The University of Osaka
Institutional Knowledge Archive

Title	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
Author(s)	田野村, 忠温
Citation	阪大日本語研究. 2009, 21, p. 21-41
Version Type	VoR
URL	https://doi.org/10.18910/3790
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

『阪大日本語研究』21（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—

Retrieving collocational information from Japanese corpora:
An attempt towards the creation of a dictionary of collocations

田野村　忠温
TANOMURA Tadaharu

キーワード：日本語コーパス、コロケーション情報抽出、日本語コロケーション辞典

要旨

この数年来、日本語のコーパスをめぐる状況は急速な展開を見せている。コーパスは今後日本語の研究に対してますますその関わりを深めていくことであろう。

コーパスの活用が見込まれる未開拓の研究領域の１つにコロケーションの調査・分析がある。この小論では、日本語のコーパスからコロケーション情報を抽出するための分析手法を考案・検討するとともに、コロケーション研究の成果物の重要な一形態となるはずの日本語コロケーション辞典の項目を試作する。

１．はじめに——前提的考察

大規模なコーパスの有効な活用が見込まれる未開拓の研究領域の１つにコロケーションの現象の調査・分析がある。しかし、分かち書きをしない日本語においては、そもそもコーパスからどのようにしてコロケーションに関わる情報を抽出すればよいのかというところから考察を始める必要がある。

この小論においては、日本語のコーパスからコロケーション情報を抽出するためのいくつかの手法の可能性を具体的な事例の分析を通して比較・検討するとともに、そうした方法によってコーパスから得られる情報に基づいて日本語コロケーション辞典の項目を試作する。

ここでは、この小論での議論全体の前提となる、コロケーションの概念とその研究に関する筆者なりの理解を一通り述べる。

1. 1. コロケーションの概念について

コロケーションは一般に語の習慣的な共起関係と説明されるが、そうした理解と関心で物事を考えていけば自ずと語と句、句と句といったより広い意味での共起関係も対象としないわけにはいかなくなる。したがって、コロケーションを語ではなく言語表現全般（語、語の連続、句、節など）のあいだに観察される習慣的な共起関係と理解すべきであることをまず明確にしておく。

コロケーションに関する一般的な慣用に従って話し書くことで言語表現が自然なものになり、慣用に従わないで話し書くことで言語表現が誤りとまでは言えなくとも不自然なものになり得る。そうした意味でコロケーションの研究には特に言語の運用に関わる方面での応用的価値が期待される。

もっとも、コロケーションは所詮単なる表現間の距離と使用頻度とに基づく概念であり、言語を理論的に捉える感覚からすれば価値も魅力も乏しい。筆者の理解によれば、コロケーションの研究は何かへの利用、応用が達成されて初めて価値を持つものであり、コロケーションをそれ自体として研究することにはあまり意味がない。そうしたことから、コロケーションの概念を厳密に規定する必要はなく、上述のような緩やかな理解にとどめておけばそれで十分である。

1. 2. コロケーションに関する情報の獲得は何の役に立つのか？

コロケーションに関する情報が得られたとき、それには２種類の用途が考えられる。

１つは、語義の精密な分析・記述のための考察材料としての用途である。特に、類義的な表現の意味の差を考える際には、当該の表現がどのような文脈でよく生起するかを観察することが大きな手がかりとなる。

もう１つは、辞書の編集、外国語の教育・学習、母国語の運用といった実用的な分野での利用である。コロケーション情報は、辞書編集においては語義の記述や例文の選択に役立ち、言語の教育・学習の場面ではより自然な教材の開発などに有効に利用され得るはずである。コロケーションの研究は利用、応用が達成されて初めて価値を持つと上で述べたが、筆者が目標として想定するまとまった形での成果物には日本語コロケーション辞典や日本語コロケーション・例文検索システムがある。この小論の最後では、そうした認識に基づいてコロケーション辞典の項目の試作を行う。

1. 3. コロケーションに関する情報がすべて有用か？

言語形式の共起関係に関する情報は広範で雑多であり、そのすべてが有用であるわけで

はない。例えば、「大学」の直後によく現れる語（形態素）をYahoo!知恵袋ベータ版データ――これを含む、この小論で使用するコーパスについては後述する――に基づいて調べると、次のような結果が得られる。

（表 1）「大学」の直後によく現れる語（形態素）――Yahoo! 知恵袋ベータ版データの場合

の、生、に、で、を、は、院、病院、受験、が、時代、卒業、へ、も、です、～

しかし、このような情報が得られてもほとんど無用であることは明らかである。

無用の、ないし、有用性の低いコロケーション情報にはいくつかの種類のものがある。

まず、文法の知識から容易に分かる情報がある。例えば、名詞には格助詞が、動詞には助動詞や接続助詞がしばしば後接するといったことは調べるまでもない当然の事実である。もっとも、細かく見ていけば、例えば動詞と助動詞の組み合わせにもいろいろと個別の制約や傾向――例えば、ある動詞はほとんど常に否定の文脈で使われるとか、受身の形で使われやすいといったこと――があり、そうしたことは文法の知識からは分からない。

次に、語義の知識から容易に分かることも調査に値しない。「バイオリン」「フルート」「ティンパニ」のそれぞれに、楽器の演奏を表す「弾く」「吹く」「叩く」のどれが通常続くかは自明である。もっとも、ここでも例えば「シンバル」の演奏をどんな動詞で表すのかといった疑問はあり得るので、「弾く」「吹く」「叩く」の語義が一通り分かっていればそれで万全ということではない。

また、何を表現する機会が多いかという事実の反映に過ぎない情報も価値が乏しい。「飲む」という動詞の目的語として各種の飲み物を表す表現のうちどれが多く現れるかが明らかになっても、実際に話し書く場面ではそのような傾向とは関わりなく、発話意図に即した表現を選ぶ必要がある。ただし、これについても、「飲む」という動詞で表される行為の範囲は非母語話者にとって自明であるわけではないから、「飲む」の目的語として可能な表現の範囲を調べること自体が無意味であるわけではない。例えば、「飲む」という行為の対象は飲み物には限らず、「スープを飲む」「薬を飲む」とも言うし、「息を飲む」「言葉を飲む」といった慣用表現もある。また、辞書や日本語教科書の例文を用意する際には、同類の表現なら使用頻度の高いものを選ぶようにするのが望ましいであろう。

コロケーション情報の抽出においては、そうした無用の情報を避け、有用性の高い情報の獲得を目指すことが課題となる。

1. 4. コロケーション情報をいかにして獲得するか？

コロケーションに関する正確な情報を得るには、大量の用例を収集して分析することが必要である。そして、特定の語句の一定数の用例を見つけるのに必要な言語資料の量に比べて、その語句が他のさまざまな語句と共起した用例をそれと同数ずつ見つけるのに必要とされる言語資料の量ははるかに大きくなる。このため、コロケーションの本格的な調査は大規模なコーパスを使って初めて可能になる。内省は正確なコロケーション情報の獲得手段として絶望的に非力であり、紙媒体の言語資料の手作業による分析は時間と労力の点で非現実的である。

しかし、コーパスを使いさえすれば有用なコロケーション情報が容易に得られるということではない。日本語のコーパスに対してどのような処理を施せば有用なコロケーション情報を獲得することができるかという問いに対する答は知られていない。したがって、コーパスからコロケーション情報を抽出するための有効な手法を模索・案出することがこの小論の中心的な課題となる。

分析方法についての具体的な検討に話を進める前に、コーパスからのコロケーション情報抽出に関わる筆者の一般的な認識をあらかじめ述べておけば次の通りである。

第１に、コーパスから獲得できる情報は、我々が真に得たいコロケーション情報のごく粗い近似でしかない。コーパスの調査は基本的に表面的な文字レベルの分析であり、それを大きく上回ることはない。統語的、意味的な情報の利用の可能性はきわめて限定的である。コーパスに基づくコロケーションの研究は、たとえて言えば、水面の波を、それを生じさせている水面下の流体の運きに目を向けることをしないで観察し分析するようなものである。

第２に、そのため、機械処理によってコロケーションに関する結論的な知見を直接に得ることは望めない。言い換えれば、コーパスを入力として与えれば完成した形の分析結果を出力してくれるソフトウェアを作ることはまず不可能である。コーパスの機械的な分析によって得られるコロケーション情報はたいてい多くの無用の情報を含む。語義分析、辞書編集、言語の教育・学習などへの応用のためにそこから有用な情報をより分ける作業は人手に頼らざるを得ない。

第３に、コロケーション情報の評価において統計的な厳密さを重視する意味は乏しい。上述の通り、コロケーションは理論的な価値の乏しい概念であり、しかも、人手による情報の取捨を前提としてもいるので、コロケーション情報を統計的に厳密に評価してもしかたがない。また、コーパスから得られるコロケーション情報は資料ごとに大きな変動を見せ、その意味でも特定のコーパスから得られる統計的情報には絶対的な価値を見出しがた

い。さらに、そういった問題とは独立に、辞書編集や教科書の作成といった実用的な応用の場面においては、言語使用の実態をもっぱら忠実に記述することが目的であるわけではないから、コーパスから得られる情報を絶対視する必要もなければそうする意味もない。コーパスから得られる情報を適宜参考にするという姿勢でよいのであり、それこそが正しい考え方でもある。

コーパスから得られるコロケーション情報が資料ごとに異なることを示す単純な例として、4種類のコーパスにおいて「西瓜」という名詞に高頻度で後続する動詞の上位各5語を次の表に掲げる。各動詞に添えた数値は用例数である。片仮名表記の「スイカ」も同時に調査対象としている。

（表2）4種類のコーパスにおいて「西瓜」によく後続する動詞

Webコーパス：	食べる10315、買う2152、ある1528、切る1095、なる1092
Yahoo! 知恵袋ベータ版データ：	食べる275、出る80、取りこぼす71、出す65、引く57
朝日新聞記事11年分：	食べる32、作る10、切る8、売る8、割る6
BCCWJ2008の書籍データ：	食べる8、盗む7、切る5、持つ4、できる3

動詞「食べる」が第1位を占めることはすべてのコーパスに共通しているが、第2位以下の動詞はコーパスによって大きく異なっている。ちなみに、Yahoo!知恵袋ベータ版データでの上位を占める「出る」「出す」はその用例の大半が出産に関する話題における「鼻からスイカが出る（スイカを出す）くらいの痛み」のような表現での使用であり、「取りこぼす」と「引く」はスロットマシンに関する話題での使用である。

引き続き2節で、この小論で調査・分析に用いるコーパスについて述べる。

2．使用するコーパス

コロケーションの調査に使うコーパスは大規模なものである必要があることから、この小論では筆者の試作によるWebコーパスを主として使用する。これはインターネット上の日本語文書を収集してコーパスとしたもので、そのデータ量は2008年10月現在で約150ギガ（＝1,500億）バイト、文字数にして約750億字である。文書数で言えば約1,500万件の文書の集積である[1)]。以下ではその一部である約10ギガバイト、約50億字分を用いる。これに加えて、ほかのコーパスも補助的に使用している。

次の（表3）で丸印を付けて示しているのが、ここで用いるコーパスである。その他のコーパス・資料はサイズの比較のために掲げている。データ量はShift-JISコードでのバイ

ト数である。

（表 3）使用するコーパスとその他のコーパス・資料

日本語データの種類	字数　（データ量）	小説との比率
小説単行本※1	20 万字	1
新潮文庫の 100 冊	2,000 万字　（40 メガバイト）	100
新聞記事 1 年分※2	6,000 万字　（120 メガバイト）	300
○BCCWJ2008 の書籍データ※3	7,300 万字　（146 メガバイト）	365
○朝日新聞記事 11 年分※4	1.8 億字　（361 メガバイト）	900
BCCWJ（構築中）※3	2.8 億字　（1 億語）	1,400
○Yahoo! 知恵袋ベータ版データ※5	16 億字　（3.2 ギガバイト）	8,000
国会会議録のデータ※6	35 億字　（7 ギガバイト）	18,000
◎自作 Web コーパス	750 億字　（150 ギガバイト）	375,000
Web 上の日本語文書※7	26 兆字　（52 テラバイト）	130,000,000

※1　手近にあった本から平均的と思われるサイズのものを選んだ。縦組みで、43 字 ×18 行 ×270 頁。かけ合わせると総文字数 208,980 字ということになるが、ページに空白部分もあるので実際の文字数はそれより多少少ない。奇しくも「新潮文庫の 100 冊」のちょうど 100 分の 1 に相当する量であった。

※2　新聞の種類や年によってサイズは異なる。最近の朝日新聞、毎日新聞の 1 年分の記事は 150 メガバイト程度になっている。

※3　BCCWJ は現在国立国語研究所において構築が進められている「現代日本語書き言葉均衡コーパス」の英語名の略称。BCCWJ2008 はその試用・評価のために特定領域研究「日本語コーパス」の組織内部で配布されているデータの 2008 年版を指す。BCCWJ は Yahoo! 知恵袋ベータ版や国会会議録のデータなども含むが、ここでは書籍のデータだけを用いる。

※4　朝日新聞 1987 ～ 1997 年の記事データ。

※5　Yahoo! 知恵袋の正式運用が開始する直前の試験段階（2004 年 4 月～ 2005 年 10 月）のデータ。国立情報学研究所から研究者に対して提供されている。

※6　国立国会図書館の Web サイトで公開されている。データ量は第 169 回通常国会（2008 年 1 ～ 6 月）までのもの。

※7　Web 上の日本語文書のうち、Yahoo!JAPAN のサーチエンジンが収集し検索対象としている文書。データ量は 2008 年 10 月時点の推計。

約150ギガバイトのWebコーパスのうち約10ギガバイトだけを用いるのは次のような理由による。すなわち、筆者の経験に基づく感覚として、コロケーションに関する統計を取ったとき、できれば着目するすべての用例数が 3 桁以上であってほしい。用例数が 1 桁では明らかに頼りなく、2 桁でも、特にその前半では、やはり不安が大きい。着目するすべての用例数が 3 桁以上ということは多くの場合最大の用例数は 4 桁以上であり、そこまでくれば統計は安定し、それ以上の用例を収集しても結果に大きな変化は生じにくい。もちろん、得られる用例数の点だけで考えれば利用可能なデータをすべて使えばよいわけであるが、現実の処理に際してはパソコンでの処理に要する時間の要素も考慮に入れる必要がある。ハードウェアの性能にもよるが、150ギガバイトのデータは単にディスクから読み

出すだけでも通常１時間以上かかる。そうしたことから、経験上、コロケーションの調査には10ギガバイト程度のデータを使うのが適当だとの判断を得た。言うまでもなく、これは大まかな目安に過ぎず、出現頻度の低い語句の場合やさらに詳しいコロケーション情報を得たい場合はより多くのデータを調査することが必要となる。また、今後パソコンの高性能化に伴って、無理なく処理できるデータの量は徐々に増えていくはずである。

３．コーパスからのコロケーション情報の抽出

さてこのたび、コーパスから有用なコロケーション情報を効率的に抽出する手法を求めて、いろいろな分析の可能性を模索してみた。有効性の確認できた方法は３種類に大別することができ、それらを共起語分析、共起語連鎖分析、共起文字連鎖分析と呼ぶ。以下では、それぞれの方法について順次述べる。

3. 1.　共起語分析――所与の語句とよく共起する語への着目

3. 1. 1.　分析の基本方針

第１の分析方法である共起語分析は、コーパスにおいて所与の語句の近傍にどのような語ないし形態素がよく現れるかを調べるものである。

ここで「語ないし形態素」というのは、形態素解析ソフトウェアMeCabが返す解析結果における分割単位のことであり、以後単に「語」と表現する。解析用の辞書としては、Windows版のMeCabの配布パッケージに含まれるIPA辞書を利用する[2)]。

ソフトウェアによる形態素解析には、解析の誤りや分析の不統一などの問題が常に付きまとうことを十分認識しておく必要がある。（表４）は、上記のソフトウェアと辞書の組み合わせでいくつかの複合的な表現を解析した場合にそれらがどのように解釈されるかを示したものである。

（表４）形態素解析ソフトウェアによる複合的な表現の解釈

大阪大学	大阪＋空港、大阪＋駅
気に入る	気＋に＋かける
調査＋によって	調査＋に＋よれ＋ば
不十分、不健康	不＋満足、不＋機嫌
自動的	機械＋的

左右に分けて配置してある表現のあいだに必ずしも文法上大きな差はないにもかかわ

らず、左半分の表現は解析において１単位（「調査によって」は２単位）と解釈され、右半分の表現は複数の単位（より多くの単位）として解釈されている。先にコロケーション情報の評価に統計的な厳密さを求める意味が乏しいことを述べ、そのいくつかの理由を挙げたが、形態素解析ソフトウェアに依存した処理を行う場合はその解析精度上の問題がもう１つの理由として加わることになる[3)]。

3. 1. 2. 分析手順の検討

先に（表１）として示した「大学」の直後によく現れる語のリストからも予想が付く通り、所与の語句の近傍に現れるすべての語を無差別に調査したのでは有用なコロケーション情報は得られない。そこで、

・所与の名詞にどのような動詞が後続するか
・所与の動詞にどのような名詞が先行するか
・所与の副詞にどのような動詞が後続するか
・所与の形容詞にどのような名詞が後接するか

などのように調査対象とする語句とその共起語の関係を限定し、それぞれの条件のもとで所与の語句の一定範囲の文脈にどのような語が高頻度で現れるかを調査する。その所与の語句のことを以後「中心語」と呼ぶことにする。正確には「中心語句」とすべきところであるが、表現の簡潔のために略記する。

分析に際しては、中心語とそれに直接関係しない語とのあいだに関係を認定してしまうことを避ける必要がある。例えば、「誘う」という動詞に先行する名詞を調べる場合、「太郎の友人を誘う」という表現において「太郎」と「誘う」が直接関係していると考えることはできない。そこで、「誘う」の前に名詞――今の例の場合、「友人」――が現れたらそれよりも前にある名詞は見ないといった処置が必要となる。そうした処置に完全を求めることはできないが、調査の種類ごとに可能な範囲で適宜処置を講じる。実際に行ったさまざまな個別の対処の詳細については省略に従う（以後においても同様とする）。

3. 1. 3. 分析事例と考察――所与の名詞に後続する動詞

手始めに、さまざまな名詞を中心語とし、それらに続く文脈中にどのような動詞がよく現れるかをWebコーパスを使って調べてみた。

次に示す（表５）は、「情熱」と「熱意」という類義的な名詞によく続く動詞上位各10語のリストである（動詞の前の数値は用例数）。下線を施してあるのは、２つの組の動詞のうちで互いに重ならないものの一部である。

（表 5）「情熱」「熱意」に後続する動詞

■情熱（…）［動詞］		■熱意（…）［動詞］	
1185	持つ	665	ある
741	傾ける	576	持つ
695	注ぐ	470	伝わる
688	ある	420	感じる
620	燃やす	213	もつ
536	感じる	161	奪う
375	もつ	148	負ける
262	伝わる	115	あふれる
227	失う	105	伝える
150	あふれる	97	示す

このリストから、「情熱」は「傾ける」「注ぐ」「燃やす」との共起に見るように、多くの場合内面の描写に用いられるのに対し、「熱意」は「伝わる」「伝える」「示す」「感じる」「負ける」などとの結び付きの高さに見るように、他者との関係において伝えられたり作用を及ぼしたりする面が「情熱」よりも強いことを容易に見て取ることができる。

（表 6 ）に示すのは「才能」と「能力」によく続く動詞である。

（表 6）「才能」「能力」に後続する動詞

■才能（…）［動詞］		■能力（…）［動詞］	
5672	ある	13275	ある
1519	持つ	10888	持つ
499	恵まれる	3027	高める
399	認める	2068	もつ
373	思う	1581	使う
355	伸ばす	1547	応じる
324	もつ	1220	有する
311	感じる	1175	失う
291	溢れる	1137	優れる
238	あふれる	1074	引き出す

このリストからは、「才能」は先天的なもので、もっぱらプラス評価の観点から使われる傾向が強いことが知られる。「能力」には「失う」やこのリストの外（第20位前後）に位置する「劣る」「落ちる」などの動詞もよく続くが、「才能」にそれらの動詞が続くことは少ない。

以上の例だけからでも、内省では容易に知り得ない情報をコーパスの調査によって容易に獲得できることが分かるが、これで満足というわけにはいかない。

まず、例えば「熱意」に「伝わる」「感じる」「負ける」「あふれる」などの動詞が続くと言っても、実際には「熱意が伝わる」「熱意を感じる」「熱意に負ける」「熱意（に）あふれる」のように名詞と動詞の意味・表現上の関係が動詞ごとに異なる。上の分析ではそれらがすべて同列に扱われている。

この問題に対処するには、名詞に助詞を加えたものを中心語とすればよい。次に示す（表7）は、「熱意が」「熱意を」「熱意に」を中心語として、それに後続する動詞を調べた結果である。

（表7）「熱意が」「熱意を」「熱意に」に後続する動詞

■熱意が（…）［動詞］		■熱意を（…）［動詞］		■熱意に（…）［動詞］	
375	伝わる	565	持つ	120	負ける
239	ある	207	感じる	65	押す
157	感じる	101	もつ	51	打つ
43	通じる	97	伝える	48	応える
28	足りる	92	示す	35	ほだされる
17	薄れる	57	込める	29	燃える
17	あふれる	46	見せる	27	あふれる
16	違う	42	失う	25	満ちる
13	こもる	35	注ぐ	22	動かす
11	冷める	31	燃やす	19	よる

3通りに分けて――また、必要に応じてほかの助詞についても――調べるという調査上の手間が必要になるが、上述の問題はこれでかなりの程度に解決が付く。

しかし、（表7）の結果を見てさらに考える必要があるのは、例えば「熱意に」に「負ける」「押す」「打つ」「答える」などの動詞が続くと言っても、そのうち「押す」と「打つ」はもっぱら「熱意に押される」「熱意に打たれる」のように受身の言い回しで用いられるという事実である。

かりに問題が受身だけなのであれば、名詞に助詞を加えたものを中心語とするという処置にならって、動詞と受身の助動詞の連続を特別に処理するようにすればよい。しかし、実際のところ、こうした名詞などと動詞の共起のあり方を左右する述語形上の条件は受身要素の有無にとどまらない。

例えば、「ついに」という副詞によく後続する動詞を調べてみると、「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た」などのように不可能の表現の形において現れるのであり、可能・否定・変化という3つの要素の存在に依存している。

また、「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。

こうした問題は共起語分析の方法では対処が困難であると言うほかなく、3.2で後述する方法において解決を図る。

3. 1. 4.　分析事例２――所与の動詞に先行する名詞

共起語分析の方法で動詞に先行する名詞を調べる場合には、上の「熱意」の例において行ったように、助詞との関係を考慮に入れることが不可欠である。例えば「任せる」という動詞に先行する名詞を単純に調査すれば「身」「仕事」「流れ」「他人」などの名詞の出現頻度が高いことが分かるが、そうした分析では前二者は「身を」「仕事を」、後二者は「流れに」「他人に」の形で「任せる」と関わるという重要な違いが不問に付されてしまう。

それは他動詞について調べる場合に限った問題ではない。例えば「眠る」という自動詞は「～が眠る」という単純な形の表現で使われるとは限らず、「朝まで」「ベッドで」のような表現の共起が可能だからである。「余る」「残る」でも、「～が余る」「～が残る」のような表現だけでなく、「～が目に余る」「～が手に余る」「～が会社に残る」「～が心に残る」のような言い回しがあるから、調査にあたっては助詞との組み合わせを考慮に入れることがやはり必要である。

（表８）は、「余る」と「残る」に先行するガ格名詞の上位語のリストである。副詞としての「余り」の用例を除外する目的のために、「余る」「残る」とも連用形以外の活用形の用例だけを対象として調べている。

（表 8）「余る」「残る」に先行するガ格名詞

■［名詞］が（…）余る		■［名詞］が（…）残る	
1270	時間	2085	疑問
307	お金	1878	跡
182	金	1700	雪
161	元気	1521	記録
136	チケット	1093	記憶
114	エネルギー	1060	問題
107	体力	1011	疲れ
107	人	961	人
96	力	931	印象
86	予算	906	言葉
75	土地	857	不満
62	米	850	何
50	生地	803	余韻
50	パワー	794	後遺症
48	食料	783	不安

「残る」の主語は下線を施して示しているような“あることがらの結果として生じたもの”である場合が圧倒的に多いという事実は事前の予想にないものであった。「ご飯が｛余った／残った｝」とか「部屋に人が｛×余って／残って｝いる」といった例文だけに着目し、“以前の時点に存在した事物の一部（または全部）が消えずに存在している”という見方を前提として内省で考えていたのではそのような用例には気付きにくく、気付いたとしても上のリストに見るような主語の範囲や頻度的な傾向を把握することはむずかしい。

3. 1. 5. 共起語分析の評価

いくつかの具体例を通して見てきたように、このような単純な調査法でも、コーパスは内省では知り得ないさまざまなコロケーション情報を提供してくれる。

もっとも、3.1.3で見た通り、コロケーションを単純な2語間の関係として調査・分析する方法には限界がある。名詞については助詞、動詞については助動詞類などとの共起を考慮に入れる必要があり、特に後者は現実に対処が困難であった。

しかし、共起語分析の問題点はそのことにはとどまらない。

例えば、「健康」という名詞に「気」「害」「付ける」「及ぼす」などの語がよく後続することが分かったとしても、それらは任意の組み合わせで「健康」に後続し得るわけではない。「健康に気を付ける」や「健康に害を及ぼす」という表現は一般的でも、「健康に気を及ぼす」「健康に害を付ける」といった表現はそうではないからである。これは、「気を付

ける」や「害を及ぼす」のような言い回しを一まとまりのものとして扱うための方策が必要であることを意味する。しかし、それは、「気を付ける」のような一般に慣用句とされる表現を特別扱いすればすむという問題ではない。例えば、「荷物」に「バッグ」「台車」「詰める」「載せる」がよく続くとしても、一般的な組み合わせは「荷物をバッグに詰める」「荷物を台車に載せる」であって、「荷物をバッグに載せる」「荷物を台車に詰める」などではない。

また、そもそも名詞に後続する動詞、動詞に先行する名詞、副詞に後続する動詞、といったさまざまな関係を分析するのにそれぞれに異なるプログラムを用意する必要があることも共起語分析の実施上の難点である。このことは、特定の種類の関係についてのみ分析する場合にはさほど問題にならないが、多くの種類の語句のコロケーション情報を分析するという場面においては大きな負担となる。名詞を中心語とする場合に限っても、動詞との関係だけでなく、形容詞、連体修飾などさまざまな要素との関係も調査する必要がある。そうした多様な２語間の関係のために一々専用のプログラムを作成して調査するのは効率が悪く、すべての関係の可能性を網羅することは現実に困難でもある。

これらの問題を解決するために、共起語分析に代わる分析方法の可能性を検討する必要がある。

3. 2. 共起語連鎖分析――所与の語句とよく共起する語連鎖への着目

3. 2. 1. 分析の基本方針

第２の分析方法である共起語連鎖分析では、コーパスにおいて所与の語句の近傍にどのような語連鎖――語レベルのN-gram――がよく現れるかを調べる。

3. 2. 2. 分析手順の検討

中心語の一定範囲の先行文脈と後続文脈のそれぞれに高頻度で現れる語連鎖を調べる。これとは異なる考え方として、前後の文脈に分けることをせず、中心語を含みその前後にわたる一定範囲の文脈を対象として語連鎖を調査するという方法の可能性も考えられる。しかし、実際に試してみたところ、得られる語連鎖のリストで前後の文脈に関わる情報が入り混じると、リストが非常に読みづらいものになることが分かった。したがって、中心語の前後の文脈に分けて調査を行うことにする。

その決定に加えて、語連鎖の認定を、それを構成する各語の代表形――用言の終止形など――に基づいて行うか、それとも、各語の出現形（異形態）に基づいて行うかということについても選択を行う必要がある。この２つの方法はそれぞれに長短がある。

語連鎖を各語の代表形に基づいて認定する場合、例えば「めっきり」という副詞の後続文脈によく現れる語連鎖の１つとして「が 減る」が得られる。一方、語の出現形に基づく認定によれば、それが「が 減り」と「が 減っ」という２通りの語連鎖として別々に認定されることになる。この例の限りでは前者の代表形に基づく認定のほうが合理的で好ましいようにも思えるが、代表形に基づく認定では「まさか」という副詞に後続する語連鎖の上位に「と は 思う」が位置付けられることになる。しかし、この語連鎖は実際にはほとんど「とは思わなかった」「とは思いませんでした」などの否定の言い回しの一部として出現するのであり、代表形に基づく認定ではその重要な情報が欠落してしまう。

また、語の代表形に基づく語連鎖の認定には、得られる語連鎖のリストにおいて、活用語を含む語連鎖がしばしば読みづらいという問題もある。例えば、「やっと」に「て くる ます た」、「どうにも」に「なる ない なる て」、「後悔」に「は する て いる ます ん」という語連鎖がよく後続するという情報が得られても、それらが「（やっと）～てきました」、「（どうにも）ならなくなって」、「（後悔）はしていません」のことであることを読み取るには少々労力を要する。平易な謎解きではあるが、多数の語連鎖に目を通すうえではかなりの負担になる。

以上のような理由から、ここでは所与の語句の近傍を先行文脈と後続文脈とに分け、各語の出現形に基づいて語連鎖を認定するという方法によってコロケーション情報を調査する。また、以下の論述において、語連鎖のリストを読みやすいものとするために、語連鎖は「て き まし た」のように語ごとに区切ることはしないで「てきました」のように表記することにする。

3. 2. 3. 分析事例

次に示す２つのリスト（表９）と（表10）は、それぞれ「めっきり」「めきめき（と）」という２つの副詞について、それぞれの前後の文脈によく現れる語連鎖を調べた結果の一部である。

これ以後の例における表中の数値は用例数を基礎としつつ、それに若干の調整を施したものである。最も重要な調整は、語連鎖が中心語に隣接して現れる場合はそうでない場合の２倍の重みを持つと見なすというものである。また、例えば「なりました」と「なりまし」のように一方の語連鎖が他方を含む関係にあるとき、両者の用例数に大きな差がなければ短いほうを省くとか、単独の助詞や助動詞は最初から調査の対象としないといった、リストに載る情報の重複や無駄を減らすための措置を講じている。

そのような調整や措置は恣意的な性格を有するが、そのことはさしたる問題ではない。

（表 9）「めっきり」の後続文脈・先行文脈に現れる語連鎖

■めっきり（…）［語連鎖］

1340	減っ
1316	ました
1018	寒く
1015	なって
819	少なく
780	減って
768	なった
768	涼しく
667	てき
611	寒くなっ
549	なりました
547	少なくなっ
503	減った
482	なってき
469	寒くなって
392	涼しくなっ
381	減り
380	寒くなり
361	なくなっ
346	寒くなってき
316	涼しくなり
312	たね
304	寒くなりました
288	てしまっ
285	てきました
282	涼しくなって
281	少なくなった
279	ましたね
255	少なくなって
248	減りました
237	秋
236	なってきました
227	てきた
224	少なくなり
222	てしまった
212	減ってしまっ
207	見
206	てしまい
202	寒くなってきました
197	涼しくなりました
180	ましたが
176	減ってしまった
170	弱く
168	になっ
163	涼しくなってき
161	なくなって
156	なってきた
156	てしまいまし
147	減ってしまい
142	少なくなりました
140	なくなり
139	見なく
136	冷え込んで
135	が減っ
129	寒くなった
127	春

■［語連鎖］（…）めっきり

3138	最近
929	最近は
307	朝晩
283	では
258	てから
216	機会が
191	朝夕
180	機会
158	ことが
154	近頃
132	最近では
129	朝晩は
116	朝夕は

（表 10）「めきめき」の後続文脈・先行文脈に現れる語連鎖

■めきめき（と）（…）［語連鎖］

193	上達
169	頭角を
122	上達し
88	力を
84	腕を
82	して
82	頭角を現し
80	てき
64	ている
63	上達して
60	腕を上げ
59	成長
57	力をつけ
54	上達する
50	上げ
47	ました
46	実力を

■［語連鎖］（…）めきめき（と）

119	最近
31	てから
25	して
20	腕
19	からは
18	ように
17	力
16	すれば
16	になって
14	腕は
13	の腕
13	指導
13	を受け

筆者の理解によれば、1.4 で述べたように、コロケーション情報を統計的に厳密に扱う必要はなく、また、情報は応用のために適宜利用すればよい。コロケーション情報の最終的な評価は、その統計自体においてではなく、それを応用して得られた成果の次元でなされるべきものであるというのが筆者の見解である。

共起語分析の場合と異なり、共起する表現として単一の語だけでなく句の形をした表現が抽出されており、中心語がどのような語句とよく用いられるかをリストから容易に読み取ることができる。調査の方法上、類似した語連鎖がリストに載り、そのために共起語分析の場合と違ってリストが冗長性を含む大きめのものになることは避けられないが、例えば「めっきり」には「減った」や「{寒く／涼しく／少なく}なっ（てき）た」など、「めきめき（と）」には「上達する」「頭角を現す」「腕を上げる」などがよく後続することが分かる。

次の（表11）は、名詞「熱意」に助詞「が」「を」「に」を加えたものを中心語としてその後続文脈によく現れる語連鎖を調べた結果である。

動詞を含む語連鎖については、助詞・助動詞・補助動詞などとの組み合わせが多数抽出されており、当該の動詞が「熱意{が／を／に}」との組み合わせにおいてどのような形でよく使われるかをリストから知ることができる。

（表 11）「熱意が」「熱意を」「熱意に」に後続する語連鎖

■熱意が（…）［語連鎖］		■熱意を（…）［語連鎖］		■熱意に（…）［語連鎖］	
417	伝わって	862	持って	267	負け
318	感じ	339	感じ	118	押され
227	感じられ	159	もって	117	打たれ
213	伝わってき	145	伝え	112	負けて
164	あれば	141	ている	94	ました
136	伝わる	102	示し	82	れて
134	てき	102	感じました	79	負けた
123	伝わってくる	94	込めて	72	応え
86	通じ	86	持ち	63	動かされ
83	伝わった	86	感じる	59	れた
79	伝わり	84	持っている	58	感動
78	てくる	82	持つ	56	心
72	感じられる	72	伝える	52	満ち
71	感じられない	70	持った	50	押されて

3. 2. 4. 共起語連鎖分析の評価

共起語連鎖分析は２つの点において共起語分析にまさる。

第１に、言語表現の慣用をより詳しく観察することできる。すなわち、例えばよく使われる動詞のリストが得られるだけではなく、それぞれの動詞がどのような形の言い回しで用いられるのかという情報が同時に得られる。

第２に、調査方法の一般性、汎用性が高い。調査に当たっては中心語に任意の表現を指

定するだけでよく、共起語分析の場合のように名詞に後続する動詞、動詞に先行する名詞、副詞に後続する動詞といった場合ごとに異なるプログラムを用意する必要がない。ある名詞とよく共起する動詞単体のリストだけを得たいといった目的の場合には一点集中式に分析を行う共起語分析のほうが有利であるが、コロケーション情報の一般的な抽出方法としては共起語連鎖分析のほうがはるかに効果的である。

共起語分析と共起語連鎖分析に加えて、もう１つの分析方法を試してみた。ただし、さまざまな事例について分析を行ってみたところ、得られるコロケーション情報は共起語連鎖分析のそれに近く、しかも無用の情報をより多く含みがちであることが分かった。そこで、この第３の方法による分析については簡単に述べるにとどめる。

3. 3.　共起文字連鎖分析――所与の語句とよく共起する文字連鎖への着目

3. 3. 1.　分析の基本方針と手順

第３の分析方法である共起文字連鎖分析では、コーパスにおいて所与の語句の近傍にどのような文字連鎖――文字レベルのN-gram――がよく現れるかを調べる。

共起語連鎖分析の場合と同様に中心語の一定範囲の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文字連鎖を調べる。

3. 3. 2.　分析例と評価

（表12）は、「めっきり」の後続文脈によく現れる文字連鎖のリストの一部である。

（表 12）「めっきり」に後続する文字連鎖

■めっきり（…）［文字連鎖］					
1746	くなっ	792	少なくな	505	くなりました
1345	減っ	764	なくなっ	504	減った
1179	ました	724	涼しくな	477	寒くなって
1138	なくな	722	りました	444	くなってき
1019	寒くな	717	くなった	428	なくなった
948	くなって	622	寒くなっ	394	涼しくなっ
804	くなり	554	ってき	387	寒くなり
797	減って	543	少なくなっ	386	減り
		533	しくな	371	減ってしま

このリストは共起語連鎖分析によって得られた（表９）のリストに似ているが、そこにはなかった「くなっ」「くなって」「くなり」などの文字連鎖が含まれている。言うまでもなく、それらの第１文字の「く」は「寒く」「涼しく」「少なく」などの形容詞の一部である。

同様に、「なくな」「寒くな」「少なくな」などの文字連鎖の末尾の「な」は「なった」「なって」「なり」などの一部である。共起文字連鎖分析においては語の境界を考慮しないことから、そのようなものまで数え上げられることになる。

分析する事例によっては、そうした無用の情報をあまり含まない、共起語連鎖分析とほぼ同じ結果が得られる場合もあるが、この共起文字連鎖分析には共起語連鎖分析を明確にしのぐ特徴は見出せなかった。理屈のうえでは、共起文字連鎖分析は形態素解析の過程を必要としないことから、処理に要する時間を節約できるというメリットがある。しかし、実際のところ多くの場合処理時間のかなりの部分はコーパスからの用例の検索が占めるので、形態素解析が不要であることのもたらす処理効率上の恩恵は大きくない。また、形態素解析の誤りや不統一の問題から解放されることによる分析の質の向上の可能性については、特に評価できるほどの効果は認められなかった。

したがって、この小論における総合的な評価としては、今回試した分析方法の範囲では、共起語連鎖分析がコロケーション情報抽出のためのもっとも優れた方法であるという結論になる。

４．日本語コロケーション辞典――項目の試作

コーパスから共起語連鎖分析によって抽出したコロケーション情報に基づいて、日本語コロケーション辞典の数項目を試作してみた。試作の項目には「制限」と「制約」、「希望」と「願望」という２組の類義語を選んだ。その結果を（表13）、（表14）に示す。

実際の辞典においては全項目が五十音順に並べられるとすれば、類義語の対がこれらの表に示すような形で辞典の使用者の目に触れるわけではないが、対をこのように対照できる形に並べて見ることで２項目のあいだの違いが明瞭になり、各項目の記述内容を比較・検討しやすくなる。

これらの試作項目から、コーパスの利用によって辞典の記述を現実の日本語をよく反映した自然なものにすることができることを見て取ることができよう。当該の語の用いられる表現パターン、その文例とも、日常生活の中でしばしば見聞きする種類のものを取り上げることが可能になっている。例えば、「希望」で始まり、よく使われる複合語に「希望者」「希望小売価格」「希望退職」「希望的観測」などがあるということは言われれば確かにそうだと分かるが、そのことを内省や少数の用例調査で明らかにすることは容易ではない。また、「制約」を要素とする複合語は種類も用例も少ないといった否定に傾く事実も、内省に基づいて指摘し、あるいは、自信を持って判断することはむずかしい。

（表 13）コロケーション辞典の試作項目（1）「制限」「制約」

制限　せいげん

1　名詞
◇「制限」＋動詞
―を設ける、―を加える、―を緩和する、―を解除する
……*多くの国は外国人の長期滞在に関していろいろと制限を設けている*
―が設けられる、―が緩和される、―が解除される
……*この区域では建物の用途や規模に制限が設けられている*
―を受ける、―を超える
……*手荷物は一人**20kg**までという制限を超えると追加料金がかかる*

◇「制限」＋形容詞　※種類も用例数も少ない
―が多い、―が大きい、―が少ない

◇連体修飾句＋「制限」
～の―、～に関する―、～に対する―
……*面会の権利の制限／受験回数に関する制限／配偶者の収入に対する制限*
厳しい―
……*運動や外出についてはさほど厳しい制限は課していない*

2　複合サ変動詞語幹
自由を―する、権利を―する
……*この法案は個人の自由・権利を不当に制限するものだ*
利用を―する、行動を―する、活動を―する
……*混雑時はご入場を制限することがあります／輸入を制限することによって自国の産業を保護している国は多い*

3　複合語要素
◇「～制限」
食事―、年齢―、人数―、速度―、高さ―、字数―
……*散歩と食事制限で減量に励んでいます／年齢制限に引っかかってもう受験できない*
無―、時間無―
……*どちらかがＫＯされるまで時間無制限で闘うというルール*

◇「制限～」
―時間、―速度、―事項
……*制限時間内に問題を全部解くのはむずかしい／その男は制限速度**50**キロのところを時速**130**キロで運転して捕まった*

制約　せいやく

1　名詞
◇「制約」＋動詞
―を課す、―を設ける、―を加える
……*この法改正は報道の自由に制約を課すものだ*
―が課せられる
……*国際社会では軍事力の行使にさまざまな制約が課せられている*
―を受ける、―を満たす
……*電子媒体は紙媒体と異なり紙面やコストの制約を受けない*

◇「制約」＋形容詞
―が大きい、―が厳しい、―が多い、―が少ない
……*野党では政策の実現に制約が大きい*

◇連体修飾句＋「制約」
時間的(な)―、物理的(な)―、技術的(な)―
……*時間的な制約のためすべての要請に応じることは不可能である*
時間の―、基本的人権の―、システムの―
憲法上の―、技術上の―
～に対する―
……*表現の自由に対する制約*
厳しい―、大きな―、一定の―
……*正社員の解雇に関しては厳しい制約がある*

2　複合サ変動詞語幹
自由を―する
……*その決定は国民の自由を不当に制約するものと言わざるを得ない*
行動を―する

3　複合語要素　※種類も用例数も少ない
◇「制約～」
―条件

（表 14）コロケーション辞典の試作項目（2）「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―

コロケーション辞典の試作項目についてはこれ以上の説明や議論を要しないであろう。その出来のほどについてはご覧になった方の評価に委ねたい。

５．おわりに

日本語の大規模なコーパスを用いてコロケーションの現象を調査・分析し、その結果をコロケーション辞典の作成に応用する可能性を模索・検討した。分析の手順にはさらに精密化や効率化を図る余地があり、あるいはまったく異なる着想に基づく手法の可能性にも関心を覚えるが、ともあれコロケーションの研究におけるコーパスの有用性を十分に確認することができたものと思う。

大規模なコーパスは言語の使用に関わる統計的事実を圧倒的な説得力をもって我々の前に提示してくれる。過去には考えられなかった特性と規模を有する言語資料により、過去には考えられなかった種類と質の日本語研究へのさまざまな道が今拓かれようとしている。

注

１）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、という方法で行った。実際の処理に際しては対処を要するさまざまな問題が生じるが、そうしたことについてはここでは省略に従う。

２）MeCabはhttp://mecab.sourceforge.net/で公開されている。

３）解析用の辞書としてIPA辞書の代わりにUniDic（http://www.tokuteicorpus.jp/dist/）を使えば、このような分析の不統一の問題はかなりの程度に回避することができる。しかし、UniDic使用時における異表記同一語の認定の精度が仮名表記された語の場合に十分高くないためにコロケーションの調査にはかえって不都合をもたらすことから、ここではIPA辞書を使用した。UniDicでは、例えば「いった」の現れはあるときは「行った」、あるときは「言った」として解釈されるのであるが、それが正しい認定であることも多いが、そうでないことも多い。IPA辞書では異表記同一語の認定はなされず、「いった」は「行った」「言った」からは独立した「いった」として扱われる。

付記

本稿は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」の日本語学班研究会（大阪大学待兼山会館、2008年８月４日）および辞書編集班拡大班会議（ホテルスワ（つくば市）、2008年11月23～24日）における発表の内容を合わせまとめたものである。

（文学研究科教授）