解析と生成のための共起情報の記述方法

柏野和佳子
国立国語研究所

1．はじめに

計算機による日本語処理は，解析と，生成の二種類に大別してとらえることができる。解析における主な課題の一つは，多義語の曖昧性解消である。一方，生成における主な課題の一つは，表層表現化する際の語句の選択や組み合わせである。よって，解析のためには，その曖昧性解消に役立つ情報が，生成のためには，語選択と共起の可否判定に役立つ情報が必要になる。すなわち，解析や生成といった目的によって，必要な情報やその詳細さは異なっている。

しかしながら，それぞれに必要な情報を別々に辞書に構築していくことは，効率的ではない。なぜなら，辞書に記述すべき情報には，解析にも生成にも共通に必要なものも多いからである。ところが，従来の辞書記述の研究では，解析と生成という視点から，共通に必要になる情報と，それぞれの目的に特化した情報とを切り分けて議論することは，ほとんど行われていなかった。

そこで筆者は，解析と生成に共通に必要な情報と，個別に必要な情報とを切り分け，それらを盛り込んだ辞書記述の方法について研究を進めている。辞書情報のうち，名詞句と格助詞と述語との結びつき，また，修飾語と被修飾語との結びつきなどに関する「共起情報」は，多義語の曖昧性解消，語選択，共起の可否判定に役立つ情報を含む重要な情報である。そこで本稿では，研究の一段階として，共起情報の記述に絞って議論し，例示の網羅性を明示するための「限定，列挙，代表型」を区別する記述方法，多義語の曖昧性解消を支援するための「特有語」を区別する記述方法，語選択を支援するために，述語表現，修飾表現，名詞の下位表現などへ意味ラベルを付与する記述方法を提案する。

2．これまでの共起情報の辞書記述の実際と問題点

最初に，これまでの計算機用辞書における記述の実際と問題点を述べる。

これまでに作成され，一般にも入手可能な計算機用辞書には，『ＥＤＲ辞書』[EDR93]，『ＮＴＴ日本語語彙体系（以下，ＮＴＴ辞書と呼ぶ)』[NTT97]，および『ＩＰＡＬ』[IPA97a]がある。このうち，ＮＴＴ辞書は，日英翻訳のために開発された解析用の辞書である。あとの二つは，目的を明示していないが，解析よりの記述といえる[岩波 98]。いずれの辞書にも，共起情報として，格パターン，あるいは，格フレーム[Fillmore75]などと呼ばれるものが記載されている。

格パターンとは，どのような名詞句（どのような意味マーカ[1]をもつかで表されることが多い）が，どのような格助詞を介して述語（動詞／形容詞／名詞＋ダ）と結びつくかを示すものである。

たとえば，ＩＰＡＬ動詞辞書には「産む・生む」という動詞について，図１に示すように，２つの意味それぞれに格パターンや文例が記載されている。

```
産（う）む・生（う）む：
①動物が子や卵を作り出す。
【格パターン】
 N1[HUM/ANI]ガ N2[HUM/ANI]ヲ (N3[QUA]φ)
   N1  HUM:彼女 / ANI:馬，小鳥，鮭，魚
   N2  HUM:女の子 / ANI:子馬，卵
   N3  QUA:一人，一頭
【文例】猫が 子を 5匹 産んだ。
②それまで無かった事を生じさせる。
【格パターン】
 N1[ABS]ガ N2[ABS/HUM]ヲ
   N1  ABS:誤解，勤勉，努力，時代
   N2  ABS:誤解，成功，勝利，傑作，感動
       / HUM:天才，英雄
【文例】努力が 天才を 生む。
 *英文字3字は，IPALの意味素性（結びつく述語によって焦点が当てられる意味的な側面）である。
  HUM:人間，ANI:動物，QUA:数量，ABS:抽象物
```

図１：ＩＰＡＬ「産む・生む」の記述

図１にもみられるとおり，多義語の多くの場合，意味ごとに格パターンに違いがあることが着目され，これまで，多義語の曖昧性解消に役立つ情報として，計算機用辞書に積極的に記述されてきた。しかしながら，以下のような６点の問題点がある。

(1) 格パターンの記述を，名詞と述語，各一語の組み合わせに限ると，解析時に，同一の名詞と述語にあるコロケーションの曖昧性を解消できない「例：(足の／傘の／仕事に) 骨が折れる」[桑畑他98]。また，生成時の統語情報にも欠ける。

(2) 解析時にも，生成時にも，共起する語句の例示が共起関係の実際をどれほど示しきれているのかといった網羅性が問題になるが，網羅性に関する情

[1] 本稿では，意味マーカと呼ぶが，EDR辞書では意味概念，NTT辞書では意味属性，ＩＰＡＬでは意味素性と呼ばれるものが，それぞれ独自に構築されている。

報に欠けている。

(3) 解析時に，どの共起語が多義の弁別をし得るかの情報に欠ける。

(4) 生成時に，組み合わせ情報のない名詞句例をそのまま利用すると，「馬が卵を産む」のような文も生成されてしまう[橋本他 97]。

(5) 生成時に，語選択を意味的に行うための情報記述に欠ける。

(6) 生成時の共起の可否の判定に関して，否定の情報がない。

以上のうち，(1)と(2)が解析，生成両方の問題であり，(2)，(3)が解析の問題，残りが生成の問題である。以下，３章で解析と生成の問題を先に議論し，４章で解析の問題，５章で生成の問題を議論する。

3．解析と生成のための記述

詳細な格パターン情報の記述は，解析時にも，生成時にも必要になるものである。よって，問題点(1)について，格パターン情報として盛り込むべき事柄の検討は重要な課題であるが，先に，[柏野 98]でその検討を示してある[2]ので，ここでは，取り上げない。本章では，問題点(2)の例示の網羅性を取り上げて論じる。

3.1 例示の網羅性について

解析時と生成時では，共起例の記述に求めるものに多少の違いがある。たとえば，解析時には，多義を区別し得る情報が求められ，生成時には，多義性には関係なく，共起の可否が求められる。しかし，いずれの場合も，共起例に網羅性を求めていることに違いはない。たとえば，網羅性に欠けてしまうと，解析時には，実際に出現する共起の出現形に対応できなくなる。また，あり得る共起例が書かれていてない，ということは，すなわち，その分の生成の可能性を奪っているということになる。

しかしながら，現実問題として，共起可能なすべての表現を，網羅的に収集して記述することは不可能である。たとえば，先の図１「産む・生む」の記述例では，「彼女」が「女の子」を産む，という結びつきが例示されていた。人間の場合，生まれるものとして，他に，「男の子，子供，赤ちゃん，赤ん坊」といったものまで書けば，網羅的な記述が実現したかのように考え，「おじいさん」が生まれた，などは想定しないと思われる。ところが，「おじいさんが生まれた日は雪の日だった」といったような文では，「おじいさん」もまた「生まれる」と共起可能な語であったことになる。このようなことまで想定し，さらに，否定表現での結びつきまでも許すなら，共起可能な語は爆発的に増えてしまい，網羅的な記述が不可能であることは自明であろう。

そのため，これまでの計算機用辞書では，限りある例示と，実際の出現形との対応を少しでも広げられるようにと，意味マーカを設定し，抽象化させて共起情報を記述する，という方法がとられてきた。次節でその方法について検討する。

3.2 意味マーカ，シソーラス利用の利点と問題点

先に図１に示した「産む・生む」の格パターン情報も，一種の意味マーカを用いた記述方法をとっている。意味マーカとシソーラスを利用することにより，少ない記述例の網羅性を高めることができる。すなわち，共起する語の意味マーカが表示されてあり，シソーラス（たとえば，ＮＴＴ辞書の『単語意味属性体系』[NTT97]や『分類語彙表』[国研 96]など）を利用して，その意味マーカをもつ他の名詞群に展開することができれば，解析時には，実際の出現形のバリエーションに対応させたり，生成時には，生成可能な表現としての候補を補充したりできる。

しかしながら，これまでにも様々に議論されているように，意味マーカやシソーラスの構築は非常に難しい。

問題となる一例を示す。たとえば，『新明解国語辞典第５版』（三省堂）によると，「すする」について，「〔ところてん・そば・かゆや，熱い茶などを〕強く吸うようにして，飲み込む。」と説明されている。この時，ヲ格にくる名詞群は，食べもののうちのほんの一部でしかないので，それらを「食料」や「食品」という大まかな意味マーカでとらえるには無理がある。シソーラス上で食品や，食料の下位にある，細分類項目名を利用して意味マーカを設定しようと，『単語意味属性体系』や『分類語彙表』などのシソーラスを参照してみても，「ところてん・そば・かゆや，熱い茶」は，「豆腐分・寒天等」「麺類」「飯」「飲物」といった，異なる細分類にまたがってまばらに分類されているため，それらの細分類項名を意味マーカとして用いることもできない。「すすられるもの」という細分類項目をシソーラス上に設け，該当する名詞群を挙げておけば解決可能であるが，これは，動詞の数だけ名詞の分類を増やすようなことにつながるものであり，現実的な解決策ではない。

このような問題に対処するために，次節に，解析にも，生成にも必要になる，共起情報の網羅性を明確に記述する方法についての提案を述べる。

3.3 限定，列挙，代表型を区別した記述方法

結びつく可能性のある語がどれほどあるかは，語によって大きく異なる。慣用的な言いまわしを含め，唯一の結びつきしかないものから，たとえば，「私」や「欲しい」のように，かなりのものと結びつきが可能なものまである。よって，結びつく語が限られるものを「限定型」，典型例の列挙によってある程度網羅的な記述にな

---

[2] 単純な格パターン情報だけではなく，連体修飾句，連用修飾句，格表示といった，詳細なあらゆる統語情報が有効であり，さらに，動詞の場合，テイル形のよう形態情報も有効である[柏野98]。また，ＮＴＴ辞書には，既に副詞や助動詞の形態情報等は盛り込まれている[NTT97]。

り得るものを「列挙型」，結びつく語が多いため，典型例は代表例に過ぎないものを「代表型」と，３タイプに呼び分け，どの型の記述であるかを明示し，それにそった記述をとる方法を提案する。

3.1 節で挙げた「する」の場合は，「列挙型」として記述するタイプになる。先に，この「列挙型」について説明する。「列挙型」の場合，場当たり的に意味マーカを増設するのではなく，典型的なものについて網羅的に列挙することで示す，という方法をとる[3]。「する」については，『新明解国語辞典』の記述のように，共起し得る名詞群を列挙する，ということである。ただし，「列挙型」の場合は，他の共起例の可能性を排除するものではない。よって，既存の，もしくは新規に作成する，意味マーカやシソーラスを利用し，たとえば「豆腐分・寒天等」「麺類」「飯」「飲物」などの一部の名詞とは共起の可能性があることが予想される，といったことをさらに示すことになる。

次に，「限定型」について説明する。すでに，これまでにも指摘されていることであるが，格要素が特定の語だけをとる場合があり[NTT97]，[亀井 97]，本稿ではこれを「限定型」と呼ぶ。図２に，ＮＴＴ辞書の「構文体系」の凡例に掲げられている例の一部を引用する。

(c)格要素が特定の単語だけをとる場合には，単語そのものを“ ”で囲ってしめした（下の例では「指示」）．
(例) 仰ぐ（あおぐ）
N1が N2に N3を 仰ぐ
[N1(3主体) N2(3主体) N3(“指示”)]
特定の単語が複数ある場合には，それらを並記し，間を“/”で区切った．
(例) 当たる（あたる）
N1が N2に 当たる
[N1(3主体) N2(“火/火鉢/ストーブ/焚き火”)]

図２：ＮＴＴ辞書「仰ぐ，当たる」の記述

図 2 に示した記述と同様に，格要素が特定されるもの，つまり，共起可能なものすべてが記述できる場合は，それで全てである，ということを明示して，それらの語を記述すべきであり，抽象化させた意味マーカをつける必要はない。これが「限定型」の記述方法である。

最後に，「代表型」について述べる。「代表型」の場合は，従来通り，意味マーカやシソーラスをうまく利用することが望まれる。それによって，記述しきれない，共起可能性のある多数のものを示し得られるからである。重要なことは，その記述が，代表型であることを明示し，他とを区別することである。

[3] 典型的な共起例とはどのようなものであるかという議論は，[IPA97b]を参照されたい。また，慣用表現として収集すべきものの分類は，[亀井97]がくわしい。

4. 解析のための記述

解析のための記述は，多義語の曖昧性解消を念頭においた情報の記述である。本章では，問題点(3)を解決するために，「特有語」の区別をした記述方法を提案する。以下に，その詳細を記す。

4.1 特有語を区別した記述方法

多義語の一つの意味の場合にしか共起し得ない語を「特有語」と呼ぶこととする。一例を示す。「脱線」には，「①電車などが線路からはずれること，②話や行いが本来従うべき正しい筋道からはずれること」の２つの意味がある。この「脱線」という名詞と共起する動詞の例を以下に挙げる。

(a1) 脱線が発生した。… ①の例
(a2) 脱線が始まった。… ②の例
(a3) 脱線が多い。 … ①，②の例

上記で，「発生する」「始まる」が，「特有語」にあたる。これらは，多義の弁別をし得る語である。共起例どうしをつきあわせることによって，どの語が特有語であるかを割り出すことは可能かと考えられるが，他の意味でもあり得る共起語であるのに，たまたま一つにだけ書かれてある場合も想定されるため，「特有語」であることは明示して記述すべきであろう。

5. 生成のための記述

問題点(4)については，[橋本 97]で検討されてあるので，ここでは触れない[4]。本章では，問題点(5)の解決策として，現在思案中のアイデアを 5.1 節で述べ，問題点(6)については，5.2 節で，今後の課題として述べる[5]。

5.1 語選択を支援する意味情報

これまで，結びつく名詞については，意味マーカを付与することはよく行われてきているが，共起情報の記述時に，結びつく述語の意味タイプを示すような意味ラベルを設け，それを付与する試みはあまり議論されていないようである[城田 91]。共起情報を，名詞を中心に記述するような場合，その名詞が「列挙型」や「代表型」であるなら，結びつく述語や，修飾語に，意味ラベルをふるのが良いと考える。次に，名詞「根拠」に結びつく述語や修飾語に，意味ラベル（以下で｛｝でくくったもの）を付与した記述例を示す。

(b1) 根拠 → 述語 {弱・少}
＝ 乏しい，弱い，薄い，欠く，曖昧だ，薄弱だ

[4] すべての用例を文テーブルとして例示する[橋本他97]。
[5] 本章で述べる事柄を含む，生成のための記述の研究は，現在，情報処理振興事業協会(IPA)「独創的技術育成事業」の一環として行っている「計算機用日本語生成辞書 IPAL(SURFACE/DEEP)の研究」で進めている[村田他98]。

(b2) 根拠 → 修飾語 {弱・少}
= 乏しい，弱い，薄い，曖昧な，いくらかの
(b3) 根拠 → 述語 {強・大}
= 明確だ，はっきりしている
(b4) 根拠 → 修飾語 {強・大}
= 絶対的，確たる，大きな，強力な，最大の，有力な，しっかりした，深い，十分な
(b5) 根拠 → 述語 {出現}
= ある，得る，持つ，見つける，見出す，確立する，置く
(b6) 根拠 → 述語 {消滅}
= ない，失う，なくなる
(b7) 根拠 → 述語 {提示}
= 与える，提供する，示す，挙げる，述べる，説明する，証明する

たとえば，「根拠」について {弱・少} ということを言おうとする時，上記のような情報があれば意味ラベルを手がかりにして，「根拠が薄い」といった表現の生成が可能になる。

名詞どうしの，上位語と下位語の関係においても，意味ラベルを活用した記述を試みることで，品詞の枠を超えて，様々な表層語の選択を可能にすると考える。次に「景色」の例を挙げる。

(c1) 景色 → 述語 {良}
= 良い，素晴らしい，美しい，見とれる，見事だ，見晴らしがいい，きれいだ
(c2) 景色 → 修飾語 {良}
= 良い，素晴らしい，美しい，見とれる，見事な，見晴らしの良い，きれいな，絵になる，極楽浄土のような
(c3) 景色 → 下位名詞 {良}
= 絶景，美景，美観，名勝，景勝，佳景，好景

述語表現，修飾表現，名詞の下位表現，といった全ての表現に，同じ意味ラベルが付与されてあれば，文脈や，時には字数制限といったものにあわせて，文，句，語，のうちから適切な形式を選んで生成したり，置換したりすることが可能になるだろう。

5.2 その他生成に必要な情報

問題点(6)であげた，共起の可能性の否定情報の記述においては，記述すべき情報の選別が重要である。一つ考えられるのは，類義語に絞って記述するということである。類義表現でありながら，統語的なふるまいの違うものは多いので，それについて書くことは有用な情報に成り得ると考えられる[6]。また，文体（位相）に関する情報も，生成に役立つ情報であろう。このような類義表現の区別や，文体（位相）の情報付与についても，今後検討したい。

[6] 『使い方の分かる類語例解辞典』（小学館）など，市販の類義語辞書には，類義語と文例とを表形式にし，○，△，－を書き，共起の可否を明示する試みがされているものがある。

5．おわりに

解析と生成という視点を同時にもちながら，共起情報を記述していく方法を提案した。この方法によれば，解析用，生成用の辞書の構築が効率化されるばかりではなく，解析と生成に，共通に必要な情報と，個別に必要な情報とに切り分けることによって，これまでに見落とされていたような情報の発見を含め，それぞれに必要な情報をより明確にした記述が可能になる。

筆者は，人手による，限られた例数の共起情報の分類，整理に取り組んでいるが，一方では，大量の共起情報をコーパスから自動抽出する研究が盛んに行われている。バランスの良いデータの大量抽出のモデルとなるようなデータ作成を目指したいと考えている。

参考文献

[EDR93] 日本電子化辞書研究所(1993)『ＥＤＲ電子化辞書仕様説明書』.

[Fillmore75] Fillmore, C. (1975)『格文法の原理：言語の意味と構造』三省堂.

[IPA97a] 情報処理振興事業協会(1997)『CD-ROM版 計算機用日本語基本辞書ＩＰＡＬ－動詞・形容詞・名詞－』.

[IPA97b] 井口厚夫・猪塚元・桑畑(柏野)和佳子・山下智弥(1996)「述語の項としての用法」『計算機用日本語基本名詞辞書ＩＰＡＬ(Basic Nouns)解説編』情報処理振興事業協会, pp. 86-103.

[NTT97] ＮＴＴコミュニケーション科学研究所（監修）(1997)『日本語語彙体系』岩波書店.

[岩波98]『岩波講座言語の科学３　単語と辞書』岩波書店.

[柏野98] 柏野和佳子(1998)「曖昧性解消過程解明のための多義語の分析」『国立国語研究所創立50周年記念研究発表会資料集』pp. 69-72.

[亀井97] 亀井真一郎・田村真子・村木一至(1997)「日本語の用言相当慣用表現の意味空間における分布図」『言語処理学会第3回年次大会論文集』pp. 51-54.

[桑畑他98] 桑畑（柏野）和佳子・橋本三奈子・青山文啓(1998)「IPAL名詞辞書による多義性解消のためのコロケーションの分析」『情報処理学会論文誌』39-6 pp. 1925-1934.

[国研96] 国立国語研究所(1996)『「分類語彙表」形式による語彙分類表（増補版）』.

[城田91] 城田俊(1991)『ことばの縁』リベルタ出版.

[橋本他97] 橋本三奈子・山下智弥・桑畑（柏野）和佳子(1997)「文型テーブルを用いた統合辞書の試作」『ソフトウェア文書のための日本語処理の研究-13』情報処理振興事業協会.

[村田他98] 村田賢一・石田直子・柏野和佳子・常盤僚子・西川賢哉(1998)「計算機用日本語生成辞書 IPAL(SURFACE/DEEP)の試作」情報処理学会第130回自然言語処理研究会.