ＩＰＡＬ統合辞書による多義性解消のためのコロケーションの分析

桑畑和佳子 †[1]　　橋本三奈子 †[1]　　青山文啓‡[2]

†情報処理振興事業協会 (IPA) 技術センター　　‡桜美林大学 国際学部 国際学科

1 はじめに

情報処理振興事業協会 (IPA) では、計算機用日本語基本辞書 IPAL (IPA Lexicon) の構築を進めており、これまでに動詞辞書、形容詞辞書、および名詞辞書を開発して一般に公開している。また、各辞書の記載内容を関連付けて効率的に活用するため、辞書の統合化の検討を進めている。

IPAL の特徴は、共起情報(以下「コロケーション」と呼ぶ) や文型情報などの統語情報を詳細に記載していることである。このため、自然言語理解において問題となる多義性の解消に対して、有効な情報源となることが期待される。しかし、これまでは各辞書に記載されているどの情報をどのように利用すれば多義性が解消できるかの道筋は、必ずしも明らかではなかった。

IPAL によって多義性の解消を図る方法としてまず考えられるのは、梁ら (北陸先端大自然言語処理学講座) によって与えられたような、語義レベルのリンク情報[1] を用いる方法である[3]。「鍋が割れる」という文を例にとると、「鍋」という名詞には、器を指す意味と、中身の料理を指す意味とがあり、また「割れる」という動詞には、物が壊れる意味と、意見などがいくつかに分かれる意味とがある。それぞれに多義であるが、梁らが与えたリンク情報によれば、その「鍋」は器を指すものであり、その「割れる」は物が壊れる意味のものであることがあらかじめ示されているので、計算機による意味の特定も容易である。

このように、辞書の統合のために導入されたリンク情報は、多義性の解消のためにも有効な場合がある。しかし、リンク情報だけでは解消できない多義性が存在することに注意しなければならない。例えば、「骨が折れる」という文の場合には、「骨」に人間の骨の意味と、傘などの骨を指す意味とがあり、さらに句全体で慣用的に「苦労する」という意味を指す場合もあるため、名詞と述語とを結びつけるリンク情報だけでは、人の骨が折れるのか、傘が壊れるのか、また苦労して何かをするのか、いずれの意味にも特定することができない。

本稿では、文の多義性をもたらす要因のうち「コロケーションの曖昧性」に着目する。そして、「骨を折る」のような場合においてもコロケーションの曖昧性を解消するために、IPAL の各辞書に記載されている情報がどのように有効であるかを示す。以下ではまず、IPAL 名詞辞書のコロケーションの記載に着目して、名詞と述語が重複するコロケーションのパターンと、その重複するコロケーションを区別する辞書情報について分析する。次に、この分析に立脚して、曖昧性解消のために必要な辞書情報について考察する。

2 コロケーションの重複の分析

IPAL 名詞辞書には、コロケーション情報として、見出し語である名詞に結びつく述語を示すだけでなく、その意味を捉えるための情報[4]も記載されている [2]。これらの記載がどのような場合にコロケーションの曖昧性を解消できるかを具体的に調べるために、まず IPAL 名詞辞書のコロケーションの記載から「名詞と述語の組合せが重複するコロケーション」を抽出し、続いてその重複例同士の記載を一組ずつ比較した[5]。IPAL 名詞辞書に記載されているコロケーションは延べ 43,176 例、慣用表現は延べ 1,168 例である。そのうち、重複するコロケーションは、延べで 10,415 例、異なりで 4,578 例であった。つまり、平均して 2〜3 個重複して記載されるコロケーションが、4,000 以上もあるという計算になる。

続いて、その重複したコロケーションの重複パターンと、名詞と述語の組合せ以外の情報の記載に異同がどのようにあるかを調べるために、重複例同士のペアを一つずつ比較して、どこが同じでどこが違うかを調査した。この比較結果は、それぞれの重複の中から 2 個ずつ選ぶ組合せの数 (重複例の数を n とすると、${}_nC_2$) だけ得られる。比較結果の総数は、7,623 対であった。以下、この 7,623 対を対象とした比較結果の詳細を述べる。

2.1 コロケーション重複のパターン

IPAL 名詞辞書では、コロケーションは見出し語の区分ごとに、意味素性 [5] 別に記載されているので、違う区分内で、または同じ区分内で、別の素性と、または同じ素性と、つまり、計 4 パターンの形でコロケーションが重複して記載されていた。表 1 に、4 種の重複記載のパターンと出現数を例とともに示す。

名詞と述語の組み合わせが重複していても、実際の IPAL 名詞辞書の記載では、表 1 の例で ( ) でくくって示した部分のように、何らかの表層的な違いが記述されているものが多い。表層的な違いには、表 2 に示す 5 種類がある。

---

[1] 富士通株式会社より出向中

[2] 情報処理振興事業協会技術センターWG委員主査

[3] 梁らによって与えられたリンク情報は、各 IPAL 辞書の統合的利用を目的としたものであって、多義性の解消を直接の目的として検討されたものではない。

[4] 見出し語が立つ格以外の格助詞及びその格に立つ名詞句や、見出し語に先行するまたは後行する名詞句など。

[5] この調査は、FTP 公開中の IPAL 名詞辞書第 3.1 版 (96 年 8 月 1 日付け) より校閲を進めた、97 年 2 月 14 日時点の辞書データ (未公開版) を使用して行ったものである。尚、第 4.1 版は近日中に公開を予定している

表 1: 重複記載の 4 パターン

a.	異なる区分内で同じ素性と (異区同素)	3,608 例
(例)	01:ACT (敵軍の) 攻撃ガ始まる	
	02:ACT (議長への) 攻撃ガ始まる	
b.	異なる区分内で異なる素性と (異区異素)	2,260 例
(例)	01:MEA (大豆の) 収穫ガ多い	
	02:GRA (会議の) 収穫ガ多い	
c.	同じ区分内で同じ素性と (同区同素)	821 例
(例)	01:CON 自転車 (のサドル) ガ高い	
	01:CON 自転車 (の価格) ガ高い	
d.	同じ区分内で異なる素性と (同区異素)	934 例
(例)	01:CON (シャワーデ) 汗ヲ流す	
	01:LIQ (額ニ) 汗ヲ流す	

表 2: 重複記載における表層的な違いの 5 種類

α.	見出し語が立つ格以外の格助詞が違う (他格 1)	2,421 例
(例)	02:CON ～ガ～ニお茶ヲ入れる	
	02:PRO ～ガお茶ヲ入れる	
β.	「他格」に立つ名詞句だけが違う (他格 2)	1,145 例
(例)	02:HUM (控え室) ニ家族ガいる	
	02:ROL (彼女) ニ家族ガいる	
γ.	見出し語の先行句が違う (先行)	4,888 例
(例)	01:ACT ～ガ (原野の) 開拓ヲ行う	
	02:ACT ～ガ (技術の/新市場の) 開拓ヲ行う	
δ.	見出し語の後行句が違う (後行)	91 例
(例)	01:POT のど (の調子) ガいい	
	03:FOR のどガいい (=歌う声がいい)	
ϵ.	見出し語が立つ格助詞が違う (格助)	1,794 例
(例)	01:INT ～ガ舞台ヲ/カラ出る	
	01:SPA ～ガ舞台ニ出る	

(注) 排他的な分類ではないため、合計数は 7,623 例を越えている。

2.2 コロケーション重複の出現タイプ

続いて、種類 α ～ ϵ の表層的な違いが実際にどのように生じているかを調べた。この結果を表 3 に示す。出現タイプは全部で 24 になった。

表 3 から、一番多く重複しているタイプは 2,594 例存在する［タイプ 5］であることがわかる。先行句だけに「1」がある、つまり先行句だけに違いのある、「寿命が長い」と「(電池の／…) 寿命が長い」のようなものが一番多い、ということである。次に多いのは 1,187 ある［タイプ 1］である。これは表層的な違いが全くないコロケーションである。しかし、それは全体の 15.6% であり、残りの 84.4% は、何らかの表層的な違いが記載されているものであることがわかる。

さらに、表 3 に示した［タイプ 1］～［タイプ 24］が、コロケーションの重複記載の 4 パターン (a～d) 別にどのように出現しているかについて調べた結果を表 4 に示す。

表 4 からは、コロケーションの重複が‘異区分’にある場合と‘同区分’にある場合とで、出現数の多いタイプが違っていることがわかる。特に目立つ特徴は、［タイプ 5］(先行句だけが違うもの) は‘異区分’の方の割合が高く、他方、［タイプ 10］(他の格助詞と格助詞が違うもの) は‘同区分’の方の割合が高いという点である。「(プロペラの) 回転ガ速い」と「(頭の) 回転ガ速い」のようなものが［タイプ 5］であり、「～ガえびヲ食べる」「～ヲえびガ食べる」のようなものが［タイプ 10 ］である。

表 3: 表層的な違い 5 種類の細かいタイプ分け

「0」は各部分の内容が同じことを、「1」違うことを表す。ただし、「他格 1/2」欄については、表 2 の「他格 1」の場合を「1」とし、「他格 2」の場合を「1*」で表す。

	表層的な違い 5 種類					
タイプ	他格 1/2	先行	後行	格助	出現数	(%)
1	0	0	0	0	1,187	(15.6)
2	0	0	0	1	77	(1.0)
3	0	0	1	0	20	(0.3)
4	0	0	1	1	2	(0.0)
5	0	1	0	0	2,594	(34.0)
6	0	1	0	1	162	(2.1)
7	0	1	1	0	13	(0.2)
8	0	1	1	1	2	(0.0)
9	1	0	0	0	330	(4.3)
10	1	0	0	1	533	(7.0)
11	1	0	1	0	12	(0.2)
12	1	0	1	1	34	(0.4)
13	1	1	0	0	594	(7.8)
14	1	1	0	1	913	(12.0)
15	1	1	1	0	2	(0.0)
16	1	1	1	1	3	(0.0)
17	1*	0	0	0	524	(6.9)
18	1*	0	0	1	15	(0.2)
19	1*	0	1	0	0	(0.0)
20	1*	0	1	1	1	(0.0)
21	1*	1	0	0	551	(7.2)
22	1*	1	0	1	52	(0.7)
23	1*	1	1	0	2	(0.0)
24	1*	1	1	1	0	(0.0)
計					7,623	(100)

コロケーションが‘異区分’で重複している時は、コロケーションそのものが多義である場合と、名詞だけが多義で述語の方は多義でない場合とがあり、‘同区分’内で重複している時は、名詞に多義性は無く、述語が多義である場合がほとんどである。つまり、表 4 から、名詞が多義である場合には、文型の違いよりも先行句の違いや他の格助詞に立つ名詞句の違いの方が顕著であり、述語が多義である場合には文型の違いの方が顕著である、という結論が得られる。

3 コロケーションの曖昧性を解消する情報

これまでの分析から、重複するコロケーションを区別するのに、「先行句」「文型」「他の格助詞に立つ名詞句」の違いが利用できそうな見通しが得られる。このうち、「文型」や「他の格助詞に立つ名詞句」が区別の手がかりになることは、特に新規な知見ではない。これらはこれまで IPAL 動詞辞書や形容詞辞書でも記載されてきたものであり、また、一般的に多義性を解消しようとする際によく用いられているからである。しかし、IPAL 名詞辞書で記載した「先行句」が、重複するコロケーションを区別するものとして最も数多く出現するということは、今回の分析で初めて確認できたことであった。そこ

表 4: 重複記載パターンと表層的な違いタイプ別の出現数

() 内の数字は、その出現数が各パターンごとに占める割合を % で表したものである。

タイプ	重複記載の 4 パターン			
	a. 異区同素	b. 異区異素	c. 同区同素	d. 同区異素
1	819(22.7)	198(8.8)	0(0.0)	170(18.2)
2	7(0.2)	14(0.6)	43(5.2)	13(1.4)
3	9(0.2)	2(0.1)	6(0.7)	3(0.3)
4	0(0.0)	1(0.0)	1(0.1)	0(0.0)
5	1,698(47.1)	754(33.4)	25(3.0)	117(12.5)
6	65(1.7)	67(3.0)	22(2.7)	8(0.9)
7	8(0.2)	4(0.2)	0(0.0)	1(0.1)
8	1(0.0)	1(0.0)	0(0.0)	0(0.0)
9	63(1.7)	140(6.2)	57(7.0)	70(7.5)
10	27(0.7)	83(3.7)	298(36.3)	125(13.4)
11	2(0.1)	8(0.4)	2(0.2)	0(0.0)
12	4(0.1)	5(0.2)	19(2.3)	6(0.6)
13	190(5.3)	249(11.0)	61(7.4)	94(10.1)
14	114(3.2)	333(14.7)	248(30.2)	218(23.3)
15	0(0.0)	2(0.1)	0(0.0)	0(0.0)
16	0(0.0)	1(0.0)	0(0.0)	2(0.2)
17	375(10.4)	113(5.0)	5(0.6)	31(3.3)
18	1(0.0)	6(0.3)	4(0.5)	4(0.4)
19	0(0.0)	0(0.0)	0(0.0)	0(0.0)
20	0(0.0)	1(0.0)	0(0.0)	0(0.0)
21	220(6.1)	236(10.4)	30(3.7)	65(7.0)
22	5(0.1)	40(1.8)	0(0.0)	7(0.7)
23	0(0.0)	2(0.1)	0(0.0)	0(0.0)
24	0(0.0)	1(0.0)	0(0.0)	0(0.0)
計	3,608(100)	2,260(100)	821(100)	934(100)

で以下では、その「先行句」に的を絞り、「先行句」がどのようにコロケーションの曖昧性解消に有効であるかを [タイプ 5] (先行句だけが違うもの) の重複例で検証した結果について述べる。

コロケーションの重複が生じる主な要因の一つに、「そのコロケーション自体が比喩的意味を持つため」という場合がある [3]。[タイプ 5] の '異区分' に分類されたほとんどの例が、比喩的な意味が派生していることに関わる重複例であった。例えば、以下に示す例は、いずれも名詞の区分間にメタファーによる意味の拡張関係が認められる [6] ものである。

1) 01:GRA (口紅の／…) 色ガ濃い
05:GRA 敗戦の／…色ガ濃い
2) 01:PHE 波ガ激しい
02:ABS (感情の／…) 波ガ激しい
03:ABS (観光開発の／…) 波ガ激しい
3) 01:LOC (電車の／前方の／…) 席ガ空く
02:SPA (課長の／…) 席ガ空く
4) 01:CON (この家の／…) 天井ガ高い
02:STA (株価の／…) 天井ガ高い

以上 1) から 4) に示す記載例[6]では、「先行句」の記載を参照することによってうまく曖昧性が解消できるように見える。実際のコーパスではどのような表現が現れているのだろうか。「色が濃い」で IPA 所有のコーパス (新聞、文芸作品、マニュアルなど約 47 万文) を検索した結果、5)〜10) の実例が見つかった。

5) 《レモンティーの場合はレモンによって香味は増すが，紅茶の【色が濃】くなる。》
6) 《老油は濃口しょうゆに似て【色が濃】い。》
7) 《スペインと米国による約四百年の植民地支配の【色が濃】いのだ。》
8) 《不便を強いられる市民には疲労の【色が濃】いが、そんな中で水道だけは二日中に復旧する兆しが見えてきた。》
9) 《イカ漁シーズンを控え、漁師にはあせりの【色が濃】い。》
10) 《全戸に配られたチラシには「決して強制ではありません」と記してあるが、行政区長や隣組長を総動員した寄付集めは、実質的に強制の【色が濃】くなる。》

1) の記載にある先行句「口紅の」、「敗戦の」と、5)〜10) に現れる先行句とを比較することで、5)〜10) の「色が濃い」が視覚的な「色」に言及しているのか、状況の度合について比喩的に言及しているのか判断できることがわかる。6) には、先行句がないが、先行句を省略できるのは視覚的な「色」の方であるという情報を用いることで、その意味を特定することができる。

続いて、本稿の冒頭にあげた「骨が折れる」の記載を 11) に、その実例を 12)〜15) に示す。

11) 慣用表現: 骨が折れる
01:CON (手の／足の／…) 骨が折れる
02:CON (傘の／…) 骨が折れる
12) 《荒井さんは左足の【骨が折れ】、約三カ月のけが。》
13) 《一色さんは頭の【骨が折れ】て重体。》
14) 《ロシアを説得することが一番【骨が折れ】た》
15) 《根のない話を虚空に描き出すのには【骨が折れ】る。》
16) 《【骨が折れ】とるかも知れんな。》

12)、13) は、「左足の」「頭の」と、身体部位を表す先行句を伴っているので、01 の意味だとすぐわかる。一方、14)、15) は、身体部位を表す先行句や、「傘」などの先行句を伴わないこと、さらに「一番」などの程度副詞が現れることから慣用表現であると判断できる。以上のことから実例にある曖昧性を解消するのに、「先行句」の情報にかなりの有用性を認めることができる。

ところで、17) は、16) 同様に先行句がないが慣用表現ではない。01 か 02 であるかの曖昧性は残るが、慣用表現でないことはわかる。それは、「骨が折れている」というテイル形から判断できる。「骨が折れる」の場合、慣用表現として用いられる時はテイル形をとらない。こういったテイル形の情報は、IPAL 動詞辞書では記載していたものであるが IPAL 名詞辞書では記載していない。曖昧性を解消する辞書情報を考える時、このような情報、すなわち動詞のアスペクト情報についても留意すべきである。

さらにもう一例挙げる。「足跡を残す」の IPAL 辞書記載例を 17) に、実例を 18)〜20) に示す。

18) は「鮮明な」で形容されているいることから 01 であることが、19) は「輝かしい」で形容されていることから 02 であることが判断できる。しかし、20) は「大き

[6] 先行句に ( ) のないものはその句が必須であることを示す。

17) 01:RES ～ガ～ニ足跡を残す
03:RES ～ガ～ニ(輝かしい／…) 足跡を残す
18) 《しかも、選りに選って雪の夜、降り積もった裏庭から侵入して、かくも鮮明な【足跡を残】して去ったというわけか？》
19) 《中村校長らの不安をよそに二人は輝かしい【足跡を残】した。》
20) 《時代感覚を鋭敏に反映する作品で昭和初年代の新詩運動に大きな【足跡を残】した。》

な足跡」だけでは 01 であるのか 02 であるのか判断に迷うところである。実はこの場合は、「ニ格に立つ名詞句」の方を参照すれば良い。「新詩運動に」という部分を見れば、20) が 02 であることは明らかである。恐らく「足跡を残す」の場合は、先行句よりはニ格に立つ名詞句の方により曖昧性を解消できるものが現れると考えられる。しかし、19) のようにニ格に立つ名詞句が現れず、先行句が手がかりになることもあり、やはり先行句の利用の有効性は高いと言える。

以上のことから、これまで注目されていた「文型」や「格に立つ名詞句」例ばかりではなく、「先行句」例についての辞書記載を積極的に利用することが曖昧性の解消に役立つと言える。

4 考察

「先行句」をはじめとするコロケーションに付随する情報を参照することが曖昧性の解消に役立つことはわかったが、いくつか注意すべき点がある。まず、名詞辞書でうまく区別しやすい先行句が記載できていたとしても、実際の文章中には型通りには現れてはくれないものがあるという点である。次に、曖昧性をなくすような辞書記載が最初から困難な場合があるという点である。以下に例示する。

21a) 01:RES ～ガ(友人の／…) 成功ヲうらやむ
02:RES ～ガ(彼の／…) 成功ヲうらやむ
21b) 01:RES ～ガ(実験の／計画の／…) 成功ヲ祈る
02:RES ～ガ(彼の／…) 成功ヲうらやむ

21a,b) は、もともと 02 の意味を特定するのが困難な場合である。01 は「物事が思い通りにいき、目的を達成すること」を指し、02 は「富や地位を手に入れること」を指す。01 の方は、実際にコーパスをひくと、「企画の成功」「サミットの成功」「上映の成功」といった実例があり、21a) の先行句の記載例を訂正して、21b) のように記載すべきであることがわかるが、02 の方は、具体的な先行句が存在しにくい。コーパス中、02 の意味ととれるものが一例だけ見つかった。それは、

> 《カプリ島に別荘を持つのはナポリ市民の夢だが、島の中心地に九室もある大きな別荘を購入したことは、ミケーレ弁護士の地位と【成功を】物語るものだ。》

というものである。この時の先行句は「ミケーレ弁護士」であるが、この先行句だけを見る限りでは 01 の可能性も否定できない。02 の意味を特定するものは、「成功」に並列する「地位」という名詞句からの連想や、「島の中心に大きな別荘を購入した」という文脈でとしかいいようがない。

そもそも、IPAL 名詞辞書に重複記載されたコロケーションの中には、表層的な違いが記載されていないものも存在していた。文脈で判断するしかないコロケーションの曖昧性解消は、現在の IPAL 辞書のコロケーションの記載だけでは困難である。コロケーションの出現するテキスト全体まで見渡した辞書記載の検討は、今後の課題である。

5 おわりに

以上、文中に現れる多義性を解消する際に問題になるコロケーションの曖昧性を、重複記載の観点から検討した。重複するコロケーションを区別するものについて調べた結果、名詞と述語の組み合わせが同一になる時、述語が多義である場合は文型に違いがあることが多く、コロケーション自身が多義である時や、名詞が多義である時は先行句や他の格助詞に立つ名詞句の違いがあることが多いことを定量的に示した。このことから、コロケーションの曖昧性を解消するには、名詞と述語以外に辞書に記載されている表層的な違いを用いることが有効であることを明らかにした。特に、表層的な違いのうち、先行句の違いに高い利用価値があることを示した。

現在 IPA 統合辞書プロジェクトでは、単文に現れる単語についての詳細な情報を参照できるような文型情報欄のフォーマットを作成中であるが、本稿の検討結果を踏まえ、そのフォーマットには「先行句」の記載欄を設けている [4]。

謝辞 IPAL 共同研究のメンバーである、WG 委員、臨時 WG 委員の方々、並びに、補助作業をして下さったアルバイトの方々に感謝いたします。また、IPAL ハイパーテキスト化にご尽力下さった北陸先端大学の奥村学助教授、梁慶昇氏、望月源氏に感謝申し上げます。

参考文献

[1] 梁慶昇: IPAL 辞書の自動的ハイパーテキスト化, 北陸先端科学技術大学院大学 修士論文 (1996).

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

[3] 桑畑和佳子, 橋本三奈子, 村田賢一: IPAL ハイパーテキスト化のためのコロケーションの重複記載の分析, 情報処理振興事業協会 第 15 回技術発表会論文集, pp.147-153 (1996).

[4] 橋本三奈子, 桑畑和佳子, 村田賢一: IPAL 統合辞書: 単文を中心とした仕様, 情報処理振興事業協会 第 15 回技術発表会論文集, pp.135-145 (1996).

[5] 青山文啓: 意味素性, 情報処理振興事業協会 計算機用日本語基本名詞辞書 IPAL (Basic Nouns) 解説編, pp.31-61 (1996).

[6] 桑畑和佳子, 本多啓: 区分間の意味的関係, 情報処理振興事業協会 計算機用日本語基本名詞辞書 IPAL (Basic Nouns) 解説編, pp.211-227 (1996).