コーパスデータに基づく名詞コロケーションの辞書記述

1M－3

橋本三奈子*，桑畑和佳子*，青山文啓**，村田賢一*
*情報処理振興事業協会　**東海大学

0. はじめに

昨今、機械翻訳をはじめとする自然言語処理において、例文やコーパスを用いた分析が注目されている。情報処理振興事業協会（ＩＰＡ）技術センターでも、新聞・教科書・論説文・小説や市販辞書の用例文のコーパスを独自に作成し、辞書の記述に利用してきた。現在作成中の『計算機用日本語基本名詞辞書ＩＰＡＬ(Basic Nouns)』でも、見出し語が形成するコロケーションを豊富に収録する予定である。そこに収録されるコロケーションは、コーパスに現れた文章を言語学の研究者の判断を通して分析し抽出したものである。

本稿では、初めに、『名詞辞書』に収録するコロケーションの範囲と種類について述べる。次に、語義や用法の違いに基づいて行われる見出し語の下位区分において、コロケーションと語義の関係を初めとする問題点について論じる。

1. 慣用句とコロケーション

複数の語が密接に結びついて文中に現れるものとして「慣用句(idiom)」があるが、これは(1)個々の語の意味の総和が当該句全体の意味にならない、(2)述語の活用形や使役化・受身化・否定化・肯定化・丁寧化や、さらに修飾語の付加や省略の可能性などに制限が見られる、という特徴を持つ。ここでは、複数の語が密接に結びついて現れる関係のうち、「慣用句」とは異なり、句全体の意味を個々の語の意味へ還元することができ、文法的なふるまいに制限がないものを「コロケーション(collocation)」と呼ぶ。

名詞のコロケーションという場合には、その名詞がどのような述語の格要素となるか、その名詞がどのような語に修飾されるか、その名詞がどのような語を修飾するかという三つの視点が欠かせない。『名詞辞書』では、それぞれ、(1)項としての用法、(2)被連体修飾語としての用法、(3)連体修飾語としての用法という情報欄を設け、用法ごとに該当するコロケーションを記述する。

本稿では、このうち(1)項になる用法、すなわち次の形式のように見出し語が格助詞を介して述語と結びつくコロケーションの問題に焦点をあてる。

（［名詞］ノ）
（［形容詞］）　【見出し語】　ガ／ヲ／ニ　［動詞］／［形容詞］／［名詞］ダ

2. 『名詞辞書』におけるコロケーション

見出し語と結びつくことが可能な述語をすべて記述することは不可能に近い。そこで『名詞辞書』では、次の三つの観点から、当該見出し語の意味やふるまいを特徴づけるようなコロケーションを抽出して収録する。

(A) 構文的な問題と関連するもの

「スル」「ナル」「アル」「ナイ」という述語は、かなり広範囲の名詞と結びつくことが可能であるが、サ変動詞用法などの構文的な問題とも関わる重要なものである。

ヲスル	【発表】をする	【指輪】をする
ニスル	【逆様】にする	【医者】にする
ニナル	【孤独】になる	【美人】になる
ガアル	【選挙】がある	【連絡】がある
ガナイ	【伝統】がない	【楽しみ】がない

(B) 見出し語固有のもの

慣用句ほどの意味的・統語的制限はないが、見出し語と固定的な結びつきを示す述語がある。これらは、＝の右辺に示したように、他の表現と言い換えられるものが多い。

【判定】を下す＝判定する／判定をする
【反対】にあう＝反対される／反対を受ける

【成功】を収める＝成功する

【衝動】に駆られる＝衝動を感じる

(C) 知識ベースとして欠かせないもの

(B) に比べれば、生産性の高い結びつきではあるが、コーパスに現われる頻度の高い表現のうち、常識の範囲内で使用されるものについては、辞書記述の立場からも見逃すことができない。この方針に従って、例えば「（頭から）ズボンをかぶる」については収録を見送るが、「ズボンをはく」のような例については収録に努めた。

【ズボン】をはく／を脱ぐ／が長い

【電気】を点ける／を消す／が暗い

【失敗】を無くす／を隠す／が続出する

【注意】を集める／をそらす／が届かない

3. コロケーションと語義との関係

『名詞辞書』では、サ変動詞用法などの有無、コロケーションの違い、類義語の組み合わせの違い等に基づいて、見出し語を下位区分する。下位区分されたものを見出し語の「語義」と呼ぶことにすれば、コロケーションと語義との関係は大きく以下の二つに分類できる。

(X) 語義が異なり、コロケーションも違うもの

「点数」という見出し語には、(a)試験などの得点、(b)作品や品物の数、という二つの語義がある。

(a)（数学の）【点数】が高い／低い

(b)（絵画の）【点数】が多い／少ない

上の例では、「高い」と結びついた「点数」は(a)の語義に特定でき、「多い」と結びついた「点数」は(b)の語義に特定できる。つまり、対象テキストに当該コロケーションが存在すれば語義の違いは特定可能であり、当該コロケーションに対し、意味や翻訳を与えることに効果が期待できるはずである。

(Y) 語義は異なるがコロケーションが同じもの

一方、【育ち】という見出し語には、(a)植物などの発育の状態と、(b)人間が成長するときの環境や育てられ方など、概略二つの語義がある。ところが、両者ともに「育ちが良い／悪い」というコロケーションを持つ。

(a)（朝顔の）【育ち】が良い／悪い

(b)（太郎・）【育ち】が良い／悪い

このような場合には、当該コロケーションの存在だけでは、語義を特定することができず、下に示すように、別の語との結合を含めて語義を判断する必要がある。

4. 語義の認定における問題点

コーパスに存在するコロケーションを分析してそこから語義を認定する場合、辞書の中で記述すべきものとそうでないものとの区別が問題になる。

例えば「見物」では、「見物が出る／多い／大勢いる」という表現がコーパスに見つかる。この例だけを見ると「見物」は「見物客」の省略であり、「見物」に〈人間〉という意味を認めてよいように思われる。ところが、「見物客がけがをした」とは言えても、「見物がけがをした」とは言えない。つまり、先の例では「見物」は特定できない人間の集まりを表わしたものであり、「見物客」と同義ではないことがわかる。また、「訪問」の場合も、「見物」と同様にサ変動詞用法を持ち「訪問客」という複合名詞を作ることができるが、「訪問が大勢いる」とは言えない。したがって、「見物」という見出し語に対しては「出来事や場所を見て楽しむ不特定の人の集まり」という語義を認めるが、「訪問」にはこの語義は認めない。このように規則化できない事柄を個別に辞書に記述するという方針をたてた。

5. おわりに

『名詞辞書』は、コーパスに現れた実例から、研究者の判断で抽出したコロケーションを豊富に記述した知識ベースである。今年度末の完成をめざし，現在校閲作業を進めているところである。

参考文献

(1)M. Benson, et al., THE BBI COMBINATORY DICTIONARY OF ENGLISH, John Benjamins(1986).