提供されるデータ、テキストデータベース」のことであり、その構成については考慮されない。一方、狭義のコーパスとは、そのコーパスを構成するテキストの特徴について言及されなければならない。つまり、狭義のコーパスを構成するテキストは、言語研究のために、特定の原理や方針に従って集められたものでなければならない。Edwards (1993)の記述(1)は重要である。[1]

(1)　It is common to distinguish between corpora and textbanks. These differ in size and composition, and serve somewhat different analytic aims. Corpora are intended to be representative of some specified population of genre. Textbanks tend to be collections of available data with looser connection to each other, or focus on a restricted number of genre (including perhaps only one).

Edwards (1993: 282-3)

(1)の主張は、コーパスとはある言語や方言、またはある言語の何らかの下位集合を代表するものであり、言語研究はそのようなコーパスを利用すべきであるというものである。したがって、たまたま手元にある電子テキストはコーパスではなく、テキストバンクだということになる。Leech (1991)も(コーパスの設計や内容ではなく)ただ単に、大きさに焦点を当てるのは4つの理由から "naive" であるとしている。(2)はその4つの理由のうちの1つである。[2]

(2)　(...) a collection of machine-readable text does not make a corpus. The Brown and SEU corpora were carefully designed as systematic collections of samples, so as to have face-validity as representative of 'standard' varieties of English.

Leech (1991: 10)

現在では、英語で書かれた文学作品や新聞・雑誌記事がインターネットや CD-ROM を通じて容易に、しかも大量に入手できる。Edwards や Leech の主張は、コーパス言語学がその研究対象としなければならないのは狭義のコーパスであり、インターネットや CD-ROM を通じて入手できるそのようなテキストデータ (すなわちテキストバンク)を、直ちにコーパス言語学

の対象として利用するのは危険だということである。

公開されているコーパスにはさまざまなものがあるが、それらを内容によって分類することがある。それは、サンプルコーパス(sample corpus)とモニターコーパス(monitor corpus)の区分である。サンプルコーパスは収集されているテキスト量が一定であり、編纂時期の言語表現の全体像が反映されるように幅広い分野からバランスを考えて作成されている。Brown Corpus や LOB Corpus はサンプルコーパスである。一方、モニターコーパスは内容的な均質よりもデータ量の拡充を重視し、常に変化する言語を監視しながら、古い情報を捨て新しい言語情報を付け加え、最新の言語情報を提供しようと作成されたコーパスである。その代表的なものは3. 4節で見る Bank of English とそのオンライン検索サービス COBUILD*direct* であり、その規模は1998年7月現在、およそ3億3千万語に上る。モニターコーパスの特徴は、さまざまな分野の最新テキストを大量に保有することであり、近年の語彙研究や辞書作成には欠かすことのできないものになっている。[3]

コーパスを入手するもっとも一般的な方法は、コーパスを収録した CD-ROM を購入するか、インターネットでダウンロードするかのいずれかである。[4] ノルウェーの ICAME (International Computer Archive of Modern and Medieval English) からさまざまなコーパスを1枚の CD-ROM に収録した "ICAME Collection of English Language Corpora" (以下 ICAME-CD)が入手できる。1999年6月に刊行された第2版に収録されているコーパスは以下の通りである。[5]

コーパス名	内容や特徴	対象	語数
Australian Corpus of English (ACE)	現代オーストラリア英語のコーパス。Brown Corpus, LOB Corpus に準拠。	書	100万語
Brown Corpus*	アメリカ合衆国ブラウン大学で開発された現代アメリカ英語のコーパス。15のカテゴリー、100万語からなり、1961年の各種の刊行物が対象。	書	100万語
Corpus of Early English Correspondence Sampler	1417年から1681年までのさまざまな人物による書簡のテキストを収録	書	45万語
The Bergen Corpus of London Teenage Language (COLT)*	ロンドンのティーンエイジャーの英語コーパス。	話	50万語
Freiburg-LOB Corpus of British English (FOB)	ドイツのフライブルグ大学で1991年にスタートしたプロジェクトで、Brown Corpus, LOB Corpusの1990年代版。Brown Corpus, LOB Corpusとの比較によって、30年の推移が観測可能。	書	100万語
Freiburg-Brown Corpus of American English (FROWN)		書	100万語

コーパス名	内容や特徴	対象	語数
Helsinki Corpus of English Texts, Diachronic Part	ヘルシンキ大学で1984年にデータ構築を開始。800-1710年の古英語、中英語、初期近代英語の3期から 150万語を収録。	書	150万語
Kolhapur Corpus of Indian English	1978年の出版物から集められたインド英語の書き言葉のコーパス。Brown Corpus, LOB Corpus に準拠。	書	100万語
Lampeter Corpus of Early Modern English Tracts	The University of Wales, Lampeter 所蔵の Tract Collection (1640-1740)のフルテキストのコーパス。	話	50万語
London-Lund Corpus	主として1960年代教養層のイギリス英語の話し言葉を収録。	書	50万語
Lancaster-Oslo-Bergen Corpus (LOB)*	イギリスのランカスター大学とノルウェーのオスロ、ベルゲン大学で1970年代に開発された Brown Corpus のイギリス英語版。1961年出版のテキストが対象。	書	100万語
Lancaster Parsed Corpus	LOB Corpus の各カテゴリーから計133,000語分について品詞標識を付け、構文解析したコーパス。	書	133,000語
Newdigate Newsletters	1674年から1715年の初期近代英語期にかけて書かれた通信文書のマニュスクリプト2,100通を収録。	書	75万語
Helsinki Corpus of Older Scots	ヘルシンキ大学で制作された1450年から1700年までのスコットランド英語を15ジャンルの散文から収録。	話	83万語
Polytechnic of Wales Corpus	子供のことばを収録したコーパス。転写は正書法に基づき、音韻データは含まれない。Halliday の体系機能文法に従って、構文解析されている。	話	61,000語
Lancaster/IBM Spoken English Corpus (SEC)	現代イギリス英語のコーパス。正書法に基づくテキストと音韻に基づく転写を収録。	話	55,000語
Wellington Corpus of Written New Zealand English*	LOB Corpus に準拠し、1986年以降の出版物から収集されたニュージーランド英語のコーパス。	書	100万語
Wellington Corpus of Spoken New Zealand English	ニュージーランドのヴィクトリア大学で1985年から10年かけて作成されたインフォーマルなニュージーランド英語(話し言葉)のコーパス。日常的な話し言葉が75%と大きな割合を占めるのが特徴。	話	100万語

*印は文法標識が付与されたバージョンも含まれることを表す。

「対象」の欄の「書」は「書き言葉」(written)、「話」は「話し言葉」(spoken)を表す。

表1　ICAME Collection of English Language Corpora 収録コーパス

3　電子辞書はコーパスとして有効か

前節ではコーパスを分類し、さまざまなコーパスの概要を見た。コーパスはさまざまな基準に適合するように設計する必要があり、コーパスには広義のものと狭義のもの、サンプルコーパスとモニターコーパスがあると述べた。この意味においては、電子化された辞書は、どれだけ例文が記載されていたとしても、コーパスとして見なすことはできないことになる。この節では、辞書のCD-ROM版を用いて、どの程度の信頼性のおけるデータが得られるかを、特に数量的に考察する。まず、通時的なコーパスとしてHelsinki Corpusを取り上げ、これを*The Oxford English Dictionary*と比較することから始めよう。

3．1　Helsinki Corpusの概要

Helsinki Corpusは正式名を“The Helsinki Corpus of English Texts: Diachronic and Dialectal”といい、750年から1710年の1,000年間に書かれたさまざまなタイプのテキストを集めたもので、通時的言語資料を集成したDiachronic Partと方言資料を集成したDialectal Partの2部からなる。前者はさらに、本文中でHelsinki Corpusと呼ぶことにしている古英語(Old English: OE)・中英語(Middle English: ME)・初期近代英語(Early Modern English: EModE)からなる‘basic corpus’(160万語)と、スコットランド英語(1450–1700年)を集成したHelsinki Corpus of Older Scots(83万語)および初期アメリカ英語(1620–1720年)を集成したThe Corpus of Early American Englishからなる‘supplementary corpora’で構成される。Helsinki Corpusは今日の英語史研究には欠くことができないものになっていると言っても過言ではない。その時代ごとの収録語数は表2の通りである。

時代区分	年代	語数	割合（%）
OE1	-850	2,190	0.52
OE2	850–950	9,250	2.24
OE3	950–1050	251,630	60.89
OE4	1050–1150	67,380	16.30
OE 小計		413,250	100.00
ME1	1150–1250	113,010	18.56
ME2	1250–1350	97,480	16.01
ME3	1350–1420	184,230	30.27
ME4	1420–1500	213,850	35.13
ME 小計		608,570	100.00
EModE1	1500–1570	190,160	34.51
EModE2	1570–1640	189,800	34.44
EModE3	1640–1710	171,040	31.04
EModE 小計		551,000	100.00

Kytö (1993[2]: 2)

表2　Helsinki Corpusにおける各時代区分ごとの語数

3．2　OEDの検索と注意点

The Oxford English Dictionary（以下OED）では、用例を幅広くできるだけ多く収集し、語義解釈の区分に沿って、初出ものから年代順に用例を並べている。語義の変化を年代的にたどる「歴史的原理に基づいた」(on Historical Principles)編纂法こそが、OEDの最大の特徴であり、その内容は規範的(prescriptive)ではなく記述的(descriptive)である。1150年頃に廃語であったものを除き、それ以後使用されてきた単語を、その異形を含め616,500あまり収録している。見出し語は290,500ある。

OEDの出版の歴史は100年以上にわたる。OEDの前身となる*The New English Dictionary*の編纂が1858年に決定し、OED全巻の出版が完結したのは1928年である。その後1933年に1884年以降の新語と新語義を加えたSupplementとBibliographyが追加出版され、さらに、1972年から1986年にかけて4巻からなる新補遺が加えられた。1989年には新補遺の内容をすべて本巻に盛り込んだ第2版が出版された。今回の検索の対象としたのは、この第2版をCD-ROM化したものである。総用例数は240万以上にのぼる。以下このCD-ROMをOED2-CDと表記する。[6]　このように、OEDの出版には100年以上の歳月を要し、その間には6人の編集者と多くのスタッフが携わってきた。したがって、引用などに関して不規則性があることは避けられない。OEDを使った研究では、このことを特に留意しておかなければならない。[7]

次の表はOED2-CDの各時代ごとの引用数を示したものである。時代区分はHelsinki Corpusと同一である。

時代区分	年代	OED引用数	Helsinki Corpus語数
OE1	750- 850	1,164	2,190
OE2	850- 950	6,030	9,250
OE3	950-1050	14,438	251,630
OE4	1050-1150	2,846	67,380
OE 小計	750-1150	22,363	413,250
ME1	1150-1250	27,474	113,010
ME2	1250-1350	56,023	97,480
ME3	1350-1420	85,118	184,230
ME4	1420-1500	91,558	213,850
ME 小計	1150-1500	250,353	608,570
EModE1	1500-1570	133,413	190,160
EModE2	1570-1640	295,131	189,800
EModE3	1640-1710	251,965	171,040
EModE 小計	1500-1710	673,028	551,000

表3　OEDにおける引用数とHelsinki Corpusの語数

このような作業が可能なのは、まさしくコンピュータ技術によるものであり、OEDの真価はコンピュータを利用することによって初めて発揮されるといっても過言ではない。引用数の算出方法についての情報は、さまざまな点で有益であると思われるので、ここにその方法を記載する。

1）　OED-CD の検索ソフトで、[Search]メニューから[Quotation]を選択して現れたウィンドウ内に、検索したい年代を打ち込む。(図1は、Helsinki Corpus の OE 2期に相当する「850-950」を打ち込んだもの。)

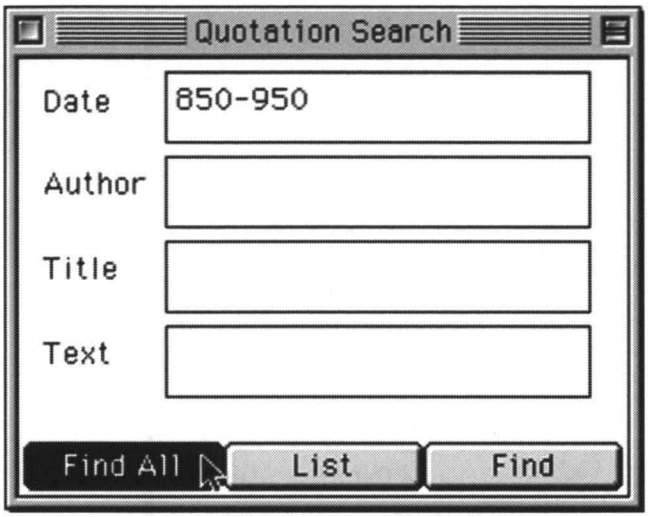


図1　OED2-CD の検索ソフトにおける引用検索ウィンドウ

Find All ボタンを押すと、引用数とともに関連するテキストが現れる(図2)。List と Text の表示はウィンドウのボタンによって切り替える。図2は引用対象の単語の一覧 (List) とその引用 (Text) の両方が表示されている。

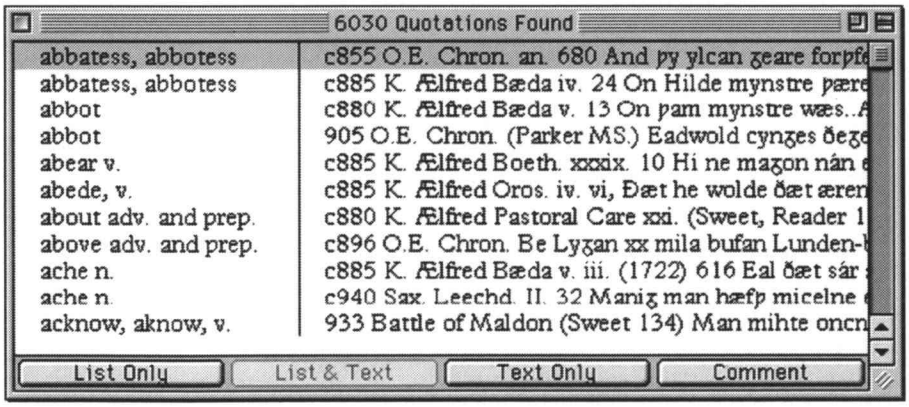


図2　850-950年の引用一覧。全部で6,030あることが分かる。

2）　引用検索結果を保存するためには、[File]メニューの[Save]コマンドを選択し、ファイル名を“850-950.quo”のように保存する。

3）　引用をテキストファイルに書き出すためには、2）で作成したファイルを利用する。[File]メニューから[Output to Text]を選択し、2）で作成したファイルを選ぶ。次に、"Quotation Text" のボックスをチェックし、[OK]ボタンを押す(図3)。

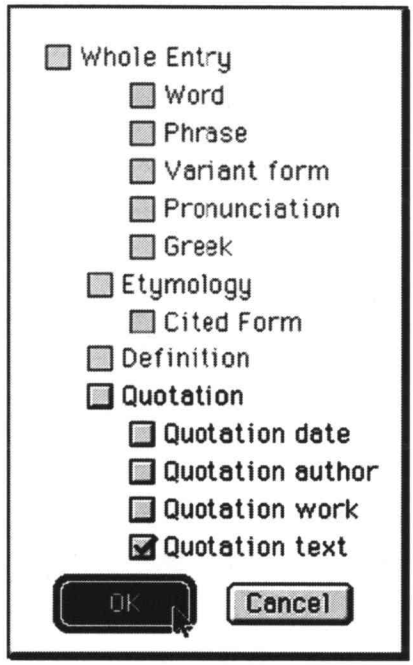


図3　引用をテキストに書き出すときの操作

注意しなければならないのは、OED2-CD には、引用年代のラベル付けがされていないものがおよそ2,500あるということ、また、単にModとだけラベル付けされているものがあるなどである。もちろん、表3には、このような例は含まれていない。OEDの引用数とHelsinki Corpusの語数を比較すれば、OEDからはかなりの例文が得られることが分かる。ここで、表3の各期の数を合計しても、各時代区分の小計と一致しないことに注意しよう。例えば、OE小計の数値は、OE1期からOE4期までのそれぞれの引用数を合計したものではなく、750年から1150年までの引用数を一括して調べたものである。なぜこのような違いが生じるかは、すぐ下で論ずる。[8]

ところで、引用数は上記の通りであるが、語数に換算すればどの程度になるのかという疑問が次に生じる。しかしながら、この疑問に答えるのは容易ではなく、いくつかの問題を乗り越えなければならない。1つ目の問題はOED2-CDの検索ソフトの設計に関わる問題である。引用をテキストファイルに書き出す方法は、上述のとおりであるが、書き出せるサイズの上限は2MBである。したがって、ある時期のテキストすべてを一度に書き出すことができない場合もある。そうすると、ある時期を分割して、別々に書き出せばよいのではないかということになるが、このようにした場合、第2の問題が生じる。例えば、上で示した方法でOE 2期の

引用数を算出すれば、6,030例あることが分かる。しかし、OE 2 期を850-900年および901-1050年の 2 期に分けて算出すれば、それぞれ4,485例と1,547例の合計6,032例となり、一度に算出したときに比べて 2 例の誤差が生じる。この現象は、OE 2 以外の時期でも生じる。なぜこのようになるのであろうか。それは次のような理由による。例えば、twelve という語彙項目の引用の一つには、"*a* 900, *c* 950" という標識が与えられているために、twelve は850-900年と901-950年の両方の時期で重複してカウントされる。[9] したがって、表 3 に挙げられた数は、絶対的なものとしてではなく、概数としてとらえられなければならない。

また、引用(quotation)の数と引用文・句・節(passage)の数が必ずしも一致していないことにも注意を要する。この理由については、注11を参照のこと。再びOE 2 期を例に挙げる。OE 2 期の引用の数は6,030であるが、実際に引用されている文・句・節の数は5,905であり、語数は53,614語である。[10] この53,614語中には厳密な意味では引用文・句・節の一部ではない語も含まれているが、誤差の範囲内にあり、1 つの引用文・句・節の平均語数は 9 から10であると言える。[11] これらの様子をまとめたものが表 4 である。なお、表 4 には、Brown Corpus や LOB Corpus がデータ収集の対象としている1961年の OED2-CD からの引用数も示してある。8,647の引用、8,274の文・句・節中に138,469語が生起している。1000年、1400年および1600年を選んだのは、Helsinki Corpus の OE, ME, EModE 期との比較に際しての、それぞれの時期の中心となる年代の平均語数の概数を知るためである。

年代	引用数	文・句・節数	文・句・節の語数	文・句・節中の平均語数
OE2	6,030	5,905	53,614	9.1
1000	10,307	10,161	77,912	7.7
1400	23,158	22,854	248,761	10.9
1600	8,413	8,305	103,764	12.5
1961	8,647	8,274	138,469	16.7

表 4　OED2-CD中の各期における引用数と語数の関係

3．3　Helsinki Corpus と OED のデータ考察

家入 (1997) は whether ... or no/noo/none 構文の whether ... or not 構文への歴史的変化を Helsinki Corpus から100例、*The Bible in English* から400例を引用し、whether ... or not 構文は18世紀の後半から優勢になったことを統計的に示している。[12] また、新井(1998)では、Helsinki Corpus と同時期の1710年までの間に約240用例が OED2-CD から収集できると述べられている。OED2-CD 全体から収集できるのは約700用例にもおよび、*The Bible in English* の約400用例を上回っている。少なくとも、数量的には、OED2-CD は *The Bible in English* に勝るとも劣っていないと言える。表 5 に示されるとおり、OED2-CD から得られる

資料でも、whether ... or no/noo/none 構文の whether ... or not 構文への交代の時期は18世紀であると言えそうで、家入の「18世紀後半」とほぼ同じ結論が得られる。[13]

	1350	1400	1450	1500	1550	1600	1650	1700	1750	1800	1850	1900	1950
or no	0	**3**	**3**	**4**	**10**	**40**	**50**	**49**	12	13	15	17	7	1
or noon	0	2	1	0	0	0	0	0	0	0	0	0	0	0
or none	0	0	0	0	1	0	0	2	0	0	0	0	0	0
or not	0	0	0	0	23	17	12	29	**23**	**23**	**59**	**133**	**66**	**140**

新井(1998: 229)

表5　OED2-CD における whether or not 構文の生起

この事実に加え、前節での数量的比較を考慮に入れれば、OED2-CD は、少なくとも通時的な考察に関しては、数量的にも内容的にもコーパスから得られる知見を補完できると言えそうである。

3．4　共時的コーパスと電子辞書、オンラインコーパスとの比較と考察

前節では、OED2-CD と Helsinki Corpus を比較して、OED2-CD から得られるデータは、Helsinki Corpus から得られるデータを補完できることを見た。この節では、電子辞書から得られるデータと、共時的コーパスから得られるデータを比較して、妥当性を検討する。

田島(1995)では、LOB Corpus と Brown Corpus を用いて、cannot but＋動詞原形(root form)構文(以下、cannot but V-root)と cannot help＋動詞-ing 構文(以下、cannot help V-ing)の使用頻度を調査し、後者が優勢的に使われていると述べている。そこで、『新編英和活用大辞典』の CD-ROM 版(以下、活用辞典-CD と呼ぶ)を用いた検索を行った。活用辞典-CD の総用例数はおよそ38万例である。[14]　(4)はその一例である

(4)　a. I cannot but think that the adoption of such a course would be a serious mistake.

b. We cannot help excluding several items from our discussion.

(『新編英和活用辞典』)

検索結果をまとめると、表6のようになる。活用辞典-CD でも、cannot help V-ing 形が優勢であることが示される。[15]

	LOB/Brown	活用辞典-CD
cannot but V-root	7/3	4
cannot help V-ing	12/6	11

表6　cannot help V-ing 形の優位性

新井（1988）は、この2つの構文についてOED2-CDを用いて調査した結果を表7のようにまとめている。表7からは、20世紀前半に優勢度が入れ替わる様子が観察できる。ここでも、通時的コーパスとしてのOED2-CDの有用性が示される。

	1400	1450	1500	1550	1600	1650	1700	1750	1800	1850	1900	1950	
cannot but V-root	0	1	0	**10**	**28**	**87**	**81**	**31**	**28**	**42**	**43**	17	5
cannot help V-ing	0	0	0	0	0	0	0	13	19	31	**39**	**22**	**26**

新井（1988：228）

表7　OED2-CDにおける2つの構文の年代別生起数

次に、オンラインコーパスであるCOBUILD*direct*を使って、今問題となっている構文の検索を試みてみる。[16] ここでは上で議論した2つの構文に加えて、cannot help but＋動詞原形（root form）構文（以後、cannot help but V-root）も調査の対象とする。次の表は、調査の対象とした構文と、COBUILD*direct*での検索式との対応を示したものである。

		cannot＋help	can@＋not＋help	cannot＋but	can@＋not＋but
cannot but V-root	cannot but V-root			✓	
	could not but V-root				✓
	can not but V-root				✓
cannot help V-ing	cannot help V-ing	✓			
	could not help V-ing		✓		
	can not help V-ing		✓		
cannot help but V-root	cannot help but V-root	✓			
	could not help but V-root		✓		
	can not help but V-root		✓		

表8　COBUILD*direct*での検索に際して使用した検索式と調査対象構文の対応

COBUILD*direct*で調査の対象としたコーパスは、イギリスの書籍コーパス（ukbooks：535万語）およびアメリカの書籍コーパス（usbooks：560万語）である。得られた結果を示したものが表9である。アメリカ英語とイギリス英語の合計を見ると、cannot help V-ingはcannot but V-rootに比べて多く用いられているものの、その優位性は絶対的なものではない。

	UK books; fiction & non-fiction (ukbooks)	US books; fiction & non-fiction (usbooks)	計
cannot but V-root	12(35%)	13(29%)	25(51%)
cannot help V-ing	13(38%)	20(44%)	33(67%)
cannot help but V-root	9(26%)	12(26%)	21(43%)

表9　COBUILD*direct* での検索結果(ukbooks および usbooks)

田島(1995)は1961年当時の Brown Corpus, LOB Corpus から得られたデータをもとに、1）歴史的にもっとも古い形式 cannot but V-root はイギリス英語では比較的よく見られるが、アメリカ英語ではもっとも頻度が低い　2）cannot help but V-root は非正用法と見なされることもあるイギリス英語では、使用頻度がきわめて低い。一方、アメリカ英語においては比較的よく用いられており、アメリカでは確立した語法だとする辞書、語法書等の記述を裏付けている、と指摘している。しかし、これらの指摘は、そのテキストの大部分が1990年代以降のものである COBUILD*direct* のデータにはそのまま当てはまりそうもない。アメリカ語法とされてきた cannot help but V-root 型も、イギリス英語にもアメリカ英語と同程度に現れる。(5)に COBUILD*direct* からの検索結果の一部を示す。(5a)が ukbooks から、(5b)が usbooks からのものである。

(5)　a. i. bused and scorned even after death, cannot help but remember the lesson that
　　ii. though not over-sensitive, could not help but understand him. With a glance at
　　b. i. are not tripping gaily through life cannot help but think that everyone else
　　ii. as a summer rain, and he could not help but marvel at the weapons of a woman

このように見てみると、1つのコーパスだけに依存すれば、誤った結論を導き出してしまうことが分かる。表10は表9と同じ用例検索を、ラジオ放送の原稿について行ったものである。コーパスを変えることによって、どれだけ傾向が変わってくるかが、さらに明らかになる。

	BBC World Service radio broadcasts (bbc)	US National Public Radio broadcasts (npr)	ラジオ原稿計	UK transcribed informal speech (ukspok)
cannot but V-root	2	2	4	0
cannot help V-ing	0	0	0	1
cannot help but V-root	0	5	5	7

表10　COBUILD*direct* での検索結果(bbc, npr および ukspok)

BBC では、cannot help but V-root を非正用法として認めていないために検出されないとも考えられる。また、ukspok の数値からは、イギリス英語の形式ばらない話体(informal speech)での cannot help but V-root の浸透ぶりがうかがえる。

4　まとめ

そもそも、コーパスを用いた研究をする場合は、研究の対象としたい言語のタイプを本当に代表したものであるといえるのかを確認しておかねばならない。さもなくば、どのコーパスを選択するかによって、全く異なる研究結果となってしまう。したがって、一つのコーパスからだけしか得られていない結果を直ちに定説とするべきではなく、傾向を探るためのものと考えるべきである。このことを念頭に置き、この論文では、辞書の CD-ROM 版を用いて、それをコーパスとして用いるのが妥当であるかどうかを考察してきたが、コーパスから得られる結果と同様の結果が導かれることを示した。ただし、あくまでも補助的な手段として使うべきであることを覚えておく必要がある。

注

＊この論文は、Nakago (2000) を大幅に加筆、修正したものである。検索や分析には Macintosh を用いることを前提としている。

1　Edwards はコーパスとテキストバンクを区別するのは“common”であると述べているが、テキストバンクという用語自体はそれほど多く用いられているわけではない。

2　Leech は(2)で、Brown Corpus と SEU Corpus について言及している。Brown Corpus は正式名を The Standard Corpus of Present-Day Edited American English といい、世界で初めての電子コーパスである。対象としているテキストはアメリカ英語の書き言葉であり、1961年にアメリカ合衆国で出版された本・雑誌・新聞が題材となっている。総語数はおよそ100万語である。Brown Corpus の登場は大きな意味を持ち、その作成方法は、その後作成

されたさまざまなコーパスの基礎となっている。本文中の表1を見れば、いかに多くのコーパスがBrown Corpusに倣って作成されているかが分かる。Brown Corpusでは、情報散文と創作散文を全部で15のカテゴリーに分け、それぞれに異なる比重を与え、その比重に従って、1つのテキストの長さがおよそ2,000語という比較的小さなサンプルを無作為に選び、それによって当時のアメリカの書き言葉を代表させた。LOB Corpus (The Lancaster-Oslo-Bergen Corpus) はBrown Corpusとテキストの抽出年代とカテゴリー区分、テキストの数と長さを同一にしているために、同年代のイギリス英語とアメリカ英語の比較ができる。

SEU Corpus (The Survey of English Usage)は書き言葉と話し言葉それぞれ100万語からなる電子化されていないコーパスで、後に1975年に話し言葉の部分がSvartvikによってLondon-Lund Corpusとして電子化された。教養のあるイギリス英語話者の話し言葉が収集されており、この設計は話し言葉のコーパスのモデル的存在となっている。SEU Corpusはのちにすべて電子化され、Quirk et al.(1973),(1985)の基礎データとなった。

[3] 近年出版された中では、Biber et al. (1999)が挙げられる。これは、日常会話、新聞、雑誌、広告、小説、学術論文など英語が使用されるあらゆる状況や条件によって生じる文法的差違を綿密に探るために、4,000万語を有する現代英語データベース(Longman Corpus Network)に基づく統計的実証を試み、現代英語の用法を検討したものである。コーパスから得られる知見をもとに、アメリカ英語とイギリス英語、書き言葉と話し言葉、レジスターの別による文法用法の差違に着目している。

[4] コンピュータ化されたアーカイブの中でもっともよく知られているものは、Oxford Text Archive (OTA) である。詳細は以下のURLを参照のこと：http://ota.ahds.ac.uk/index.html

[5] ICAME Collection of English Language CorporaのCD-ROM (ICAME-CD) についての情報は、http://www.hit.uib.no/icame/cdを参照のこと。ICAMEから、3,500ノルウェー・クローネ(約45,000円：1999年12月現在)で購入できる。ICAME-CDはISO 9660形式でフォーマットされているために、DOS/Windows、UNIX、Macintoshで読みとることができるが、添付の解析ソフトはWordSmith (Windows) やWordCruncher (MS-DOS)などでMacintosh用のものはないので、Macintoshで使用するためには、Concなど他の解析ソフトを用意する必要がある。Concの最新バージョンは1.80beta 3で、http://www.sil.org/computing/conc/beta/からダウンロードできる。なお、Mac OS 8.5以上でConcを使う場合は、同じページからType 12 Eliminatorという機能拡張書類を入手してインストールする必要がある。

ICAME-CDのジャケットには、本文中に示した18種類、1,400万語以上のコーパスが収録されていると記載されているが、ICAME-CDには次の2つのコーパスも収録されている。

International Corpus of English, East-African Component

International Corpus of English (ICE)は世界各地の英語を研究する目的のコーパスであ

り、中核コーパスとして18種類、各約100万語の地域変種がある。18の地域は、英語を母語したり、公用語として第2言語とする国と地域から選ばれ、International Corpus of English, East-African Component は、1990年代のケニアとタンザニアの書き言葉と話し言葉のコーパスである。話し言葉が50万語、'written as spoken' が10万語、書き言葉が70万語ある。

Innsbruck Computer-Archive of Machine-Readable English Texts (ICAMET)

中英語の散文(フルテキストが128)、1386年から1688年にかけて書かれた手紙254とその他の中英語と近代英語のテキストからなる。詳細は、インスブルック大学が発行するThe MANUAL OF ICAMET (ISBN 3-85124-163-0)を参照のこと。

6　新補遺の内容を本巻に盛り込んだOEDの第2版と同時に、CD-ROM版も発売されたが、これには新補遺が盛り込まれていなかった。この論文で使用しているのは、新補遺を盛り込み1993年に出版されたMacintosh版のCD-ROMである(Windows版の出版は1992年)。1999年10月にはOxford English Dictionary (Second Edition) on CD-ROM Version 2.0 として、検索ソフトを充実させたCD-ROM(Windows版)が出版された。現在のところCD-ROM version 2.0のMacintosh版は発売されていない。なお、2000年3月からオンラインでのOEDの利用が可能となった。詳細はhttp://www.oed.com/public/publications/online.htmを参照のこと。

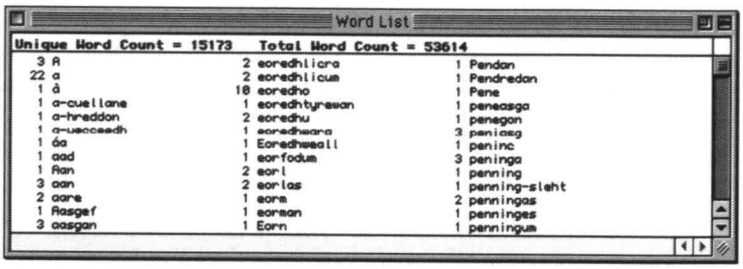
7　OED2-CDでは、調べたい時代、作者、タイトル、語・句のテキストなどの引用検索を本文の図1に示すウィンドウで行える。ただし、作家名や作品名は省略形で登録されているものが多く、登録形式は必ずしも首尾一貫していない。例えば、Shakespeare は Shakes. や Shak. でないと検索されない。書名の例では、Dickens の Dav. Copp. の引用として139例があるが、この他に、David Copperfield の書名での引用が1件ある。同様に、Austin の Pride & Prej. 139例の引用に対して、Pride & Prejudice として6例がある。このように、いくつかのヴァリエーションがあるので、図1のListボタンを押して、適切な検索キーを見つけるとよい。

8　新井(1998: 227)によるOEDの各期の引用数の一覧は、この事実を考慮していないように思われ、単なる合計値が示されている。

9　*a* 900というフィルターはその引用が900年以前に書かれたことを意味し、*c* 950はおよそ950年頃ということを意味する。*a*, *c* はそれぞれ ante, circa を指す。

10　引用数と引用文・句・節数が一致しない理由については注11を参照のこと。本文中に示した方法で書き出したテキストファイルには、<Th>a や <asg>eara などの記号列があり、これらはそれぞれ pa および ȝeara を表す。テキストファイル中に含まれる語数を数え上げるためには、TextUtils というソフトウェアを用いたが、実際に数える前には、<, >, ¥などの記号を削除しなければならない。削除する際には、NisusWriter や Jedit などの正規表現が利用できるソフトウェアを用いれば作業が容易である。TextUtils はテキスト内の語数、文字数

(句読点含む)、単語の平均的長さなどを計算する。処理を実行するためには、[File]メニューから[Open and Process Files]を選択して処理の対象となるテキストファイルを選択するだけである。



TextUtilsのアイコンと計算結果(OE 2期のファイルには53,614語)

なお、TextUtilsを用いれば、同じ語が繰り返し使われている例がカウントできる。例えば、OE 2期ではswa swaが28回、tha thaが3回生起していることが分かる。TextUtilsについての情報は次のURLで得られる。http://www.sil.org/ftp/software/mac/

[11]　OE 2期のテキストファイルの中にOE 2期以外の例も含まれていることに注意。例えば、Jute[2]の定義は、(b)の通りであるが、出力テキストファイルでは(a)のようになる。

a. ¥Comon hi of <th>rim folcum <edh>am strangestan Germanie, <th>æt is of Seaxum, of Angle, & of Geatum. Of Geata fruman syndon Cantware & Wihtsatan. (Cf. O. E. Chron. an. 449 Of Ald Seaxum, of Anglum, of Iotum. Of Iotum comon Cantwara, and Wihtwara..& <th>æt cyn on West Sexum <th>e man nu <asg>it hat Iutna cynn.)

b. *c*900 tr. *Bæda's Hist.* i. xv. (1890) 52 Comon hi of þrim folcum ðam strangestan Germanie, þæt [is] of Seaxum, of Angle, & of Geatum. Of Geata fruman syndon Cantware & Wihtsætan. (Cf. *O. E. Chron.* an. 449 Of Ald Seaxum, of Anglum, of Iotum. Of Iotum comon Cantwara, and Wihtwara..& þæt cyn on West Sexum þe man nu ȝit hæt Iutna cynn.)

また、引用(quotation)数と引用句・節(passage)数が一致しないのは、guiltの定義(c)はOE 2期の引用検索(quotation search)ではカウントされるものの、テキストファイルには出力されないからである。

c. intr. To commit an offence or trespass, to sin.

*c*825, *c*897, *c*1000 [see guilting *ppl. a.*].

なお、引用文・句・節数をカウントするためには、出力されたテキストファイルをエディタソフトに読み込み、1つの引用句・節が1行に収まるように横幅を設定し、行数をカウントればよい。

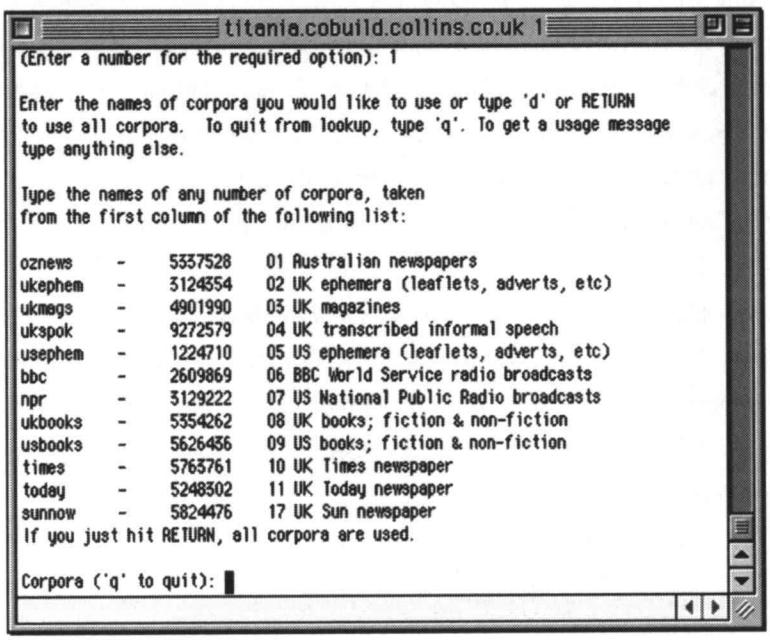
12　次のURLも参考になる。「コンピュータコーパスを利用した英語発達史研究」http://comh.soken.ac.jp/houkoku97/09204102/09204102.html

13　whether の変異形(variant)としてwapar, wheper, whethir, wether, whither, wheder, whedir を含む。

14　活用辞典-CD は『新編英和活用大辞典』(1985, 研究社)の CD-ROM 版である。初版が1939年、増補版が1958年に出版されたが、旧版の約20万例に対して、1985年の新版ではほぼ8割の用例が新しく記述され、総用例数は38万例である。試しに検索、計算してみると、38万例中、動詞 take を使った用例は283例、2,220語、get は545例、4,010語であった。活用辞典-CD をコーパスとして利用する場合に問題なのは、「…研究社が長い年月をかけて集積してきた現代英語の実例資料(corpus)を活用して、現代の要求に合致した新しい用例を大量に追加」(まえがき)と書かれているのみで、年代、引用箇所などの出典情報が全くないことである。

15　LOB Corpus, Brown Corpus からのデータは田島(1995)による。

16　COBUILD*direct* は有料のサービスであるが、WWW と telnet では "j" で始まる語をデモ版として検索できる。デモ版のログイン名とパスワードはともに cobdemo である。COBUILD*direct* の年間契約料は、500イギリス・ポンド(約88,000円：1999年12月現在)。



Telnet で COBUILD*direct* に接続し、コーパスを選択する。

上図は、telnet で COBUILD*direct* にログイン後に現れる画面で、ここで使用するコーパスを選択する。COBUILD*direct* で利用できるコーパスは、「英国の雑誌」(ukmags)、「米国のラジオ放送」(npr)など、全部で12種類あるが、全部を使用する場合には、何も入力せずにリター

ン・キーを押すだけでよい。一部のコーパスのみを検索の対象とする場合は、コーパスの略語(上図の左端に出ている記号列)を入力(例えば、BBCのラジオ放送のコーパスの場合はbbc)し、リターン・キーを押す。複数のコーパスを選択する場合は、times todayのようにコーパスの略語を並べて入力すればよい。各コーパスの後の数字は、それぞれのコーパスの語数を表す。この後の画面では、検索式を入力していく。表8における検索式のcan@ はcan, couldなどを一括で検索することを意味する。実際には、cannotと一語に綴らず、表9および10の調査対象コーパスのうち、can notと分けて書くのはわずか1例が検出されただけだった。COBUILD*direct*の詳細については、COBUILD*direct* User Guideのページを参照のこと。
http://www.cobuild.collins.co.uk/cdguide/svenguide.html

参考文献

新井洋一 1988.「英語辞書の英語学研究への応用」、齊藤俊雄、中村純作、赤野一郎(編)『英純作、赤野一郎(編)『英語コーパス言語学――基礎と実践――』、211-232、東京、研究社出版.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*, Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan 1999. *Grammar of spoken and written English*, London: Longman.

Butler, Chris. 1998. Using computers to study texts. In *Projects in linguistics: A practical guide to researching language*, ed. Alison Wray, Kate Trott, and Aileen Bloomer, 213-223. London: Arnold.

Crystal, David. 1997[4]. *A dictionary of linguistics and phonetics*, Oxford: Blackwell.

Edwards, Jane A. 1993. Survey of electronic corpora and related resources for language researchers. In *Talking data: Transcription and coding in discourse research*, ed. Jane A. Edwards and Martin D. Lampert, 263-310. Hillsdale, N. J.: Lawrence Erlbaum Associates.

家入葉子 1997.「通時的コーパス使用によるor notの研究」『コンピュータコーパスを利用した英語発達史研究』(平成8年度科学研究費補助金(08207106)研究成果報告)

Lawler, J. and Helen Arista Dry (eds.) 1998. *Using computers in linguistics: A practical guide*, London: Routledge.

Leech, Geoffrey. 1991. The state of the art in corpus linguistics. In *English corpus linguistics*, ed. Karin Aijmer and Bengt Altenberg, 8-29. London: Longman.

McEnery, Tony and Andrew Wilson. 1996. *Corpus linguistics*, Edinburgh: Edinburgh University Press.

中郷　慶 1999.「コーパス言語学の現状と課題」『愛知淑徳大学研究紀要』第38号、187-202.

Nakago, Kay 2000. Some notes on using electronic texts in the study of language. *Linguistics and philology* 19 (Synchronic and diachronic studies on language: a Festschrift for Dr. Hirozo Nakano): 491-505.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik 1972. *A grammar of contemporary English*, London: Longman.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik 1985. *A comprehensive grammar of the English language*, London: Longman.

Renouf, Antoinette. 1987. Corpus development at Birmingham University. In *Looking up: An account of the COBUILD project in lexical computing.* ed. John Sinclair, 1-40. London: Collins ELT.

Stubbs, Michael. 1997. Whorf's children: Critical comments on critical discourse analysis (CDA). In *Evolving models of language: Papers from the annual meeting of the British association for applied linguistics* 1996, 100-116. Clevedon: Multilingual Matters.

田島松二(編) 1995.『コンピュータ・コーパス利用による現代英米語法研究』東京、開文社出版.