日本語コーパスとコロケーション
――辞書記述への応用の可能性――

田野村忠温
大阪大学

【要旨】 コーパスは従来のタイプの言語研究の精密化にも大きな力を発揮するが，コーパスの特性を生かすことにより旧来の言語研究資料では考えられなかったさまざまな種類の研究が可能になる。

本稿は，大規模なコーパスを使って初めて満足な調査・分析が可能になる言語現象の1つであるコロケーションを主題とし，日本語コーパスの分析によって得られるコロケーション情報が日本語の一般的な辞書ないしコロケーション辞典の作成にどのように生かせるかという応用的な関心に基づいて考察を行う。具体的には，筆者が‘circumcollocate’と呼ぶ現象の分析，述語の有標率の分析，類義的な慣用句型の意味・用法の分析におけるコーパスの有用性について述べる。

コロケーションの分析には大規模なコーパスが必要となる。本稿では，筆者が2008年に作成した巨大なWebコーパスを使用する。その規模は約750億字，ファイルサイズにして約150ギガバイトであり，平均的と思われる小説単行本の30～40万冊に相当する*。

キーワード： 日本語コーパス，コロケーション，辞書，circumcollocate，Webコーパス

1. はじめに

コーパスは，言語研究資料として大量性と処理の柔軟性という2つの著しい特性を併せ持ち，言語の研究に無限の恩恵をもたらす。コーパスから単に多数の用例を収集して従来のタイプの研究をより精密で実証的なものにすることもコーパスの有益な用途であるが，コーパスの特性を生かすことにより旧来の言語研究資料では望むこともできなかったさまざまな種類の研究への道が拓かれる。

大規模なコーパスを使って初めて信頼性の高い調査・分析が可能になる言語現象の1つに，コロケーション（collocation）がある。拙論（2009）――以後「前稿」とする――では，日本語のコーパスからコロケーション情報を抽出する手法につい

* 本稿は平成18～22年度文部科学省科学研究費補助金特定領域研究「日本語コーパス」の計画研究の1つである「コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発」（研究課題番号18061004，通称「日本語学班」）による研究成果の一部であり，国立国語研究所共同研究プロジェクト「コーパス日本語学の創成」研究発表会（国立国語研究所，2010年2月1日）や科学研究費補助金特定領域研究「日本語コーパス」の複合辞研究会（筑波大学，2010年2月13日）などでの発表内容に基づく。また，査読者の方の意見に基づいて一部の記述を詳しくした。関係の機関，各位に謝意を表したい。

て検討し，コーパスの分析によって得られた情報に基づいてコロケーション辞典の項目を試作した。本稿では，そこでの議論を出発点とし，より進んだコーパスの利用によって日本語のコロケーションに関する我々の理解をさらに深め，辞書記述の質を高めることができることを 2，3 のテーマの考察に基づいて示す。

2. 考察の前提と分析方法

本題に入る前に，その前提となる背景的事情について，前稿で述べた議論を踏まえそれに多少補足する形でできるだけ手短に述べる。2.1. でコロケーションの研究に関わる私見，2.2. ～ 2.4. で日本語コーパスからのコロケーション情報抽出の方法とその応用を取り上げる。

2.1. コロケーションの概念とその研究

筆者はコロケーションを言語表現どうしの習慣的な共起関係全般として広く理解する。コロケーションの概念を語と語の関係に限定すべき理由はなく，したがって，句や節，あるいは，構成素を成さない語連続もコロケーションの観点から観察すべき対象の範囲に含める。

コロケーションに関わる情報は，言語の事実を明らかにするという意味での狭義の言語学的研究にも役立つだけでなく[1]，辞書や語学の教科書の編集といった応用的な方面での利用価値が大きく期待される。コロケーションはしばしば言語表現の自然さに関与し，コロケーションの慣用に従うことで表現が自然なものとなり，慣用に従わなければ文法的ではあるにせよ不自然な表現となり得る。本稿では日本語の辞書ないし特にコロケーション辞典の作成にコーパスに基づくコロケーションの分析がどのように生かせるかということを意識の中心に置いて考察を行う。ここで想定する辞書は伝統的な体裁のものにとどまらず，インタラクティブな例文検索システムやコロケーション表示システムのようなものの開発も将来的には期待されるところである。

コロケーションについては従来複数の評価法が提案されており，また，そのそれぞれに“癖”があり無条件に信頼して使えるわけでないこともよく知られているが，本稿の応用的な見地からすれば評価法の問題をあまり厳密に考える意味はない。その主な理由としては，第 1 に，そもそも辞書や語学教科書の記述においては，コーパスの調査によって得られたコロケーション情報を忠実に反映させればよいというものではないということがある。機械的な分析で得られた情報に目を通し，記述の優先順位を調整したり同類の表現をグループ化して冗長性を解消したりといった人手による作業が欠かせない。第 2 に，コロケーションの既存の評価法は単純なコロ

[1] 語彙や文法の研究の多くの局面で表現の共起に関わる情報が重要な役割を果たすことは周知の事実である。また，堀・浮網・西村・小迫・前川（2009）や服部（2010）に見られるコロケーション自体の通時変化の様相の解明も狭義の言語学的研究に属する興味深い研究テーマである。

ケーション理解を前提としており，本稿で行うような多少複雑な性質の分析には通用しない。そしてまた，既存の評価法に代わる万能で信頼性の高い評価法を考案することもおそらく望めない。コーパスからのコロケーション情報の抽出は所詮文字の連続に過ぎないコーパスから表面的な分析によって情報をかき集めてくるという粗い作業であり[2]，精密化を図ろうにも自ずと限界がある。以上のような本質的な理由に加えて，第 3 に，どのような評価法でコロケーションを分析するにせよ，得られる結果は用いるコーパスによって少なからぬ食い違いを見せるという現実的な問題もある。

2.2. 使用するコーパス

コロケーションの分析には大規模なコーパスが必要となる。ある表現 A とほかの不特定多数の表現 B_1，B_2，B_3，…とのあいだの共起傾向を調べようとすれば，A と B_i（i=1, 2, 3, …）の共起例がそれぞれ大量に獲得できる必要がある。通常の分析なら表現 A の用例が例えば 3 桁台の数で見つかれば足りるとしても，コロケーションの分析においては A と B_i（i=1, 2, 3, …）の共起例がそれぞれ最低限 2 〜 3 桁はほしいので，全体として A の用例数は少なくとも千，万のオーダーである必要があることになる。

本稿では，コロケーションの分析に，筆者が 2008 年に作成した Web コーパスを使用する。これは，インターネット上の大量の日本語文書を収集・集積した巨大なテキストで，規模は約 750 億字（ファイルサイズで言えば 150 ギガバイト），平均的と思われる小説単行本の 30 〜 40 万冊に相当する。文書数で言えば約 1,500 万件の文書の集積である。本稿での分析にはその 3 分の 2 に当たる 500 億字（100 ギガバイト）分ないしその一部を使用する[3]。

---

[2] これにはいくつかの理由があるが，ここではその最大の問題を挙げることにすれば，コロケーションの評価は通常もっぱら表現間の距離に基づいて行われる。このため，「〜ネズミが食べた〜」「〜ネズミを食べた〜」「〜ネズミと食べた〜」などの形をした用例において，「ネズミ」と「食べた」は文法・意味上それぞれ異なる関係にあるにもかかわらず，同等の共起関係として扱われる。また，距離にのみ基づくということは表現の構造を考慮しないということでもある。「ライオンが眠っている」と「ライオンが眠っているシマウマを襲う」という 2 文のうち，後者では「ライオン」と「眠る」のあいだに直接の関係はないが，それでも前者と同じく両語の共起した例として扱われる。「風が開けた窓から入って来る」という文は「風が（誰かが）開けた窓から入って来る」とも「風が開けた窓から（何かが）入って来る」とも取れるが，「風」と「開けた」「入って来る」との共起関係の認定はそうした構造上の曖昧性に関わりなく一律に行われる。コーパスに基づくコロケーションの分析は，本来無用の情報の混入を容認し，しかしそうした“不純物”が結果に与える影響は少ないだろうとの予測ないし期待に基づいて行われるものと言ってよい。

[3] 前稿でも述べたが，Web コーパスの作成は，まずサーチエンジンにさまざまなキーワードないしキーワードの組を順次与えて検索し，それによって得られる URL が指している文書を取得し，そこから HTML タグなどの不要な情報を除去する，という手順で行った。その際，サーチエンジンに与える検索キーワードによって，得られる文書の性格は異なってくる。基本的には，総体として大きな偏りがないと思われるキーワード群を用いる方法によるものとし，それに加えて，特定の話題や文体に偏ったキーワード群に基づく文書収集も行った。2 つの

インターネット上の日本語文書を言語の研究に用いることには，従来広く使われてきた書籍，雑誌，新聞などの出版物の電子テキスト（あるいは，出版物に基づいて作成されたコーパス）を資料とする場合にはない長所も短所もある。コーパスの評価はその利用目的によって基準が違ってくるが，コロケーションの分析という観点からすれば，インターネット文書の大きな価値は実際上無限とも形容し得るその大量性にある。Web コーパスの上記の規模は，例えば従来の日本語研究でコーパスとして広く使われてきた『CD-ROM 版　新潮文庫の 100 冊』（新潮社, 1995 年）に比べればその 3,000 ～ 4,000 倍に当たる。これは，Web コーパスを使えば用例が 1 万例得られるところでも，『新潮文庫の 100 冊』では用例がわずか 2 ～ 3 例しか得られないことを意味する。

インターネット文書の短所としては，特殊な言葉遣いの出現や書き誤りの多さといった問題が予想されやすいが，Web コーパス使用の経験上，コロケーションの分析を含む定量的な調査にとって真に問題となるのはむしろ同一データの重複出現である。インターネット上にはさまざまな事情で同一の文章，段落，文，句が複製されて繰り返し現れる。Web コーパスを用いてコロケーションを調査すると一般性の低い特定の表現が共起表現のリストの上位に現れることがあり，そのようなときは調べてみるとほぼ確実に複製データの重出が原因となっている。Web コーパス作成時に重出データを排除することは現実的にも原理的にも無理であり，したがって，分析結果に含まれる無用の情報は目視によって除外するしかない。複製データの重出に比べれば軽微ながら類似した問題がほかにあり，後に 4.5. で具体的なコロケーション分析の文脈において述べる。

2.3. コロケーション情報抽出の手法

分かち書きをしない日本語の場合，コロケーション情報の抽出に際してまず考えなければならない問題は，何を単位としてコロケーションを分析するかということである。前稿では「共起語分析」「共起語連鎖分析」「共起文字連鎖分析」と名付けた 3 通りの方法を試し, 第 2 の共起語連鎖分析の方法が最適であるとの結論を得た。

第 1 の共起語分析は，大まかに言って，所与の語句――以後「中心語」とする――の近傍にどのような語がよく現れるかを調べる方法である。語（ないし形態素）の認定は本稿では形態素解析ソフト MeCab と形態素解析辞書 UniDic を用いて行う [4]。第 2 の共起語連鎖分析は，中心語の近傍にどのような語連鎖（すなわち，語レ

方法によって収集したデータの量の比率は 2:1 であり，本稿での分析には前者の方法で収集したデータを用いる。なお，作成に時間をかけさえすればコーパスの規模をいくらでも大きくすることができるが，今日のパーソナルコンピュータの処理能力を考慮すれば現在の程度の規模が現実に処理可能な上限に近い。

[4] MeCab と UniDic はそれぞれ http://mecab.sourceforge.net/，http://www.tokuteicorpus.jp/dist/ で配布されている。なお, 前稿執筆時には形態素解析辞書としては MeCab に添付されている IPA 辞書を使った。UniDic も試した結果，前稿の注 3 に記した理由により目的とする分析には適さないと判断したからであったが, それは UniDic に関する筆者の理解不足に過ぎなかった。

ベルの N-gram）がよく現れるかを調べるものである。共起語分析と共起語連鎖分析の相対評価について結論的に言えば，前者では不十分で，後者を採用する必要がある。その詳細は前稿で述べたが，例えば「ついに」という副詞によく後続する動詞を共起語分析の方法で調べてみると，「耐える」が比較的上位に位置することが分かる。しかし，それは我々の求める情報の一部でしかない。なぜならば，「ついに」と「耐える」の共起は，ほとんどもっぱら「ついに耐えられなくなった」「ついに耐えきれなくなって」などのように可能・否定・変化という 3 つの要素の存在に依存しているのであって，「ついに」が「耐える」と無条件に習慣的に共起するわけではないからである。したがって，「ついに」とよく共起するのは「耐える」という単独の語ではなく，「耐えられなくなった」「耐えきれなくなって」のような語連鎖だと言わなければならない。

第 3 の共起文字連鎖分析は，中心語の近傍にどのような文字連鎖（すなわち，文字レベルの N-gram）がよく現れるかを調べるものである。これは共起語連鎖分析に近い結果をもたらすが，語や形態素の境界を考慮に入れない分析であることから得られる結果の冗長性が高い（詳細は前稿を参照願いたい）。形態素解析を要しないことによる処理効率上のメリットも期待したほどではなく，総合的に共起語連鎖分析が最も望ましい分析方法であるという結論を得た。

2.4. コロケーション辞典項目の試作

前稿では，上述のような分析手法の検討の後，コーパスの分析によって得られたコロケーション情報に基づいて日本語のコロケーション辞典における数項目を試作した。「希望」「願望」という類義的な 2 語の試作項目を次頁に掲げる。内省や少数の用例のみに基づく考察でこうした記述を準備することが容易でないことは明らかであり，かりにそれが可能だとしても作業の効率は非常に悪いはずである。

以上のことを踏まえ，以下の各節で本稿の考察について述べる。

3. ‘circumcollocate’——概念と考察

最初に取り上げるテーマは，筆者が‘circumcollocate’と呼ぶ概念の導入と具体的な事例に基づく考察である。

3.1. 問題の所在

前稿での共起語連鎖分析においては，中心語の文脈を先行文脈と後続文脈に分け，そのそれぞれに高頻度で現れる語連鎖を別々に調査した。前後の文脈を区別しない分析では，得られる語連鎖のリストが雑然として見づらいものになるからである。例えば，「めっきり」という副詞とよく共起する語連鎖には「最近は」「朝晩」「減った」「寒くなってきた」などがあるが，分析の結果として得られる共起語連鎖のリストにそうした表現が入り交じった形で現れていては非常に見づらい。そこで，中心語の前後の文脈に現れる語連鎖をそれぞれ独立に扱うことにしたわけである。

コロケーション辞典の試作項目　「希望」「願望」

希望　きぼう

1　名詞

◇「希望」＋動詞

—を持っている、—が持てる
……*若い人たちが読んで希望が持てるような本にしたい*

—を実現する、—を捨てる
……*希望を捨てずにいっしょに頑張りましょう*

—が叶う、—が実現する
……*希望が叶ってほんとうによかったですね*

(ご)—に添う、(ご)—に添える
……*なるべくご希望に添えるよう努力します*

◇連体修飾句＋「希望」

～したいという—
……*定年後も働きたいという希望*

生きる—
……*彼らの活動は何百万もの人々に生きる希望を与えた*

強い—
……*本人の強い希望により予定より早く退院した*

◇その他

(～の)—としては
……*私の希望としては経費を10万円以下に抑えたい*

(ご)—があれば
……*ほかにも希望があれば窓口の担当者に相談してください*

夢と—、勇気と—
……*その映画は子どもたちに夢と希望を与えた*

2　複合サ変動詞語幹

～を—する
……*私の妹は商社への就職を希望している／結婚後も就労を希望する女性が増えている*

(～を)(ご)—のかた、(～を)(ご)—のお客様
……*パンフレットをご希望のかたは電話でお申し込みください*

(ご)—の日、(ご)—の時間、(ご)—の商品、—の職種
……*ご希望の時間をお知らせください／希望の職種に就けるよう努力している*

3　複合語要素

◇「～希望」

匿名—、参加—、進学—、購入—、入居—
……*私は家庭の事情で進学希望が叶わなかった*

◇「希望～」

—者、—小売価格、—退職
……*数に限りがありますので希望者は早めに申し出てください／うちの会社は業績の悪化で希望退職の募集を始めた*

—的観測
……*希望的観測では工事は年内に終わると思う*

願望　がんぼう

1　名詞

◇「願望」＋動詞

—を抱いている、—を持っている、—がある
……*私は自分の将来に関して２つの願望を抱いている*

—を実現する、—を達成する、—を叶える
……*私もさまざまな分野で自分の願望を実現していきたい*

—が叶う、—が成就する
……*ついに長年の願望が叶って、スイスにスキーに行けることになった*

◇「願望」＋形容詞

—が強い
……*願望が強すぎると人生はかえってうまくいかない*

◇連体修飾句＋「願望」

～したいという—、～になりたいという—
……*大空を鳥のように飛びたいという願望は昔から人々の心にあった*

単なる—、勝手な—、個人的な—
……*ただ平和を祈念するだけでは単なる願望にすぎない*

強い—
……*では、やりたいことがより強い願望に変わるのはどんなときですか？*

◇その他

—や欲求、—と期待、—と意志
……*自由でありたいという願望や欲求はあらゆる人に共通するものだ*

2　複合サ変動詞語幹　※比較的まれ

～ことを—する
……*素晴らしい未来がこの国に訪れることを願望してやみません*

3　複合語要素

◇「～願望」

結婚—、変身—、自殺—

しかし，前後の文脈に分けてコロケーションを考えることには実のところ問題がある。そのことを指摘した前稿の一節を次に引用する。

> 「飲む」という動詞の目的語としてよく現れる名詞には，容易に予想できる通り，「水」「酒」「お茶」「コーヒー」「ミルク」などがある。ところが，そうした名詞のうちどれがよく現れるかは，「飲む」がどのような形の言い回しで使われるかに依存する。例えば，Yahoo! 知恵袋ベータ版データ[5]の場合，「〜を飲みます」という言い回しにおいては「水」「薬」「酒」「コーヒー」「お茶」などがその順に高頻度で現れるのに対し，「〜を飲みません」という言い回しでは意外にも「ミルク」が第１位となる。これは，同データでは，子どもがミルクを飲まないという話が育児に関する相談，やり取りによく出て来ることによる。また，「〜は飲みません」では「酒」が第１位となる。さらに，普通体の「飲む」「飲まない」においてはまた様子が違ってくる。このように，「飲む」行為とその対象物という同一の関係においても，述語の肯定・否定，名詞に付く助詞，文体の種類などの複合的な条件によって表現の共起の傾向が異なってくるのである。

3.2. circumcollocate の概念

前稿では指摘しつつも正当に対処できずに終わっていたこの問題は，「飲む」の共起表現として「ミルクを__ません」「酒は__ません」のような不連続な語連鎖を認定することによって解決を見る。すなわち，コロケーションの観察の対象として，連続的な語連鎖だけでなく不連続な語連鎖も視野に入れるということである。

「ミルクを__ません」「酒は__ません」のように中心語を前後からはさむ形の不連続な語連鎖を‘circumcollocate’と命名する。本来ならば日本語での議論の文脈では日本語の用語としたいところであるが，満足の行く表現が思い当たらないので拙論（2010）で導入した英語の造語をここでも用いる。便宜上先行文脈中の共起表現を‘precollocate’，後続文脈の共起表現を‘postcollocate’と呼ぶことにすれば，circumcollocate は

(1)　circumcollocate = precollocate + postcollocate

として図式化することができる。

3.3. circumcollocate の例

さて，3.1. に引用した前稿の一節で論じた例に限って言えば，「飲む」の辞書

[5]「Yahoo! 知恵袋ベータ版データ」とは，ヤフー株式会社が運営する「Yahoo! 知恵袋」（利用者が知識を共有するための問答式掲示板サイト）の正式運用開始前の段階で書き込まれた問答の内容である。同データは，国立情報学研究所の Web サイト http://research.nii.ac.jp/（「『Yahoo! 知恵袋』データの提供について」のページ）を通じて配布されている。

項目に「ミルクを__ません」「酒は__ません」のような circumcollocate があることを情報として含める意味は大きくないかも知れない。しかし，現に辞書，特にコロケーション辞典への記載に値すると考えられる circumcollocate はさまざまな種類の表現について観察される。以下に，Web コーパスの分析によって得られた circumcollocate の例をいくつか示す。分析は，2.3. で述べた共起語連鎖分析の方法を基礎とし，それを不連続な語連鎖も扱えるように拡張したものによった。使用した Web コーパスの量は 500 億字（100 ギガバイト）分である。

まず，次に示すのは動詞「尽きる」を含む慣用的な表現の一部である。

(2)　a.　{力／万策／愛想が}　尽きた
　　b.　{興味　{が／は}　／疑問は／悩みは}　尽きない

「力」「万策」「愛想が」「興味　{が／は}」「疑問は」「悩みは」などが「尽きる」の主語ないし主題としてよく現れるわけであるが，それらが一律に「尽きる」と慣用的に共起するわけではないことを（2）は示している。すなわち，(2a) に見る通り「力」「万策」「愛想」は「尽きる」の肯定の過去形，(2b) に見る通り「興味」「疑問」「悩み」は「尽きる」の否定の非過去形と共起する顕著な傾向がある。したがって，単に「尽きる」の慣用的な共起表現に「力」「万策」「愛想」「興味」「疑問」「悩み」などの名詞があると述べただけでは不十分であることになる。本稿の見解は，「力__た」や「悩みは__ない」といった不連続の語連鎖を「尽きる」の circumcollocate として認定すべきだということであり，そうした情報をコロケーション辞典に含めることは precollocate や postcollocate の情報と同様に，しかも場合によってはそれ以上に，価値があるはずである。

次は同じく動詞の「許す」を含む慣用的な表現の例である。

(3)　a.　{お時間／天気／体力}　が許せば
　　b.　{プライド／良心}　が許さない
　　c.　{予断／妥協／〜の追随}　を許さない
　　d.　{先制／逆転}　を許した

(3a) と (3b) は「許す」の主語，(3c) と (3d) は「許す」の目的語が関わる表現群であるが，それらに含まれる不連続な語連鎖の circumcollocate としての意味合いはあらためて説明するまでもないであろう。(3a) と (3b) の比較で言えば，「お時間が許…」と来れば仮定の表現，「プライドが許…」と来れば否定の表現が後によく続くことは母語話者の感覚としてもよく分かる。(3c) と (3d) の対比についても同様である。したがって，「許す」の circumcollocate として「お時間が__ば」「プライドが__ない」「予断を__ない」「先制を__た」といったものが認められることになる。

次の例は形容詞の文語連体形「多き」に関わる例である。

(4)　a.　恋多き {女／男／女性／人／乙女}
　　b.　{悩み／夢} 多き {年頃／青春／人生／日々}

「A 多き B」という形の表現が口語的な「A の多い B」の形で言い換えられることは言うまでもないが，逆に後者の形の表現が自由に前者の形に言い換えられるわけではないことに注意されたい。車の多い道のことを「車多き道」と表現したり，小骨の多い魚のことを「小骨多き魚」と表現したりすれば，文法上の問題はないにせよ日本語の慣用を外れた表現となる。したがって，「多き」については，「恋＿女」「悩み＿年頃」などの不連続語連鎖がその circumcollocate として認定されることになる。

次は副詞「まっすぐ（に）」に関わる例である。

(5)　a.　{この道／通り／商店街／信号} をまっすぐ {行く／進む}
　　b.　（〜の）{目／顔} をまっすぐ {見る／見つめる}
　　c.　{背筋／足／手} をまっすぐ（に）{伸ばす／する}

ここでは，「まっすぐ（に）」の circumcollocate として「この道を＿行く」「（〜の）目を＿見る」「背筋を＿伸ばす」などを認定することになる。なお，日本語の語順の自由性により，例えば「背筋をまっすぐ伸ばす」は「まっすぐ背筋を伸ばす」と言うこともでき，後者においては「背筋を伸ばす」は連続しており，もはや circumcollocate ではないことになる。このように語順の交替によって circumcollocate の認定が変わってくる場合がある。応用的な観点からすれば，そのこと自体が circumcollocate の概念を考えるうえでの障害となることはないが，辞書項目に共起表現の情報をどのように記載すべきかという問題については一考を要することになる。

3.4. circumcollocate に関する補説

さまざまな表現の circumcollocate を観察していくとその情報を辞書の記述に反映させるうえで考えるべき問題がいろいろと出て来るが，ここでは circumcollocate の概念に関する理解を深めるために 2 つの点について考える。

その第 1 の点は，一見 precollocate に見えるものの中にも，circumcollocate として理解すべきものがあるということである。

次に示すのは上で取り上げた動詞「尽きる」を含む別の慣用的な表現である。

(6)　{〜の一言／〜冥利} に尽きる

この下線部「〜の一言に」「〜冥利に」はいずれも表面的に見れば「尽きる」の単なる precollocate である。しかし，(6) の表現においては「尽きる」が通常肯定形であるという意味で，中心語「尽きる」の後続文脈に制約が課せられることに注意しなければならない。したがって，ここで我々は「〜の一言に＿ϕ」「〜冥利に＿ϕ」――ここでの ϕ は否定辞の不在を示すものとする――という形の

circumcollocate を見ていると理解する必要がある[6]。

これとは対照的に，例えば動詞「合う」の共起表現としての「ぴったり」という副詞を例に取れば，これは「ぴったり合う」「ぴったり合わない」のように「合う」の後ろに否定辞が続く用法も続かない用法も一般的である。したがって，少なくとも肯定・否定の極性に関して言えば，「ぴったり」は「合う」の純粋な precollocate であると言えることになる。

第 2 の点は，一見コロケーションの観点から見る必要がない現象でも，circumcollocate の概念を踏まえて細かく観察すれば，そこに新たな発見があり得るということである。そのことを，前出の動詞「許す」を例に述べる。

(7)　絶対許さない

この表現は，単に副詞「絶対」が述語「許さない」を修飾するだけのもののように見える。すなわち，「絶対」は広い範囲の述語と共起する一般的な表現であり，(7) をコロケーションの観点から考える余地はないかのようである。

しかし，事実がそれほど単純でないことはまず次の対比から確かめられる。

(8)　a.　絶対　{行く／行かない}
　　b.　絶対　{?許す／許さない}

(8a) の「絶対行く」と「絶対行かない」はいずれもよく使われるが，(8b) の「絶対許す」という肯定の表現は「絶対許さない」という否定表現に比べて比較にもならないほど用例の頻度が低い（「許す」の前に加えた ? はそのことを示している）。したがって，「絶対」はあらゆる述語と慣用的に組み合わせられるわけではなく，「絶対許さない」という表現はコロケーションの観点からそれなりに注目すべき表現である――すなわち，「許す」の circumcollocate として「絶対__ない」という不連続語連鎖を認定する意味があり得る――ことになる。

さらに言えば，動詞「許す」に限っても，そのさまざまな形での現れとそれらを慣用的に修飾する複数の副詞のあいだに相関が見られることも指摘に値する。次に示すのは，Web コーパスで「許せない」「許されない」「許すことができない」という 3 つの否定述語のそれぞれを修飾するのに最もよく使われている副詞である。

(9)　a.　{絶対（に）／どうしても／本当に}　許せない
　　b.　{絶対に／決して}　許されない
　　c.　{断じて／絶対に}　許すことができない

この対比が示すように，文法上の可能性としては広い範囲の述語を修飾する副詞

[6]「尽きる」を形態的に「尽き - る」と分析し，「る」を非過去の接辞として理解すれば，動詞（の語幹）「尽き」が「{～の一言／～冥利に}　に__る」という circumcollocate を持つということになる。このように分析・表現するならば，当該の circumcollocate が単なる precollocate のように見えるということはなくなる。

群であっても，しかも，それらが同一の動詞「許す」を含む述語を修飾する場合に限っても，観察を細かくしていけば，それぞれの副詞の使用頻度，表現の慣用度には述語の形に依存した違いがある。こうした事情は本稿の用語で言えば circumcollocate の見地から理解すべきものにほかならない。

4. 述語の有標率

文法的な要素に関わるコロケーションは特にコリゲーション（colligation）とも呼ばれる。英語の例で言えば，ある種の動詞は to 不定詞を従えるが，ある種の動詞は -ing 形を従えるといった現象である。ここではいろいろと考えられる日本語のコリゲーションの現象のうち，日本語の辞書編集や文法研究に生かせる可能性のあるものとして述語の有標率の問題を取り上げる。

4.1. 問題の所在

個々の用言はどのような形で使われるかに関して偏りがある場合がある。例えば，「見当たる」という動詞はほとんどもっぱら否定の文脈で使われる。「そぐう」「たゆむ」「離れる」（「〜のことが頭を離れない」のような表現の場合）「承知する」（「〜したら承知しない」などの表現の場合）なども同様である。また，「ござる」はほとんどもっぱら「ます」を伴う丁寧体で用いられる。「致す」「参る」（寺社参詣，降参の意の場合を除く）「申す」などにも同様の傾向がある [7]。

既存の辞書も，「見当たる」のような語についてしばしば「多く打消しの形で用いる」のような注記を施している。ここでは，コーパスを用いてあらゆる用言の用法を網羅的に調査・分析することにより，そうした注記を充実させるための参考情報を得る可能性を探る。

用言の用法の傾向に関する情報は，辞書におけるより適切な例文の選択にもつながる。例えば，次に示すのは既存の辞書が「大別」という語の項目に挙げる例文である。

(10)　a.　「東日本と西日本に―する」（『大辞林』第 2 版）
　　b.　「二つに―する」（『広辞苑』第 5 版）
　　c.　「夏服と冬服に―する」（『大辞泉』）
　　d.　「二つの部門に―する」（『明鏡国語辞典』）

いずれの辞書もこのように能動態の例文を挙げるが，Web コーパスの軽い分析によれば「大別」は受け身の形での用例の比率が約 92% と非常に高い（この比率に補足すべき事情があることは後に 4.5. で述べる）。伝統的な体裁の辞書における

[7] さらに細かく見ていけば，「たゆむ」は多くの場合「たゆまず」「たゆむことなく」のような形で後続の述語を修飾するのに使われるとか，「申す」は「私は〜と申す者です」のように連体修飾節では問題なく普通体で現れる場合があるといった個別の事実が明らかになる。

例文の選択には，文字数の制約があり，また，単純で分かりやすい例文を示すという配慮も求められるであろうから，用例の比率だけに基づいて例文を選べばよいというものではないが，例文選択時に考慮すべき重要な検討材料の 1 つであるはずである。実際，例文の短さや単純さという条件を優先する必要のない種類の辞書の場合には，用法の現実に即した例文を挙げるほうが望ましいことは明らかである。

4.2. 調査の方法

さて，従来の辞書における「多く打消しの形で用いる」のような記述は，項目執筆者の内省や先行する辞書の記述を手がかりとした個々の用言ごとの検討に基づいて準備されてきたものと思われる。ここでは，Web コーパスを用い，コーパスにおけるあらゆる動詞・形容詞について，それらが各種の助動詞類とどのような共起傾向にあるかを一挙に調査・分析する可能性を試みる。

調査は，多数の述語形を代表する一定のパターンを設定してそれに合致するすべての用例を収集し，その中で文法要素がどのように分布しているかを分析するという方法によった。形態素解析にはここでも MeCab と UniDic を使用した。

使用したパターンは，次に示す句点で終わる 2 通りの文末パターンである。ほかに複合サ変動詞と形容動詞についても同様の調査を行ったが，ここでは省略する。

(11)　a.　和語動詞（(ら）れる）（ている）（否定）（過去）（丁寧）。
　　b.　形容詞（否定）（過去）（丁寧）。

ここに示した助動詞類の順序は大まかな傾向であり，否定，過去，丁寧の順序は入れ替わることもある。例えば，「行かなかったです」では構成要素が動詞 - 否定 - 過去 - 丁寧という順に並んでおり（11a）のパターンにそのまま合致するが，「行きませんでした」のように動詞-丁寧-否定-丁寧-過去という形になっているようなものも同時に数える。また，「ている」には短縮形「てる」も含めて数え，「ておる」は「ている」として併せて数えた。使役の「（さ）せる」は用例が多くないことなどからここでの調査対象からは外した。

分析は，（11a）のパターンの場合であれば，それに合致する個々の動詞のすべての用例において，「（ら）れる」，「ている」，否定，過去，丁寧の要素のそれぞれがどのような比率で含まれるかを調べるという方法によって行った。例えば，動詞「読む」の現れには次のようにさまざまな形のものがあるわけであるが，

(12)　a.　読む，読んだ，読まれる，読んでいた，読まれている，読まれていました，…
　　b.　読まない，読まなかった，読まれない，読んでいなかった，読まれていない，読まれて｛いなかったです／いませんでした｝，…

(12b）に属する用例が，（12a）または（12b）に属する用例全体においてどれだけの比率を占めるかを調べることによって，「読む」の否定に関わる有標率を計算するということである。

以上の方法による分析は厳密には 4.5 で述べるようにいくつかの問題があるが，それでも内省や小規模な資料では得られない情報をコーパスは提供してくれる。

4.3. 動詞の有標率

例えば，コーパスの分析によって得られる，動詞の「(ら) れる」に関わる有標率についての統計の一部を示せば表 1 の通りである。(11a) に示したパターンに合致する用例が Web コーパス 50 億字（10 ギガバイト）分中に 1,000 例以上ある動詞のうち，「(ら) れる」を伴う用例の比率の高い動詞上位 10 語を挙げている（以後の例もこれに準じて示す)。

表 1 「(ら) れる」形の比率の高い動詞

	無標用例数	有標用例数	有標率 (%)
見受ける	511	8,113	94.1
任す	233	1,541	86.9
強いる	471	2,613	84.7
親しむ	374	1,800	82.8
見舞う	451	2,017	81.7
癒す	1,197	5,130	81.1
締め切る	1,504	5,805	79.4
引き込む	450	1,699	79.1
思い遣る	230	781	77.3
騙す	500	1,522	75.3

「(ら) れる」には受け身以外の用法もあるが，リストの限りでは自発と考えられる「思い遣られる」を除き，表 1 はおおむね各動詞の受け身の比率を表していると見てよさそうである。

こうした情報を辞書の記述にどう反映させるかは辞書編集の作業に際して個別に検討する必要があるが，手作業の労力と時間を費やすこともなくすべての動詞に関する有標率の統計が得られるところがコーパスによる分析の大きな強みである。

表 1 には有標率の特に高い動詞を示したが，理屈のうえでは，有標率の低い（無標率の高い）動詞やさらには中間的な程度の有標率の動詞に関する情報も意味を持ち得る。ただ，実際には，有標率の低い動詞に着目する意味は大きくない。と言うのは，自動詞はもちろんのこと，他動詞の多くはコーパスの分析によれば「(ら) れる」の有標率が非常に低いからである。例えば，「買う」「習う」「誇る」「待つ」「持つ」などの動詞の有標率はいずれも 0.1 ～ 0.2% 程度にとどまる。「(ら) れる」以外の要素についても状況は同様で，例えば否定について言えば，大多数の動詞の否定に関する有標率は 1% 未満ないし数パーセントにとどまる。

「～ている」形の比率の特に高い動詞は表 2 の通りである。

表 2 「〜ている」形の比率の高い動詞

	無標用例数	有標用例数	有標率（%）
似る	651	23,801	97.3
持ち合わせる	113	2,090	94.9
秘める	122	2,252	94.9
潜む	154	1,701	91.7
満ちる	541	4,870	90.0
備わる	167	1,431	89.5
間違う	940	7,638	89.0
満ち溢れる	153	1,232	89.0
載る	2,146	15,631	87.9
待ち構える	191	1,225	86.5

「〜ている」にはよく知られた二義性があるので，表 2 はその点に注意して受け止める必要がある。

冗長を避けるために，否定，過去，丁寧の各要素に関する有標率については用例数を含む統計を表の形で掲げることは控える。まず，否定について言えば，否定形の比率の特に高い動詞としては「見当たる（99.8%）」「計り知れる（99.6%）」「構う（99.5%）」「知れる（99.5%）」「堪る（94.7%）」「差し支える（94.3%）」「足りる（90.2%）」「済む（81.0%）」「敵う（74.4%）」「絶える（74.4%）」などがある。ただし，「構う」は「かまわない」，「知る」は「〜かも知れない」，「堪る」は「（〜したくて，〜されては，…）たまらない」，「済む」は謝罪の「すまない」といった具合に固定的な言い回しの存在が有標性を引き上げるものが多く，そうしたものを除外すれば統計の上位に来る動詞は入れ替わることになる。また，「差し支える」の有標度が高くなっているのは「差し支えない」の誤解析によるものと見られる [8]。

過去形の用例の多い動詞としては，「締め切る（94.3%）」「漕ぎ着ける（86.7%）」「寄り付く（81.7%）」「亡くなる（80.6%）」「閃く（79.4%）」「見付ける（79.3%）」「言い渡す（79.0%）」「逸れる（77.1%）」「驚く（77.0%）」「思い立つ（76.9%）」などがある。「寄り付く」は株式用語である。

丁寧体の比率の高い動詞としては，「致す（99.7%）」「申し受ける（99.4%）」「御座る（99.1%）」「存ずる（98.6%）」「承る（98.6%）」「締め切る（97.7%）」「申し上げる（96.7%）」「禁ずる（95.0%）」「申す（93.2%）」「頂く（91.1%）」などがある。こ

[8]「差し支えない」は伝統的な品詞分類で言えば名詞 - 形容詞とも動詞 - 助動詞とも理解できる曖昧表現で，しかし，多くの場合前者の解釈を取るべきだと思われるが，MeCab + UniDic では一律に後者として解析される。なお，念のため説明を加えておけば，「差し支えない」が曖昧であることを理解するためには，それを「ません」を含む丁寧体にしてみればよい。名詞 - 形容詞の「差し支えない」は「差し支えありません」（「差し支えございません」）となるのに対し，動詞 - 助動詞の「差し支えない」は「差し支えません」となる。また，前者に助詞をはさめば「差し支えはない」などのようになるが，後者では「差し支えはしない」のようになる。

れらのうち，「締め切る」や「禁ずる」（これは MeCab + UniDic が与える終止形であるが，私見では下一段活用の「禁じる」としたほうがよい）は敬語と関係のない動詞であるという点で，他と比べて異質である。これはインターネット上のサイトに「募集を締め切りました」「無断転用を禁じます」のような告知が書かれていることが多いことによる。すぐ上で「締め切る」の過去の有標率が高いことを見たが，それも同じ理由によるものと考えられる。

4.4. 形容詞の有標率

コーパスの分析によれば，形容詞の有標率は動詞の場合に比べて全般に低い。(11b)のパターンに含まれる否定，過去，丁寧のいずれの要素についてもそうである。

否定に関する有標率の高い形容詞は表 3 の通りである。

表 3　否定形の比率の高い形容詞

	無標用例数	有標用例数	有標率（%）
難い	266	1,445	84.5
好ましい	1,384	590	29.9
少ない	32,503	9,908	23.4
珍しい	6,232	1,885	23.2
相応しい	1,982	475	19.3
悪い	36,666	5,820	13.7
宜しい	8,815	1,154	11.6
正しい	6,959	857	11.0
可笑しい	19,476	2,060	9.6
詳しい	2,811	180	6.0

「想像に難くない」という慣用表現のために有標率が高くなっている「難い」を除けば，すべて 30% 未満の有標率にとどまっている。形容詞には互いの否定的な概念――論理学の用語を用いてより正確に言えば矛盾ないし反対の概念――を表す関係にある対が揃っているものが少なくない。このため，形容詞「大きい」を例に取れば，対になる「小さい」が存在することの結果として，「大きくない」という否定表現の出番は多くない。このことが形容詞の否定の有標率を低くしている一因であると考えられる。

過去に関する有標率の高い形容詞は表 4 の通りである。ここでも，有標率は高いものでも約 50% にとどまっており，動詞の場合とは様相が異なる。これは，動詞と形容詞のあいだの時間的な性質の違いを反映したものであろう。

表 4　過去形の比率の高い形容詞

	無標用例数	有標用例数	有標率（%）
楽しい	24,826	26,578	51.7
美味しい	22,692	23,709	51.1
面白い	46,235	28,562	38.2
温かい	908	545	37.5
嬉しい	65,317	30,735	32.0
微笑ましい	1,388	598	30.1
心地良い	3,339	1,376	29.2
忙しい	4,922	2,012	29.0
眩しい	1,487	586	28.3
寒い	8,617	3,242	27.3

丁寧体の比率の高い形容詞には，「嬉しい（59.5%）」「待ち遠しい（59.2%）」「心強い（56.2%）」「有り難い（49.4%）」「美味しい（48.1%）」「羨ましい（45.3%）」「楽しい（44.4%）」「寂しい（34.0%）」「辛い（32.6%）」「微笑ましい（32.3%）」などがある。ただ，形容詞には「暑いです。」のような形での言い切りを不自然に感じる人が多いというよく知られた事実があり，そしてまた，「〜暑いですね。」「暑いですから」のように終助詞や接続助詞が後続すればその不自然さが解消するといったことがあるので，（11b）のパターンに合致する形容詞の用例だけに基づいて丁寧に関する有標率を計算して考えることに無理がある。

4.5. 問題点

ここで試した述語の有標率の調査にいくつかの問題があり得ることは，以上の議論の過程でも触れることがあった。それらを含めて問題点をあらためて整理して述べれば以下の通りである。

まず，具体的な分析方法に関わるものとして，対象とする用例の範囲を文末の言い切り位置に限定することの問題がある。4.4. の最後で形容詞の丁寧体での言い切りに関わる問題に触れたが，用言と助動詞類の組み合わせの関係が文内の位置に依存する場合はほかにもある。例えば，「曲がった道」とは言っても「道が曲がった」とは言わない。また，4.1. で「大別」は受け身の形での用例の比率が約 92% と高いことに触れたが，それは句点の直前の文末位置に関する調査の結果であり，「〜は 2 つに大別することができる」のような埋込み文での使用を含めると，受け身の比率は多少とも下がる。こうしたことから，（11）に示したパターンを適宜拡張して調査したり，あるいは，非文末のパターンを独立に設定して別途調査したりする必要がある。

次に，コーパスに基づく分析の限界として，多義的な表現の語義・用法を区別して調査することがむずかしいという問題がある。例えば，4.1. で触れたように「参る」は丁寧体での使用率が非常に高いが，それは「行く」「来る」で言い換え得る

普通の移動を表す表現の場合のことであって，寺社への参詣を表すときには「きのう寺に参った」のような普通体の表現も一般に使われる。そのような語義・用法の区別を扱うコーパス処理は望みがたい。文法的な要素の多義についても同様で，「（ら）れる」が自発・可能・受身・尊敬のどの意味を表すのか，「～ている」が完了・継続のどちらの意味を表すのかといった情報を分析に組み込むことはむずかしい。この種の問題に関して有効な解決方法は存在せず，コーパスの機械処理に付随する制約のもとで工夫しながらコロケーションの分析を進めるという形を取らざるを得ない。

コロケーションの分析に他の種類のコーパスでなくWebコーパスを使うことに起因する問題は以上のような問題に比べれば限定的である。2.2. ではWebコーパスにおける複製データの重複出現の問題に触れたが，4.3. で見た，「締め切る」や「禁じる」の用例において丁寧体の比率が予想外の高さを示しているという問題は複製データの重出によるものではなく，インターネット上における特定の言い回しの頻出によるものである。しかし，Webコーパスにそうした問題があるにしても，コロケーションの安定した分析のために必要となる規模のテキストをインターネット文書に頼ることなく用意することは不可能である。したがって，特定の表現の重出や頻出のもたらす影響の可能性に注意しつつWebコーパスを利用するのが現実的な対処であることになる。

5. 類義の慣用句型の用法

日本語文法の研究で「複合辞」の名で呼ばれる概念がある。筆者の理解によれば，複合辞の概念には理論上の問題が多く，しかし，日本語教育や自然言語処理といった応用の方面では現に重要な意味を持つ。「複合辞」という名称も必ずしも適切とは言えないので，代わりに「慣用句型」ないし単に「句型」という表現を用いることにする[9]。以下では，類義的な慣用句型の用法の違いを，コロケーション（ないしコリゲーション）の観点から考える。

5.1. 問題の所在

類義の慣用句型の意味・用法は往々にして重なる部分が大きく，各句型間の微妙な差を明らかにすることは容易ではない。

一般に，類義表現の意味・用法の違いを明らかにするための有効な手段の1つは，それぞれの表現がどのような表現と共起するかを観察することである。ここでは，慣用句型の共起傾向をWebコーパスに基づいて調査することにより，各句型の性格の特徴を考えるための手がかりを得る可能性を探る。

[9] 複合辞の概念に関する筆者の見解については拙論（2002）および拙論（2008）を参照願いたい。

5.2.「〜なければ（〜なきゃ）」「〜ないと」「〜なくては（〜なくちゃ）」「〜ねば」

この標題に示す 4 つの否定条件の句型はいずれも同じように使われるものとして説明されることが多い。「〜ねば」は古めかしくて文体的に堅く，「〜なきゃ」「〜なくちゃ」は口頭語的性格が強いといった母語話者にとっては自明の文体的特徴が補足的に述べられる程度である。

初級の外国人学習者向けの単純化した説明としてはそれでよいであろうが，4 つの句型の違いが単に文体的なものにとどまらないことはコーパスを使うまでもなく内省に基づく考察によって知られる。以後，「〜ねば」以外の 3 つの句型を一括するときは「ナイ系」とする。

第 1 に，周知の通り「ない」は動詞「ある」の否定形としての一面を持つ。すなわち，文法規則上は「あらない」となるはずのところが単に「ない」と表現される。他方，「ねば」に含まれる「ぬ」はそのような例外的性格を持たず，「ある」との組合せは規則的に「あらぬ」と表現される。

このことから，ナイ系の否定条件では可能な次のような表現において，「なければ」その他を単純に「ねば」で置き換えることはできないことになる。

(13)　a.　{事実／自由／こう} で {なければ／ないと／なくては}（コピュラ述語）
　　b.　楽しく {なければ／ないと／なくては}（形容詞述語）
　　c.　小銭が {なければ／ないと／なくては}（存在表現）
　　d.　消印が押して {なければ／ないと／なくては}（「〜てある」）

同様のことを「〜ねば」を用いて言うには，「ある」を加えて「あらねば」とする必要がある。このように，「〜ねば」とナイ系は文体的な相違以前に文法上明確に異なるところがある。

第 2 に，存在を表す動詞「いる」と「ぬ」の組合せは普通体においては避けられる傾向が強く，「いぬ」という形の使用はまれである。このことを反映して，「いねば」という言い回しの用例は非常に少ない。ちなみに，「おる」については，Web コーパスの調査によれば，「ない」「ぬ」との結合に関してそうした大きな偏りは見られない。

第 3 に，意味の面での違いも指摘できる。「ば」「と」「ては」の表す条件の違い（その詳細をここで論じる用意はないが）に応じて，「〜なければ（〜なきゃ）」「〜ねば」対「〜ないと」「〜なくては（〜なくちゃ）」のあいだに違いがあると思われる。そのことを示す対比を一例挙げれば次の通りである。

(14)　a.　大雨さえ {降らなければ／降らねば} 出発する。
　　b.　大雨さえ {?降らないと／??降らなくては} 出発する。

さて，以上のようなことに加えて，4 つの句型のあいだには Web コーパスの分析によって初めて知られるコロケーション面での違いもある。期待したほど多くのことが明らかになったわけではないが，第 1 に，「〜ねば」は「廃絶されねば」「究

明されねば」などのように受け身の述語に続く傾向が強い。複合サ変動詞に限って言えば「せねば」と「されねば」の用例数の比率は 100:31（用例数はそれぞれ 9,722, 2,990）であった（「されねば」の古い形である「せられねば」55 例もあるが，それを計算に入れても結果には影響しない）。ナイ系で受け身率の最も多いのは「〜なければ」であったが,「しなければ」「されなければ」の用例数の比率は 100:7.5（用例数 154,100，11,563）であり，「〜ねば」の受け身傾向は明確である。

また，「〜なきゃ」「〜なくちゃ」については，日常的な行動に関わる次のような語連鎖に続く用例が特に多い。

(15)　勉強し，がんばら，食べ，我慢し，〜してあげ，〜もし（例:掃除もしなきゃ）

以上のような分析結果を知った後で, それが「〜ねば」あるいは「〜なきゃ」「〜なくちゃ」の文体的な特徴を反映したものだと考えて納得することは容易であるが, 当該の句型がどのような述語と共起する傾向があるかを内省に基づく考察によって具体的に言い当てることは困難であろう。

5.3.「〜に比べて」「〜に比べれば」

これらの慣用句型の違いも内省では捉えにくいものである。関連の句型としてほかに「〜に比べると」や「〜に比べたら」もあるが，分析によればそれぞれ「〜に比べて」「〜に比べれば」に近く，ここでは標題の 2 句型だけを取り上げる。

Web コーパスから得られる用例をコロケーションの観点から分析してみると，両者の用法の傾向に違いのあることが容易に知られる。

まず，「〜に比べて」の後続文脈に最もよく現れる述語は「高い」「低い」「多い」「少ない」などの形容詞であるのに対し，「〜に比べれば」の後続文脈に最もよく現れる語連鎖には意味的な偏りがあり，「（まだ）ましだ」「大したことはない」「微々たるものだ」「かわいいものだ」「楽だ」などが最上位を占める。このことから,「〜に比べて」が一般的な比較を表すのに対し，「〜に比べれば」は「ましだ」その他の述語とともに使われて，《問題が軽微である》,《困難の度合いが低い》といったことを表現するのに使われる傾向が強いことが分かる。

「〜に比べれば」は,「ましだ」の類の述語とともに使われる場合だけでなく, ほかの一般的な述語との組合せで使われるときも，その表現意図は二者の単純な比較にはなく，便宜上かりに命名すれば「条件付きの断定」とでも呼ぶべきものであることが多い。「条件付きの断定」とは,《無条件にはそうは言えないが，“〜との対比において”という条件のもとではそう言える》ということを表すということである。肯定される内容は，好ましいことがらであることが多い。具体例に即して言えば,

(16)　零下 60 度にも達するシベリアの冬に比べれば日本の冬は暖かい。

という文は，日本の冬が暖かいと無条件に言えるわけではないが，さらに寒いシベリアの冬との対比においてならば暖かいと言えるということを表している。

「〜に比べて」と「〜に比べれば」の先行文脈によく現れる表現はおおむね一致する。いずれの慣用句型も「昔」「以前」「日本」「東京」「欧米」「〜のとき」「〜する場合」のような時間・場所・場合などの表現に後続する用例が特に多い。そのような中で特徴的と言えるのは，「〜に比べれば」についてのみ，「{〜の／〜する} 痛み」「{〜の／〜する} 苦しみ」という表現に続く用例が多いということである。

(17)　麻酔無しで歯を削られる痛みに比べれば，レーザー使用中の痛みも，それ程でもない。

5.4.「〜を通して」「〜を通じて」

この 2 つの慣用句型も違いが分かりにくい。文字表記上でこそ 1 文字の差――さらに言えば，濁点の有無の差――しかないが，一方は「とおす」という和語の単純動詞を含むのに対して，他方は「つうじる」という一字漢語を語幹とする複合サ変動詞を含み，語源的に無縁の関係にある。それにもかかわらず，両句型の意味・用法は大きく重なる。

筆者自身この 2 句型には苦手意識があり（違いが分からず選択に迷う），過去に発表した拙論を調べてみても，次の通り同様の文脈で両方の句型を無差別に使っている。

(18)　a.　コーパス（電子媒体の言語資料）について，具体的な事例研究を通してその利用の価値を明らかにし，（拙論「コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発」『人工知能学会誌』第 24 巻第 5 号，2009 年）

b.　今後さまざまな事例研究を通じて，研究のテーマや目的，使用する資料に即して具体的に検証されるべき重要な課題だと言える。（拙論「サ変動詞の活用のゆれについて・続」『日本語科学』第 25 号，2009 年）

c.　全期間を通して「な」の比率が概ね 10% 以下であり，（拙論「大規模な電子資料に見る現代日本語の動態」『待兼山論叢』第 42 号，2008 年）

d.　五段化率が全期間を通じて 10% 未満のものを除く（拙論「サ変動詞の活用のゆれについて・続」『日本語科学』第 25 号，2009 年）

「〜を通して」「〜を通じて」ともに，(18a)，(18b) のように手段・経路を表したり，(18c)，(18d) のように期間を表したりする。したがって，理想的にはそうした用法上の区別も考慮に入れたいところであるが，残念ながらコーパスの機械的な分析ではそれは望めない。そこで，用法上の区別は考慮しないで，各句型と名詞の共起関係をコロケーションの見地から分析してグラフ化したのが図 1 である。この図は，Web コーパス 500 億字（100 ギガバイト）分において両句型の用例数の合計が 1,000 以上（かつ両句型の用例数がともに 7,000 以下）である名詞について，それぞれの名詞が各句型とどのような頻度で共起するかを示したものである。

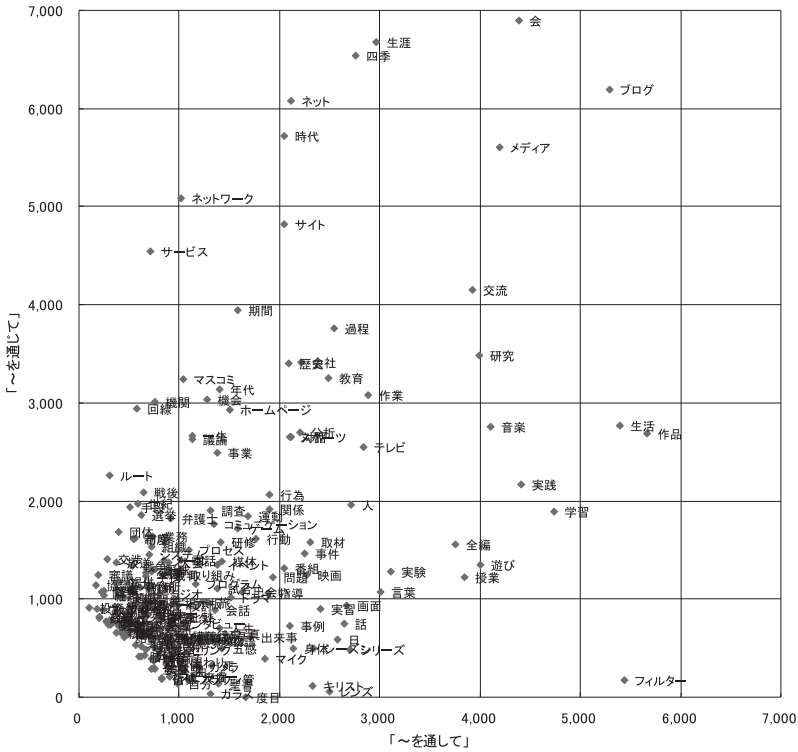

図１　「～を通して」「～を通じて」の用例数の相関

この図の観察を通して――あるいは通じて――，両句型のある程度の違いはおぼろげに見えてくる。「ルート」「ネット」のように情報伝達の経路全体を表す名詞の場合は「～を通じて」が多く使われ，「フィルター」「レンズ」「マイク」「画面」のように（多く受信方向の）伝達経路上に介在する透過物を表す名詞の場合は「～を通して」が多く使われるといったことが分かる。

いずれにしても両句型の違いの全貌は明らかと言うにはほど遠いが，いろいろと分析を試みてみたところによれば，「～を通して」「～を通じて」の違いをもっぱら先行する名詞，すなわち precollocate の観点から見ていたのでは不十分で，circumcollocate の観点からの検討も必要である。図１の限りでは，「テレビ」「ブログ」などの名詞には「～を通して」「～を通じて」の両句型が用例頻度の点で特に偏りなく組み合わせられる（「テレビ」は「～を通して」2,839 例，「～を通じて」2,551 例，「ブログ」は「～を通して」5,289 例，「～を通じて」6,195 例）。しかし，それは両句型が同じように使われているということを意味しない。すなわち，「～を通して」と「～を通じて」とでは後続する表現の内容に差が見られるのである。

(19)　a.　「テレビを通して」に高頻度で後続する特徴的な語連鎖：
　　見る，伝わって来る，知った
　　「テレビを通じて」に高頻度で後続する特徴的な語連鎖：
　　国民に，全国に，声明を，発表，報道
　　b.　「ブログを通して」に高頻度で後続する特徴的な語連鎖：
　　たくさんの，色々な，方，人，お友達，知り合った，出会い，仲良く
　　「ブログを通じて」に高頻度で後続する特徴的な語連鎖：
　　意見，発信する

(19) に見るように，「テレビを通じて」「ブログを通じて」は情報の発信，「テレビを通して」は情報の受容，「ブログを通して」はブログを媒介とした人々との出会いを表すときによく使われているということが傾向として確認される。

5.2. ～ 5.4. で取り上げたどの事例に関しても，コロケーションの観点からのコーパスの分析によって得られる情報を慣用句型の用法分析に生かす可能性を探ってみたにとどまるが，内省に基づく考察では得がたい知見をコーパスがもたらしてくれることはもはや十分に明らかであろう。

6. おわりに

以上，日本語の辞書記述の改善やコロケーション辞典の編集にコーパスをどのように生かすことができるかという応用的，実用的な問題意識に基づいて行った考察について述べた。

日本語の研究におけるコーパスの整備と活用は他の主要言語に比べて従来大きな遅れを取ってきたが，国立国語研究所を中心とする『現代日本語書き言葉均衡コーパス（the Balanced Corpus of Contemporary Written Japanese，略称 BCCWJ)』の構築（2010 年度末に完成予定）などにより状況は急速に変わりつつある。今後コーパスが日本語研究においてますます大きな意味を持つようになり，研究の質を高めるとともに，可能な研究の領域を広げていくことは確実であろう。

参　照　文　献

田野村忠温（2002）「辞と複合辞」玉村文郎（編）『日本語学と言語学』: 49–60．東京：明治書院.
田野村忠温（2008）「複合辞の本性について――その構成と単位性――」児玉一宏・小山哲春（編）『言葉と認知のメカニズム　山梨正明教授還暦記念論文集』: 489–497．東京：ひつじ書房.
田野村忠温（2009）「コーパスからのコロケーション情報抽出――分析手法の検討とコロケーション辞典項目の試作――」『阪大日本語研究』21: 21–41，大阪大学大学院文学研究科日本語学講座.
Tanomura, Tadaharu (2010, in press) Retrieving collocational information from Japanese corpora: Its methods and the notion of “circumcollocate”. In: Peter Grzybek, Emmerich Kelih and Ján Mačutek (eds.) *Text and Language: Structures·Functions·Interrelations*, 213–222. Wien: Praesens Verlag.
服部匡（2010）「大きさを表す形容詞類の選択傾向とその推移――『○○性が～』などの場合――」田野村忠温・服部匡・杉本武・石井正彦『コーパスを用いた日本語研究の精密化と新しい研究領域・手法の開発 IV』: 51–62．科学研究費補助金特定領域研究「日本語コーパス」日本語学班.

堀正広・浮網茂信・西村秀夫・小迫勝・前川喜久雄（2009）『コロケーションの通時的研究　英語・日本語研究の新たな試み』東京：ひつじ書房.

執筆者連絡先：
〒 560-8532 大阪府豊中市待兼山町 1-5
大阪大学大学院文学研究科

［受領日　2010 年 5 月 10 日
最終原稿受理日　2010 年 6 月 5 日］

Abstract

Japanese Corpora and Their Lexicographic Applications, with Special Emphasis on Collocation

Tadaharu Tanomura
Osaka University

Although Japanese has been lagging behind the other major languages of the world in the utilization of electronic corpora in linguistic studies, the situation is changing rapidly due to several factors including, notably, the ongoing construction of a balanced corpus of the language at the National Institute for Japanese Language and Linguistics.

This paper focuses on collocation, a linguistic phenomenon which can be analyzed reliably only by using large corpora, and explores the possible roles which corpora may play in the compilation of a dictionary of Japanese, be it a dictionary of an ordinary kind or a collocational dictionary. The three collocational aspects of Japanese examined by way of corpus analysis are: 1) the concept of 'circumcollocate', 2) the degree of markedness of verbs and adjectives, and 3) the semantic differences between synonymous idiomatic grammatical phrases. The paper will demonstrate the ways in which corpora may have lexicographic significance in each of those domains.

A large corpus is required for the retrieval of collocational information. The paper uses a Web corpus, constructed by the author in 2008, which consists of approximately 75 billion characters. This is equivalent to 150 gigabytes in file size, or three to four hundred thousand Japanese novel books of average size.