椙山女学園大学研究論集　第24号（第１部）1993

コーパス言語学の現在

－SEUを中心に－

深　谷　輝　彦

Corpus Linguistics in English Language Research

Teruhiko FUKAYA

１．序　　論

英語学者の仕事を，英語母国語話者が英語について持っている知識の解明と定義する。そしてその知識を記述する仕事に取りかかるとき，英語学者にこの上ない驚きと感動をもたらすのは英語知識の多種多様性である。英語の基本的語順が主語＋動詞＋目的語であるという直観的に自明な知識がある一方で，英語の一語一語について英語話者が頭の辞書に蓄えている緻密な知識もある。

後者の知識を例示するために，コロケーション（連語関係）及び語彙の頻度についてSinclair（1987: xvii）からの英語事実を取り上げる。形容詞と名詞が結びつくとき，意味的整合性に加えて語と語の間のコロケーションが影響する。（１）の表は，縦の形容詞と横の名詞の間の結合可能性を○（＝OK）と×（＝NO）で示している。

（１）

	luck	facts	evidence
hard	○	○	○
strong	×	×	○
tough	○	×	×
sad	×	○	×

ここで注目すべきは，hardとstrongは意味が非常に近いにもかかわらず，前者は表中のすべての名詞と結合可能であるのに後者のstrongはevidenceとのみ共起するという制約である。（１）の分布には意味的適格性条件に語同士の相性条件が関与していることがわかる。英語の母語話者は個々の語自体の語彙情報と併せて連語関係の知識を活用できる。

語彙の頻度についてわれわれは何らかの形で直観を持っていると意識できても，それを明示的に述べることはできない。Sinclairらは約二千万語規模のCOBUILD Databaseを材料にしながら次の頻度表を作成している。

（２）10万語につき１回：secretly, jet, core, agony, preparation, privileged
　　１万語につき１回：speak, green, maybe, quickly, sorry

1000語につき１回：two, over, get, because, see, don't
100語につき１回：in, that

さらに頻度が高い語としては the, and, of が挙げられている。この頻度表によれば，言語使用にあたり母語話者は辞書にある語を平均的に利用しているのではなく，特定の語に繰り返しアクセスしていることを読みとれる。これと同じ策略を二語三語の句レベルでも適用していると Altenberg（1991：77）が主張する。彼は London–Lund Corpus の中で繰り返し起こっている語の組み合わせ68000タイプを抽出したうえでその分析を試みている。

（３）a. The most frequent word combinations:
(e. g.) you know/[m], [m]/I think/sort of/you see
b. The most frequent word combinations of more than three words:
(e. g.) thank you very much/are going to be/I think it was/to be able to /I don't know what

Altenberg の主張が正しいとすると，脳の辞書には話しことば用に上記のような句表現が在庫しており，我々はそれを繰り返し取り出していることになる。さらに検証を必要とするにしても，なかなかおもしろい仮説である。

英語話者の知識についてもう一例観察しておきたい。英語文章構成法では伝統的に and や but で文を始めるべきでないとされる。これは果たして英語作文の実態を反映した規則なのだろうか，あるいは根拠のない規範的規則にすぎないのだろうか。イギリス英語の書きことば百万語から成り立つ LOB Corpus で文頭の置かれた and と but を検索してやると，興味深い言語事実が浮かびあがる。両語ともに大部分のテキストで文頭に起こる例を観察できる点では共通しているが，その頻度を比べると but が1547例，and が875例とかなり大きな相違点をみいだせる。言い換えれば，and よりは but のほうが文頭に配置することに抵抗を感じないと言える。テキストの種類ごとにその生起回数を調べると，文頭の and/but は科学小説に最も多く現れ，逆に政府刊行物や企業文書では二千語につき一度起こるか起こらないという低い頻度を示している。(Davidson 1992: 21)

以上の記述が英語を母国語とする人々の言語知識表示のなかでどういう位置を占めるかは別にしても，そのアプローチは従来の英語学研究にない全く新しい分析を提供する。これまで言語運用だからというだけで正面から研究されてこなかった話しことば書きことばの実際を，コンピューターを武器に解明しようというコーパス言語学が現在多くの言語学者の注目を集めている。生成文法がもっぱら直観に依存するのにたいし，コーパス言語学者の中には直観を疑問視するひとも多く自然な発話や文章こそ研究対象とするべきだという。本稿では，コーパス言語学の基本的性格を述べた上でその妥当性を検討するところから議論を始める。次に，SEU（Survey of English Usage, University College London）で現在構築過程にある国際英語コーパス（The International Corpus of English）を詳細に紹介し，コーパスの目的，デザイン，構築過程を例証したい。第三番目にはコーパスを活用した研究事例をみる目的で Fukaya（1992）を要約する。そして最後に，コーパス言語学の最新動向をその問題点と共に考える。

2．コーパス言語学の基本的性格とその妥当性

議論を始めるために，コーパス言語学を便宜的にコーパスに基づく言語研究と定義しておく。では，コーパスとは何か。大量の研究用言語資料体である。但し，この答えには付帯条件を二つ科す必要がある。第一に，コーパスとは何の計画もなく行き当たりばったりで集めた例文の集合では決してない。コーパス構築には必ず目的があり，それに沿ってコーパス設計が行われ，研究対象の言語またはテキストを適格に代表すべきである。最初の本格的コーパスとされる Survey of English Usage は，イギリス英語のうち教養ある人々の文法的レパートリーを記述目標としたので，ラブレターから大学の講義やパーティの会話まで幅広い言語使用域の英語が含まれている。このコーパス中の英語使用者も多彩で，弁護士，ジャーナリスト，教員，技術者，医師などが登場する。Svartvik & Quirk（1980：9）によると，Survey Corpus 作成者の間で(i)特定言語の文法的レパートリーをすべて駆使できる話者はいない，(ii)内省だけに頼っていては英語の文法的文体的体系全体を記述しきれない，という合意ができていた。

第二番目の条件とは，コーパスといえばコンピューターによる解続が可能なテキスト集合体を意味する点である。この条件があるからこそ，コーパス言語学者は二重の恩恵に浴することができる。第一のそれは，大量の言語データの保存，検索，さらに分類を短時間で処理する能力である。後述する tag 付与や parsing と呼ばれる統語分析も，コーパスが電算機に乗っているから実行可能だといえる。第二番目には，言語データの移動可能性が大きく高められた点を強調したい。VDU やプリンターを通した出力に加えてコンパクトディスク等によるデータの移動を考えると，コーパスは世界中の言語学者に強力かつ潜在力のある研究材料を提供するものと思われる。(Leech（1991：76）参照)

コーパス言語学を生成文法及び伝統文法と比較しながら，その理論的基盤の妥当性に議論の焦点を向けることとする。比較対象としてまず生成文法を選ぶと，(4)のような対照を描き出すことができる。

(4)	生成文法	コーパス言語学
分析	直観	直観とコーパス
言語能力と運用	能力	能力と運用
母語話者	理想的話者／聴者	不完全で多様な母語話者
範疇	決定論的	非決定論的

(4)について若干の補足説明をしておく。生成文法の Chomsky とは異なり，コーパス研究者たちは理想化された母語話者を設定せず，その知識の可変性を共時的にも通時的にも認める。コーパス構築にあたり言語使用域に関心を払うのは，この言語知識の非画一性ゆえであることは言うまでもない。

言語記述を進める際にその理論がたてる範疇を，Lakoff のプロトタイプ（prototypes），Quirk の漸次的移行（gradience），Coats のファジー範疇（fuzzy categories）として捉える立場がコーパス言語学の特徴である。英語の接続詞を等位接続詞と従位接続詞という二つの範疇に分類するとき，and が前者の，if が後者に属する典型的接続詞だといえる。しか

しながら，for は従位接続詞にやや傾きながらも両範疇の中間位置にあるとする。(Quirk *et al.* 1985: 90) Leech は，このような柔軟な範疇を設ける言語理論からすると，直観は言語能力を忠実に反映しておらずむしろその本来の姿をゆがめていると考える。別の言い方をすると，直観だけに頼るとプロトタイプ的データにのみ目を向けてしまうか，逆に言語の実際を無視した反例ばかりあげることになりかねない。そういう危険を避けるために実際の言語運用から生まれたコーパスを丹念に観察することを，コーパス言語学は強く勧める。そしてコーパスから得られた事実をそれぞれの理論的枠組みで説明しようとするときも，決してＡかＢかと二者択一にせずＡとＢの中間段階にある現象を積極的に認知していこうとする。

では伝統文法とコーパス言語学はどう比較できるんだろうか。コーパス言語学の優位性を求めることができるのは何にもまして話しことばテキストの利用可能性であろう。二十世紀後半に飛躍的に発達した録音技術のおかげで，話しことばを録音しその音声的特徴を詳細に記述することが困難なくできる。しかも話しことばが我々の言語生活のなかで独占的に機能していることを再認識するとき，言語記述に話しことば分析が不可欠である。既述の Survey Corpus は百万語の総数のうちその半分が話しことばに割り当てられている。話しことばの重要性を英語の照応現象で例証できる。someone のような不定代名詞を前方照応で受ける代名詞に何を選択するかしばしば議論されるところである。伝統文法学者に尋ねると，男性ならば he，女性ならば she，男女不明または両方を指すならば he と答えるだろう。問題は最後の代名詞 he で，女性差別表現としてやり玉にあがることもしばしばである。そこで話しことばでの処理方法をコーパスに探ると，they/their/them による照応関係が自然な形で成立している。不定代名詞の単数複数よりも男女両性を中立的に指示する they 系列語の使用が現代語法である点は，話しことばを守備範囲にいれたとき明らかにみてとれる。

これに関連して，例えば伝統文法学者 Jespersen は印刷されたことば特にフィクションからの引用を多く採用する。しかしコーパス言語学者が Jespersen に面会できていたら，きっともっと幅広い言語使用域やジャンルのテキストを考慮すべきだと助言していたにちがいない。(Svartvik & Quirk 1980: 9) 伝統文法学者に好都合な引用例文だけに目を向けていると，バランス感覚を欠いた文法記述に陥ってしまう。英語ならば英語の語彙文法構造全体を浮かび上がらせるためにはどうしてもコーパス観察を欠くことができない。「英語の全体像」という指向を bet を例に説明しよう。コーパスで bet の用例を集めると最も頻出するのはその前に I が先行する表現である。この場合，I と bet の間にはよく法助動詞が共起してくる。(例：I'll bet, I'm willing to bet) bet それ自体の意味は that 以下で述べる陳述に対する確信を表すことである。しかしコーパスは bet の別の典型的用法をも忘れることなく映し出す。主語の位置に you をとる bet である。I を取るときとは異なり，that 節が後続せず I と bet に挟まれる助動詞も起こらない。you bet だけで表現が完結しているうえに，意味は yes の強調を表す。伝統文法もコーパス言語学もともにテキストを基本単位とするという共通項を持ちながらも，言語全体の構造と機能をより正確により綿密に説明しようと試みるときコーパス言語学に優位性があるとせざるをえない。

これまでコーパスが万能薬であるかのように議論を展開してけれども，現在広く使われているコーパスは長所もあるが短所もあって完全ではない。アメリカ英語を代表する

Brown Corpus は1961年時点の書きことば百万語から成るが，それと対応させて構築されたのがイギリス英語の LOB Corpus である。両コーパスともに話しことばを含まず，また語彙について調査を行うためには百万語では規模が小さい。幸いなことに，コーパスの大きさはコンピューター技術の発展と共に拡大傾向にあり，1990年代に入ってから作られている The Bank of English や The British National Corpus は億の単位に突入している。体系言語学の Halliday の考えでは，体系文法記述にとっても二千万ぐらいの文，語数にして二億語サイズのコーパスが望ましいという。その一方で SEU の Greenbaum らは，英語変種間の文法比較の目的には百万語のコーパスでよしとする。

もう一点注目しておきたいのは，書きことばの入力は OCR（光学式文字認識）によりかなりの省力化がすすんだが，他方話しことばについては依然端末のキーボードから録音テープを聞きながら入力する方式が継続しており，膨大な時間，労力，費用がかかる。以上の様々な課題を考え合わせると，コーパス構築にあたりその目的が最優先されなければならないという主張を認めても，それに伴う時間，経済的裏付け，さらに人材の確保という非言語的条件をどこまで満たせるかが影響することも否めない。以下では，英語変種の比較という目標を掲げるまさに国際的なプロジェクト，国際英語コーパスを詳述する。

3．国際英語コーパス

現在ロンドンの SEU を中心に世界二十三か国の大学研究機関の参画を得ながら国際英語コーパス（The International Corpus of English，略称 ICE）プロジェクトが進行中である。参加国としては，英国，米国，豪州はいうまでもなくアジア・アフリカ諸国で英語を第一または第二公用語とする国が網羅的に含まれている。その目的は，第一に世界にあまねく広がりをみせる英語変種を実証的に比較するためである。派生的には国際語としての英語を確立するための基礎研究の役割を果たすものと期待されている。さらに，各国がそれぞれコーパス構築を行う結果，各変種の英語研究がおおいに前進するだろう。音声，文法，語彙，談話研究の基礎資料になるだけでなく，英語教育，英語政策，自然言語処理での応用研究にも有効であろう。

このプロジェクトが生まれた背景にはもうすこし複雑な問題意識があるようである。英語の規範としてイギリス英語あるいはアメリカ英語をそのほかの英語より上位におくという傾向が弱まり，むしろそれぞれの地域英語がいわば独立宣言を始めたというのが現状である。このまま進むと，正しい英語はそれぞれの国または地域で決定することに当然なり英語変種間のコミュニケーションが成立しない状況さえ想定できる。一例を挙げると，インド英語話者のなかに milk bottle でミルクの入っていない瓶とミルク一瓶の両方の意味を担わせるひとがいる。Greenbaum（1990：82）は，曖昧さを避けるために milk bottle と a bottle of milk を区別したほうがよく，そのほうが他の英語変種の用法とも合致すると，力説している。様々な英語間の共通点，相違点をコーパスに基づき把握し，国際共通語たる英語の地位を保持したいという狙いが ICE プロジェクトに込められている。Greenbaum（1991：4）のことばを借りると，このコーパス研究が英語の過剰な多様化を防止する言語政策につながり，すくなくとも書きことばの面だけでも英語の国際性を維持したい，という発言にいたる。

以上の目的に沿いながら，国際英語コーパスは次のようなコーパス設計案を下敷きにしながらまさに今現在その構築が日々すすめられている。第一にコーパスの規模は中核コーパス（core corpus）およそ百万語とする。各国独自の企画で，中核コーパスの加えて拡大コーパス（expanded corpus），特殊コーパス（specialized corpus），非標準（英語）コーパス（nonstandard corpus），モニターコーパス（monitor corpus）のいずれかが追加的につくられる。第二番目の特徴は，コーパスは話しことば三百，書きことば二百，合わせて五百のテキスト（各テキスト二千語）から成り立つ点である。コーパスをテキスト範疇で区分したときに各範疇が約二万語になるよう配慮されている。話しことばは話者が一人か複数かで，書きことばは印刷されたか否かで分類される。テレビやラジオ番組の台本化されている英語は話しことばと書きことばの両面を持つとされる。更に詳しいテキスト構成については Greenbaum（1991：5）を参照。第三番目には，入力するテキストは1990年から1993年に話され書かれた英語に限定する。次に，ICE は各国の標準英語または教育を受けた人が使う英語を記述対象とする。よってコーパスに登場するのは，最も広い意味での英語の professional というグループに分類できる話者並びに筆者である。この原則は Survey Corpus でも適用されたものであり，ICE はそれを踏襲したことになる。第五番目の特徴に挙げられる点は，教会での説教，詩，法律文書，ファクスメッセージ，コンピューター通信，広告等で生産される英語は除外するという合意である。各テキスト収集にあたり，性別，年齢，地域，職業，学歴などという社会言語学的変異要因を念頭におくべきだとされる。例えば話者の性別について，男性同士，女性同士，男女の間というように変化をつける。最後に，ICE では四つの特殊コーパスが計画されている。すなわち，EC 諸国語から翻訳された英語，国籍の異なる話者の間で用いられる英語，上級英語学習者の英作文，EC の公式文書にある英語である。

ICE に参加しているコーパス言語学者は，どのような過程を経ながらコーパスを作り上げるのだろうか。第一段階はテキスト収集である。話しことばはテキストについては録音作業を行う。テキストに関連する情報（例えば，話しことばでは話者，日時，場所）を詳しく記録する。すでに述べたように，書きことばは光学式文字読み取り装置の発達に伴い，その入力労力が軽減されてきている。しかしすべての書きことばジャンルが読み取り可能という訳にはいかない。学生の試験答案などは手書きのため，キーパンチ方式で入力せざるをえない。話しことばは入力者が録音テープを繰り返し聴きながらキーボードを打つのだが，この仕事が実に骨の折れるまた大量の時間をとる部分である。特に，入力者はテキスト発生場面に居合わせずもっぱらテープ音声に依存せざるをえないので，英語母語話者でさえその理解にはかなりの困難をともなう。Halliday が説くように，我々が意味を語彙文法レベルさらには音声レベルで実現するとき，そこにはイデオロギー，ジャンル，言語使用域，文法的比喩という色々な意味の糸が絡み合う。テキスト現場に参加していた聴者はそれらの意味に直接ふれることができる。けれども，コーパス入力者は音声モードと自らの言語知識や経験のみしか利用できないという不利な条件下にいつもおかれる。話しことばは現在でも入力上なかなかてごわい存在である。

マークアップの第三段階では，原テキストがもつ情報をできる限りコーパスに反映させるために，記号付与が入力者によって行われる。書きことばの字体，パラグラフの切れ目や引用，標準英語からの逸脱などが記号化され挿入される。話しことばの典型である対話

に対して，話者の同定，ポーズ，オーバーラッピング，笑い声という要素がテキストに追加される。Greenbaum（1991：7）より具体例を引用する。

（5）
```
<$A\> <#\>Could you explain why it's not until the twelfth of January
nineteen eighty eight<,> that you are signing forms<,> to<,> various
consultants saying that you want to borrow money<,,>

<$B\> <#\>I should've<,> been given the chance to look at the forms<,,>

<$A\> <#\>Take as much time as you want to look at them<,>

<$B\> <#\>The form isn't a form<,> in relation<,> to one particular<,,>
financier <#\>It's a general application<,,>

<$A\> <#\><O_>three or four words<O/><,,>

<$B\> <#\>Yes sir<,>
```

Key:

```
<$A\>, <$B\>:  Speaker identification
       <#\> :  Sentence boundary
       <,>  :  Short pause
       <,,> :  Longer pause
       <?>  :  Uncertain transcription
       <O>  :  Untranscribed text
       <}>  :  Editorial emendation (for tagging and parsing purposes)
       <->  :  Editorial deletion (for tagging and parsing purposes)
       <=>  :  Self-correction or repetition
       <.>  :  Incomplete word
```

マークアップに続く過程はtag付与と呼ばれ，テキスト中の各語に品詞を与える。tag付与者の実際の仕事は，自動tag付与プログラムが選んだtagの是非を判定する作業である。複数の可能性が画面に提示されたときには一つ妥当なtagを選んでやる。tagの付いた文例を以下に示す。

（6）
```
hospitality_NN is_BEZ an_AT excellent_JJ virtue_NN ,_,
but_CC  not_XNOT  when_WRB  the_ATI  guests_NNS  have_HV
to_TO sleep_VB in_IN rows_NNS in_IN the_ATI cellar_NN
!_!
```

(Leech & Fligelstone 1992:125)

各語に品詞が与えられると，次の段階はその情報を基に統語分析を行うparsingである。parsingを終了すると，テキストは例えば(7)のような表示を受ける。

（7）
```
[S[NP[D The_ART D][N house_NL1 N]NP][VP[V has_VH3 V]
[NP[A blue_ADJ A][N walls_NN2 N]NP]VP]S]
```

(Leech & Fligelstone 1992:125)

ここで注意しておきたいのは，tag付与とparsingはそれぞれ独立した作業ではない点である。もしtag付与がより高いデリカシーで行われれば，それに応じてparsingは容易になると思われる。しかし，現実的にはtag付与者に過度に複雑な文法分析を期待すること

はできないから，二つの仕事のあいだのどこかでバランスをとることになる。

国際英語コーパスのデザインとその構築過程を概略的にみてきた。当然のことながら次にわく疑問は，コーパスを活用するとどういう新しい研究ができるのだろうか，ということである。この疑問に答えるために，Fukaya（1992）の要点を以下で述べコーパス言語学研究例としたい。コーパス言語学全体の動向を押さえるには，Johansson & Stenström (1991), Aijmer & Altenberg (1991), Svartvik (1990), 赤野，吉村＆藤本（1991ａ，ｂ）が便利である。

4．コーパス言語学研究例

Fukaya（1992）の出発点は，（8）のペアに対してこれまで与えられてきた説明に不満を抱くところからである。

（8） a. On what did you base your prediction ?
b. What did you base your prediction on ?

生成文法も含めて従来の文法書は，（8 a）が形式ばった，（8 b）が形式ばらない表現だという記述で終わってしまう。記述文法 Greenbaum & Quirk（1990）をみると，形式ばった文体では，前置詞がその目的語に先行し（8 a）の語順をとる。もしそうでなければ，目的語のみが文頭に置かれ前置詞は文末に据え置かれると言う。生成文法の枠組みで（8）を扱う Hornstein & Weinberg（1981）は，（8 b）の動詞と前置詞（on what）が動詞－補部という強い関係にあることを指摘してくれるが，（8）の交替に関与する他の要因には言及がない。

しかしながら，このような説明が前置詞文末残留現象を単純化しすぎていることを示す証拠をいろいろコーパスから拾うことができる。（9）にある二文を比べると，（9 b）の方が明らかに自然な文である。

（9） a. You should restrict yourself to words with which you are familiar and you can use confidently.
b. You should restrict yourself to words you are familiar with and can use confidently.

(Quirk *et al.* 1985: 1253)

（9 b）をよしとし（9 a）を不自然な文にしている要因は等位構造における平行性である。（9 b）ではそれが動詞句でみごとに成立しているが，（9 a）では前置詞が関係詞のすぐ前の位置にあるため平行性が崩れている。（9）の観察からできる推測は，（8）の交替には単に formal か否かという文体上の要因に加えて他にも様々な文法的機能的要因が影響しているのではないか，ということがある。

どのような要因が（8 b）を選ばせるのかという問題意識をもちならが，Survey Corpus から前置詞が文末に据え置かれている文を体系的に抽出しそれを機能主義的立場から分析し得られた結果の一部を紹介したいと思う。文法性のみを問題にするとき，wh 疑問文も

関係節もともに前置詞が目的語と共にある場合，文末に残留する場合を許す。

(10) a. At which house did you leave the car ?
b. Which house did you leave the car at ?
(11) a. The old house about which I was telling you is empty.
b. The old house (which) I was telling you about is empty.

それにもかかわらず，コーパス調査では(10b)タイプの文が圧倒的多数派で(10a)タイプはその頻度が極めて低い。対照的に，(11)のような関係節文では(a)型と(b)型の間に頻度数でおおきな差はない。言い換えると，関係節は前置詞の可能な両位置を偏向無く利用しているのに，wh疑問文になると前置詞残留が強く好まれる。

この頻度格差は問題の構文が果たす機能構造に帰することができよう。関係節の主要な機能は一言でいうと指示物の特定ということである。関係節により聞き手は先行詞が何を指示するのか特定し易くなる。そして，前置詞が関係代名詞の直前に起これば，その先行詞が後続する命題のなかでどういう文法的役割を演じるのか決定するのを援助する。

(12) a. ??That's the place which I'd like to spend as much of my retirement as possible in.
b. ?That's the place which I'd like to spend my retirement in.
c. That's the place to spend retirement in.

(Crystal 1984: 62)

(12a)では in と the place の結びつきをつけるのがその間の距離ゆえ困難なため，前置詞を which の前に配置することが強く望まれる。(12c)ではその関係づけが容易なため，前置詞残留が許可されている。結局，関係節の場合，前置詞が関係代名詞に先行する機能的動機づけを持っている。これが wh 疑問文になると機能的事情が異なる。wh 疑問文は通常話し手と聞き手が wh 語として実現している語を除いた命題を共有しているときこそ用いられる。その wh 語が疑問の焦点を構成し，他方残りの部分は後景化されているといってよい。従ってここでは関係節でみられたような結びつけの機能は作用せず，(10a)の語順には機能的動機づけがないという結論である。

コーパス観察の有効性を立証する例をもう一つ挙げたい。関係節が修飾する先行詞の文法機能と前置詞の生起する位置の相関を探ると，(13a)より(13b)タイプの関係節が多いことに気づく。(14)には Survey Corpus から例文を列挙する。

(13) a. There are few things about which a child is more curious than what his parents are up to while he is asleep.
b. There are few things a child is more curious about than what his parents are up to while he is asleep.
(14) a. There was a barrage they didn't tell us about and... (W−16−8: 353)
b. "There's a harmonium in there that I'm very fond of," said Emma.

(W−16−2: 15)

c. there was a particular man there she wanted to work with and...
(S−1−3: 346)

d. I should ask him if there are any seminars you ought to go to (S−1−5: 29)

There 構文のあとに関係節が続くと，前置詞＋関係代名詞の語順が 1 に対して前置詞残留が 4 位の頻度で実によく起こる。Deane（1991）は，前置詞残留の機能的条件として前置詞の目的語が話題であると同時に焦点でもなければならない，と論じている。(14) の事実と Deane が提案する条件はどこで接点をみいだすのだろうか。

(14c) を (15) と再録しなおし，それを材料に問題の there 構文＋関係節という文型が前置詞残留に強い機能的動機づけを与えているという仮説を説明したい。

(15) there was a particular man there she wanted to work with

前半部の there 構文は基本的に a particular man を談話舞台に乗せる機能を果たし，その結果 a particular man は話題という地位を情報の流れのなかで得る。後半部分に注目すると，a particular man は関係節末の（前置詞）with の目的語に対応するのだから重要な情報ということができる。まとめると，a particular man は前半部から話題，後半部から焦点という機能的性格づけを与えられる。これは Deane 条件をみごとに満足させる。よって，(15) が前置詞を後に据え置くのは機能的に見て極めて自然な文法措置だといえる。

さらに詳しい議論は Fukaya（1992）に譲ることにしても，英語研究コーパスの有効な活用法を例示することができたと思う。それと同時に，従来の文法研究が既製の範疇に言語事実を押し込めようとするばかりで，重要なデータを多く見逃していたことも明らかにしてきた。直観だけに頼っていると，人間で言えば骨だけ研究したことにしかならず，人間を人間らしく見せている部分を切り捨ててしまっていることになる。英語を英語らしく見せている部分にもっと注目し，英語のデリカシーを大切にした言語研究がいま求められている。

5．コーパス言語学の動向

コーパス言語学の今後の動向について以下では二つの流れを指摘したい。第一はコーパスサイズの驚異的拡大である。1960年代は百万語，1970年代に一千万語を超え，今1990年代はいよいよ億の単位の語数が実現しようとしている。ではなぜこれほど語数を集めるのか。多ければ多いほど必要な詳しい情報をコーパスから引き出すことが可能になるからである。

文法面と語彙面から億単位サイズのコーパスがどれ程の力を発揮するのか具体例をあげておきたい。regret の後に to 不定詞をとる環境で，to に続く動詞はどういう種類かという問題を調べる。二千万語レベルのコーパスでは regret to Verb という動詞型の文が十二例のみで，say/learn/inform の三種類しか検索できない。ところが，一億語コーパスでみると，登場する動詞が discover/tell/announce/confess/report と拡大する。そうなれば，regret to のあとには伝達動詞が起こると自信をもって記述できる。語彙面ではサイズ拡大

の効果がてきめんに出る。競馬用語で nap を使うと 'to name (a horse) as likely to win a race' の意味である。二千万語のコーパスではこの意味の nap は現れない。しかし，一億語コーパスには登場し，面白いことにすべて受動態の形を取っている。やはり，億サイズのコーパスは，いままでのそれとは一味も二味も違う英語像を描いてくれるものと思われる。

二番目の流れとして英語変種研究の発展がおおいにのぞめそうである。そして，この分野こそコーパス研究が英語学，言語学に貢献できるところでもある。なかでも国際英語コーパスに期待が集まっている。既に述べたように，Survey Corpus の伝統に習い，この新コーパスは英語全体を代表するようにテキストの構成に十分バランスを考慮してある。また，国際英語コーパスと通時的英語研究用のヘルシンキコーパスを組みあわせた研究は新しい側面をコーパス言語学に持ち込みそうである。

Leech が指摘するように，言語の各レベルごとの情報を付与されたコーパスを体系的に組織する必要性が高まっている。現在のところ，音声特徴や品詞のついたコーパスは利用可能だが，コーパスの統語分析は進行中である。そのほかの音節，形態，意味，談話レベルとなると全く未開拓の領域である。

以上で述べてきたように，まだ今後の発展を待たねばならない部分を多くかかえつつもコーパス言語学は新しい研究方法と領域を言語学にもたらしつつある。我々のように英語を母語としない英語学研究者にとっては実に魅力ある分野である。

注

* 本稿は椙山女学園大学の海外研修制度のもとで1991年4月より1992年3月までイギリスの Survey of English Usage, University College London で行った研究の一部であり，椙山女学園ならびに私学振興財団の助成を得てなされたものである。

References

Aijmer, K. and B. Altenberg (1991) *English Corpus Linguistics,* London: Longman.

赤野一郎，吉村由佳，藤本和子（1991a）'Corpus Linguistics の現在の動向と問題点（1）コーパスとその構築' 京都外国語大学 SELL, pp. 1–45.

赤野一郎，吉村由佳，藤本和子（1991b）'Corpus Linguistics の現在の動向と問題点（2）tag 付与，各種プロジェクト，コーパスの活用事例' 京都外国語大学研究論叢 XXXVII, pp. 1–23.

Altenberg, B. (1991) 'The London–Lund Corpus of Spoken English: Research and Applications', in the proceedings of the 7th annual conference of the UW centre for the new OED and text research.

Coats, J. (1983) *The Semantics of the Modal Auxiliaries,* London: Croom Helm.

Crystal, D. (1984) *Who Cares about English Usage,* Harmondsworth: Penguin Books.

Davidson, L. (1992) 'Using Large Text Data–banks on Computers', in P. Roach (ed) *Computing in Linguistics & Phonetics,* London: Academic Press, pp. 5–34.

Deane (1991) 'Limits to Attention: A Cognitive Theory of Island Phenomena', *Cognitive Linguistics* 2, pp. 1–63.

Fukaya, T. (1992) 'Motivations for Deferred Prepositions', MS.

Greenbaum, S. (1990) 'Standard English and the International Corpus of English', *World Englishes* 9, pp. 79–83.

Greenbaum, S. (1991) 'ICE: the International Corpus of English', *English Today* 7. 4, pp. 3–7.

Greenbaum, S and R. Quirk (1990) *A Student's Grammar of the English Language,* London: Longman.

Hornstein, N. and A. Weinberg (1981) 'Case Theory and Preposition Stranding', *Linguistic Inquiry* 12, pp. 55–91.

Johansson, S and A–B. Stenstrom (1991) *English Computer Corpora,* Berlin: Mouton de Gruyter.

Lakoff, G. (1987) *Women, Fire and Dangerous Things,* Chicago: The University of Chicago Press.

Leech, G. (1991) 'Corpora', in K. Malmkjær (ed) *The Linguistics Encyclopedia,* London: Routledge, pp. 73–80.

Leech and Fligelstone (1992) 'Computers and Corpup Analysis', in C. Butler (ed) *Computers and Written Texts,* Oxford: Blackwell, pp. 115–140.

Quirk, R. (1968) 'Descriptive Statement and Serial Relationship', in R. Quirk *Essays on the English Laguage Medieval and Modern,* Bloomington: Indiana University Press, pp. 167–183.

Quirk, R, S. Greenbaum, G. Leech and J. Svartvik (1985) *A Comprehensive Grammar of the English Language,* London: Longman.

Sinclair, J. (ed) (1987) *Collins COBUILD English Language Dictionary,* London: Collins.

Svartvik, J. (1990) *The London–Lund Corpus of Spoken English,* Lund: Lund University Press.

Svartvik, J. and R. Quirk (1880) *A Corpus of English Conversatin,* Lund: Lund University Press.