コーパスを利用した語法研究

コーパスを利用した語法研究

言語体系研究部第１研究室　山崎誠・鈴木美都代

要旨：

１．コーパスとは元来，言語分析のために集められた言語資料を意味するが，近年の日本語研究においては，とくに，コンピュータで取り扱うことを前提にした大規模な電子化データをさすようになってきた。

２．現代語研究にとっては，コーパスは内省とともに有力な研究手段の一つとなる可能性を秘めている。今後コーパスを利用した新しい研究方法の開発が期待されている。

３．コーパス利用と密接に関連のある文法研究において，どのように例文が採られているかを調査した。その結果，実際の論文では，出典のある例文は約３割であること，また，出典の内訳では，文芸作品を中心としたフィクションが多いことがわかった。

４．個々の論文においては，作例中心タイプのものと，実例中心タイプのものとの二つの類型があるようである。

５．ケーススタディとして，連体修飾句の中の「が」「の」の使い分けについて，書名データと雑誌データの二つを調査し，おおまかな傾向をみた。

６．どちらのデータにおいても，主語（あるいは直接目的語）が「の」で表される場合，述語との間にはほとんど他の語句が介在していないことが確かめられた。

７．どちらのデータにおいても，述語の品詞が形容詞・形容動詞の場合，主語（あるいは直接目的語）が「の」で表されやすいことがわかった。これは，データにおいて，形容詞・形容動詞が述語に来る場合，主語（あるいは直接目的語）との間に他の語句が介在する割合が低いことと関連しているものと思われる。

キーワード：コーパス　例文　ガノ可変　書名　「現代雑誌九十種の用語用字調査」

１．はじめに

近年，日本語研究において，コーパスに対する関心がたかまっているが，実際に大規模なコーパスを使いこなすような状況には至っていない。これは，安心して使えるデータが整備されていないことが大きな要因だろうが，データ利用の価値に対する研究態度の問題でもある。また，コーパスを利用してどのような研究の地平が見いだせるのか，その見通しが明確でないという点もコーパスの地位をあいまいなものにしている。本稿では，コーパスを，言語の記述的研究のための手段として正当に位置づけたうえで，その普及のための素地となる，論文における用例の採り方を調査する。また，コーパスを利用した日本語研究のケーススタディとして，いわゆる「ガノ可変」の実態をとりあげる。

２．コーパスとは

２．１　本来の意味でのコーパス

「コーパス」とは，その本来の意味で言えば，言語分析のために集められた（言語）データのことである。これだけでは，何の変哲もない定義である。言語研究には，通常，資料やデータが欠かせないものであるから，分析のためにデータを集めるは，ごく当然のことで，ことさらそれを「コーパス」と呼び換えなくてもよいように思われる。

しかし，現在，言語研究において「コーパス」という言葉を用いるときには，上記の定義にさらに二つの条件が加わる。ひとつは，「電子化」，もうひとつは「大規模」である。

２．２　電子化コーパス

コーパスは，本来の意味ではその媒体を選ばないはずであるが，現在，コーパスというと，ほとんど確実にコンピュータで扱えるものをさす。これは，データ量の多いことによる扱いの利便性を考慮すれ

ば当然のことである。また，コーパスに関してはコンピュータで扱うゆえのさまざまな技術的な問題点や，それに対する創意・工夫などが存在する。（注１）コンピュータを利用することは，コーパスを利用する言語研究にとって本質的なことではない。コーパスの本質にかかわるのは，次の「大規模」という条件である。

２．３　大規模なデータ

たいていの言語は，多くの使用者がそれを用いる機会を持ち，過去から現在，そして未来へと継承されるものであるから，およそ言語使用の総体を把握することは不可能である。しかし，なるべくそれに近づくべく，着目する言語要素の使用実態を網羅することが，コーパスに期待される要件である。ここで前提にされているのは，実証的な研究態度であり，これは，戦後，国語学において発展した現代語研究において，とくに重要な研究態度の一つである。そして，使用実態を網羅的に把握するためには，データの量はおのずと大規模化に向かっていく。

２．４　現代語研究とコーパス

現代語は，古代語と違って，有限で固定的なデータを持たない。古代語研究は，すでに過去のものとなった言語について，もっぱら資料に基いて分析するものであるから，データはあきらかに有限である。また，異本などの存在もあるが，資料は，それ自体が閉じた世界を形成している。もちろん，古代語の研究とて，その資料のみの分析にとどまるのを目的とするわけではなく，当時の言語共同体における言語使用の実態や言語体系にせまるのが，その目的であろう。

しかし，現代語研究の場合は，様相がかなり異なる。データは「どこにでもある」のである。極端な話，研究者自身が内省によってデータを作り出すことができる。自然科学においては，研究者自身がデータを「作り出して」しまうことは，研究倫理にもとる行為とされるが，現代語研究では，内省は後述のように，ごく普通の研究手段として用いられている。

内省を用いないとすると，どういうデータを選んだらよいのか。現代語研究においては，データは「どこにでもある」のだから，本当に何を選んでもよいかというと，実はそうではない。分析目的に適したデータを適切な分量選び出さなければならない。古代語研究において，資料にあらわれた言語の特徴や時代性をきびしく吟味しなければならないように，現代語のデータにおいても，資料批判が必要である。現代の電子化コーパスも，その資料的特性を十分考慮したうえで利用すべきであろう。

２．５　内省と非文

内省による研究手段が成功をおさめたひとつの要因は，非文の存在である。正常な例文とそれと対になる許容できない例文との両方をしめすことによって，それらの文で対立された条件成立の可否を問うわけである。非文と正しい文とのペアによる分析方法は，違いがわかりやすい反面，言えるか言えないかについて人によって判断が分かれることもある。（注２）また，許容度の度合いによって，非文にも段階がつけられる場合がある。

非文の利用は，コーパス的な研究方法とは相容れないものでもない。日本語教育における誤用例分析は，いわば，日本人なら非文とするところの例文の分析であるし，言語情報処理において，言語解析における精度をあげるためには，誤った解析結果（＝非文）を分析することが重要となってくる。

２．６　日本語研究における電子化コーパス利用の動向

電子化コーパスを利用した日本語研究が普及してきたのは，１９９０年代になってからである。その要因は，もちろんパソコンの普及であるが，大量データとして新聞記事がＣＤ－ＲＯＭ化されて市販されたことも大きなきっかけとなっている。文部省の科学研究費補助金を受けたタイトルを見ても，「コーパス」あるいはそれに類似の用語を含むものが，平成６年度で４件だったのに対して，平成７年度１２件，平成８年度１３件と増えている。しかし，その内訳は，ほとんど情報処理系の研究であったり，人文系の研究でも，英語を対象にしたものであり，いわゆる国語学の分野は２，３件にとどまっている。

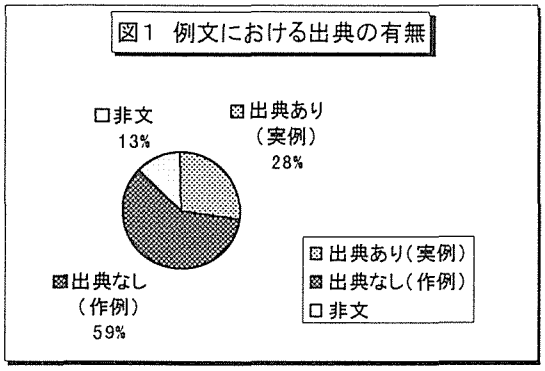
２．７　コーパス言語学への期待

電子化コーパスには，よく，その汎用性・共有化が求められるが，（注３）それ以上に重要なのは，コーパスを利用した新しい研究方法の開発である。コーパスは，言語研究にとっての一手段にしかすぎないが，これを利用することによって，従来あまりかえりみられなかった観点からの研究方法の開拓が期待されている。たとえば，田野村（1994）は，次のように指摘している。「ある言語形式の用法に関

わる統計的な偏りの現象は内省だけでは正確な把握が困難であるが，大量の用例を調査・分析することでその実相が浮かび上がってくる。」(p. 51)

３．実例と作例

日本語研究において，コーパス利用の素地となる例文の採り方はどうなっているか。それを例文における実例と作例の取り扱いという点から観察してみよう。実際の論文で用いられる実例と作例の用いられ方をみるために，次のような小調査を行った。「国語年鑑」１９９５年版および１９９６年版の雑誌論文の「文法」の分類に掲出された論文の中から，ランダムにその１０分の１を抜き出して，用例数，作例数，および出典数を調査した。抜き出された論文は２７編であるが，そのうち１編は，全く例文が出てこなかったため，これを除外し，残る２６編を対象とした。まず，全体の内訳を見てみよう。



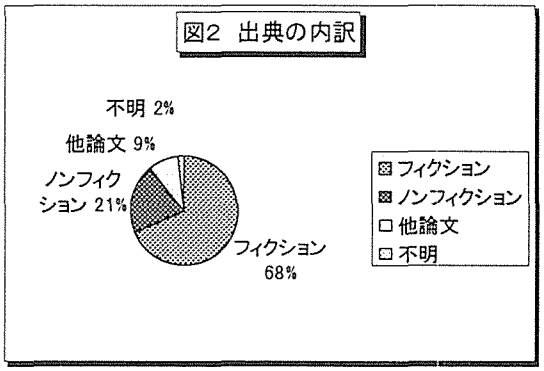


図１は，２６編全体の例文を，出典のあるもの，出典のないもの，および非文にわけて，その割合を示したものである。全例文のうち，出典のあるものは，約３割弱であることがわかる。

また，図２は，出典の内訳をおおまかに示したものである。ここで，「フィクション」とは，おもに小説，随筆などの文芸作品，「ノンフィクション」とは，実用書や新聞・雑誌，「他論文」とは，他の論文からの例文の引用をさす。図２から例文の典拠は，約７割がフィクションであり，他論文からの引用も，その原典ではフィクションが出典であるものが多いため，実際はこの割合はもう少し高い。

以上は，全体を概観したものであるが，個々の論文における実例と作例の出現のしかたはどうなっているのだろうか。図３をみてみよう。

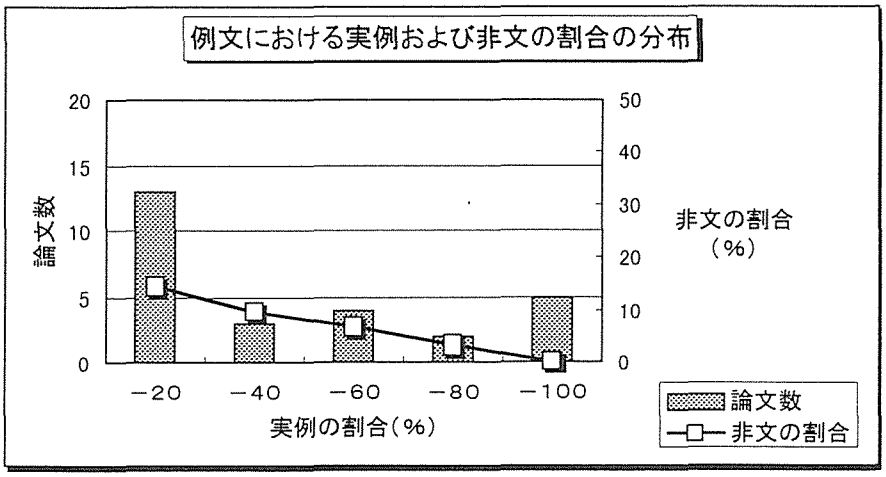


図３から，実例の占める割合が２０%以下という論文がいちばん多く，その次に多いのが８０%から１００%の論文であることがわかる。この傾向が調査対象の数を増やしたらどうなるか興味深いところである。すなわち，作例中心で若干の用例をまぜるタイプの論文と，用例中心でそこに若干の作例をまぜるタイプの論文が，典型的なスタイルとして取り出せそうである。さらに，ここで着目したいのは，非文のしめる割合である（図３の折れ線)。非文は，ほとんど作例であるためか，作例中心の論文で占

める割合が高くなっており，例文中の用例の割合が増えるとともに減っていくことがわかる。

４．連体修飾句中の「が」と「の」の使い分け

４．１　「が」と「の」の交替

コーパスを利用して，どんな言語事実の分析ができるか，その例として，次のような例を取り上げる。

日本語には，「ガノ可変」（三上(1953)）と呼ばれる文法的現象がある。

Ａ　雨が降る日　　Ｂ　雨の降る日

このように，連体修飾句中の主語（あるいは直接目的語）を表すのに，「が」「の」のどちらを使っても言える場合がある。すべての場合にそうなるわけではなく，「の」が使えない，あるいは使いにくい場合がある。それがどういう条件によるものか，従来指摘されているのは，次の２つである。

（１）連体修飾を受ける位置にある語句が名詞的な性質を持っていない場合（この場合，連体修飾そのものが成立しなくなる）

（２）主語（あるいは直接目的語）とそれを受ける述語との間に，他の語句が介在する場合（注４）

（１）の例として，「手紙ガ来タノガ遅レタノダ」をみてみよう。ここには，「が」が２回出てくるが，最初の「が」が「の」に置き換えられるのに対して，２つめの「が」は置き換えられない。これは，「来たのが」の「の」が，いわゆる準体助詞であり，名詞的な性質を持っているのに対し，「遅れたのだ」の「の」は，「のだ」の形でひとつの助辞的な形式となっているためである。（例は三上（1953）による。）

（２）の例として，「太郎がきのう学校に置き忘れた本」をみてみよう。この文の「が」を「の」に置き換えると，かなり不自然な感じがする。

これらの指摘は，おおむね妥当なものであるが，実例ではどういう傾向がみえるか。連体修飾句中の「が」と「の」の分布を二つの言語データを使って調べてみた。

４．２　書名における連体修飾句中の「が」と「の」使用実態

連体修飾句が比較的多く現れやすいものに，本の題名がある。そこで，「出版年鑑１９９７」に掲載されている全書籍（1996 年発行の全書籍）から，そのタイトルに連体修飾句中の「が」および「の」を含むものを抜き出して調査した。一定の条件のもとにデータを整備した結果，対象となったデータ数は 1,001 であった。ちなみに，書籍の総数は 60,462 である。全体の内訳を表１に示す。また，先の（２）の傾向がどうなっているかをみたのが表２である。

表１　書名における連体修飾句中の「が」と「の」

助詞	書名数	割合(%)
が	７６６	76.6
の	２３５	23.5
全体	１００１	100

表２　述語との間の語句の有無と「が」「の」の使い分け

述語との間の語句	あり	なし
が	２１３	５５３
の	８	２２７

表２によると，「の」で示される主語（あるいは直接目的語）とそれを受ける述語との間に何らかの語句のあるものが８例あるが，おおよその傾向は上記の４．１（２）に合致する。（注５）

４．３　述語の品詞と「が」と「の」使い分け

次に，述語にはどんな語がくるのかをみてみよう。表３（次ページ）によると，動詞が圧倒的に多く，全体の 92.7%を占める。形容詞・形容動詞は 7.2%。名詞２例というのは，名詞述語文に相当する例が二つあったためである。名詞を除き，述語が動詞である場合における「が」と「の」の割合と，述語が形容詞・形容動詞である場合における「が」と「の」の割合が逆転していることに注目したい。

では，述語には具体的にどのような語が多いのか。用例数が１０以上あったものは次の１６語である。「愛する(8/5)，ある(4/33)，いる(1/16)，教える(28/0)，語る(32/2)，知る(14/1)，好きな(2/8)，薦める(14/0)，住む(1/10)，付く(20/3)，出る(9/5)，ない(1/29)，なる(64/4)，見える(11/5)，見る(7/13)，

表３　述語の品詞と「が」「の」の使い分け

品詞	動詞	形容詞・形容動詞	名詞
が	７４８	１６	２
の	１７９	５６	０

わかる(173/3)」（　）内は，／の前が「が」の用例数，後ろが「の」の用例数である。この中で，「が」よりも「の」の用例数のほうが多いのは，「ある，いる，好きな，住む，ない，見る」の６語である。これらの中には，「～のある風景」「～の住む町」など，一種のパターン化した形式による書名が，「の」の用例数を増やしている例が見受けられる。

なお，述語のテンスおよび肯定否定についても，「が」「の」の使い分けについてクロス集計をしてみたが，いずれも，全体の傾向とほぼ同じ割合になり，有意差のある傾向は認められなかった。（χ^2値はそれぞれ，1.63，1.77である。）

４．４　雑誌九十種データにおける連体修飾句中の「が」と「の」使用実態

前節でみた書名という言語データは，ほとんどが名詞句のみの形，すなわち，文の形をしていなかった。では，実際に文として使用された用例における連体修飾句では，「が」「の」の使い分けにどういう傾向が見られるだろうか。国立国語研究所が１９５１年から１９６１年にかけて実施した，現代雑誌九十種の用語用字調査の用例カードを使ってその傾向をみてみよう。表４に全体の用例数と内訳を示した。表１の書名データの場合と同様，「が」の用例数のほうが多いが，表１とくらべると，全体にしめる割合は低くなっている。また，表５からは，表２で確認された，述語との間に何らかの語句があると「の」になりにくい傾向が認められた。（注６）

表４　雑誌九十種データにおける連体修飾句中の「が」と「の」

助詞	書名数	割合(%)
が	８６１	61.6
の	５３７	38.4
全体	１３９８	100

表５　述語との間の語句の有無と「が」「の」の使い分け

述語との間の語句	あり	なし
が	４７１	３９０
の	２１	５１６

また，表３でみられた，述語の品詞が形容詞・形容動詞の場合，「の」が選択されやすいことについても，同様の傾向が認められた（表６）。なぜ，このような傾向があるのかは，違った角度からの考察が必要だろうが，これが，実は表５で示した，述語との間に他の語句があるかどうかということとも関連していることを指摘しておこう。

表６　述語の品詞と「が」「の」の使い分け

品詞	動詞	形容詞・形容動詞	名詞
が	７７８	６３	２０
の	３６６	１７１	０

表７　述語との間の語句と述語の品詞との関係

述語との間の語句	動詞	形容詞・形容動詞	名詞
あり	４５８	１８	１０
なし	６８６	２１６	１０

表７に示すように，述語の品詞が動詞の場合は，その約４割に，主語あるいは直接目的語と述語との間に他の語句が介在しているが，述語の品詞が形容詞・形容動詞の場合は，他の語句の介在が 7.7%しかない。他の語句が介在しにくいということは，すなわち，「の」の使用が許されやすい条件の一つがそなわっているということである。

４．５　連体修飾句の述語において「の」をとりやすい語

具体的に用例が１０例以上あった語は，次の２１語である。

「ある (50/68)，いい(1/14)，言う，(7/15)，行く(10/1)，いる(31/2)，多い(9/9)，行う(8/3)，来る(12/4)，少ない(3/8)，する(21/10)，付く(8/13)，強い(1/10)，できる(4/16)，出る(18/5)，ない(20/51)，なる(41/3)，入る(9/4)，持つ(12/7)，やる(12/5)，良い(0/15)，悪い(0/15)」( ) 内は，／の前が「が」の用例数，後ろが「の」の用例数である。

この中で，「の」の用例数が「が」を上回っているのは，「ある，いい，少ない，付く，強い，できる，ない，良い，悪い」である。ほとんどが形容詞であるが，動詞も「ある，付く，できる」の３語まじっている。このうち，「ある」は書名のデータでも「の」の優位性が認められたが，「付く」は，書名データでは，「が」が多く現れた語であり，「できる」も書名データでは９例あり，全部「が」の例である。「付く」，「できる」については，これらが書名データや雑誌データでの偶然の結果なのかどうか確かめる必要があるだろう。(注７)

５．おわりに

ここでケーススタディとしてとりあげた，連体修飾句中の「が」「の」の使い分けの傾向は，言語事実とコーパスの性質・規模との関係をみるうえで，ひとつの参考事例となるだろう。今後，他の種類のデータについてもこの調査をすすめ，コーパスの種類と言語事実との関連を明らかにしたい。なお，４．１（１）にあげた，連体修飾を受ける名詞の性質と「が」「の」の使い分けについては，今回は分析していない。これも，今後の課題としたい。

注

注１　たとえば，近藤(1992)に，電子化テキストの取り扱いに関する問題点が指摘されている。

注２　ドメニコ・ラガナ(1988)に，内省を利用した例文が実際にどれくらい許容度があるかを調査した報告がある。

注３　「国語学」178集に，「電子化テキストの国際的共有」という特集がある。

注４　ここで，「主語」「直接目的語」「述語」というのは，着目している連体修飾句内でのことである。以下も同様である。

注５　８例の内訳は，「良く」３例，「いちばん」１例，「気に」１例（「気になる」を分割してできたもの)，並立句３例である。並立句とは，「時間の使い方のうまい人・下手な人」のような書名のとき，「時間の使い方が」と「下手な人」との間にある「うまい人」のような語句をさす。

注６　他の語句が介在した２１例の内訳は，以下のとおり。「あまり，大きく，多く」などの副詞的修飾語句が１９例，「職業を，我が国輸入総額に」のような別の補語が２例。また，語句の長さでは，１語の介在が１７例，２語の介在が４例あった。これは，塚本（1991）の「『が』で表示された主語或いは直接目的語と述語との間に，他の補語や副詞類などの構成要素が長く介在すればするほど，その文全体の文法性は低くなる。」という指摘とも合致する。

注７　「付く」の「が」８例，「の」１３例という分布は，χ^2検定では有意差とは認められない。

使用データ

「出版年鑑１９９７」第２巻 目録編　出版ニュース社　1997.5.23 発行

現代雑誌９０種の用語用字調査の用例カード　国立国語研究所言語体系研究部第１研究室蔵

参考文献

近藤泰弘(1992)　文法研究と電子化テキスト　「国語学会平成４年度春季大会要旨」

田野村忠温(1994)　丁寧体の述語否定形の選択に関する計量調査―「～ません」と「ないです」―　「大阪外国語大学論集」　11

塚本秀樹(1991)　日本語における格助詞の交替現象について　愛媛大学法文学部論集文学科編　24

ドメニコ・ラガナ(1988)　「これは日本語か」　河出書房新社

豊島正之ほか(1994)　［テーマ別研究］電子化テキストの国際的共有　「国語学」178

三上章(1953)　「現代語法序説」　刀江書院　(引用は1972年のくろしお出版の復刻版による。)