ＩＰＡＬハイパーテキスト化のためのコロケーションの多義性の分析

7Ｌ－2

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

ＩＰＡ技術センターでは、これまでに動詞辞書(861語)、形容詞辞書(136語)、並びに、名詞辞書(1,081語)を開発し、公開している。名詞辞書では、一つの見出し語を区分したうえで、共起する述語（コロケーション）の用例を意味素性[1]別に分類して記載している。ＩＰＡＬハイパーテキスト化[2]に際して、動詞辞書、形容詞辞書と名詞辞書との間に自動的にリンクを張る時、該当区分を決定するのにこのコロケーションの情報を参照する。この時、コロケーションの中に同一のものが複数個あるものがあると、リンク先が一意に定まらなくなる。そこで、最初に、名詞と述語のペアが慣用表現と重複するものを調べた結果、72例の慣用表現が重複するものであることがわかった。続いて、その72例と記載されている全てのコロケーション43,362例とを足して異なりを算出したところ、37,590であった。そして、重複するものが、延べで10,561、異なりで4,619もあることがわかった。つまり、2、3個重複するコロケーションが4,000以上もあるということである。本稿では、重複例を分析してわかった重複のパターンと重複の要因について述べる。

2　コロケーションの重複記載のパターン

ＩＰＡＬ名詞辞書では、名詞コロケーションに関する情報として、その名詞と結びつく述語を名詞の区分ごとに意味素性別に収録している[3]。また、単語の連鎖全体が一つの意味をもつ慣用句は、コロケーションから除外し《慣用表現》の欄で取り上げている。例えば、「鼻が高い」と言う時、実際に鼻の高さが高いことを言う場合と、「得意になっている」という意味で用いられる場合とがある。前者が「鼻」のコロケーションであり、後者が「鼻」の慣用表現である。このような時、同じ名詞と述語のペアが、コロケーションの例としても、慣用表現の例としても、重複して記載されている。同様に、「足を洗う」「頭を下げる」「骨を埋める」「しっぽを振る」などもコロケーションとしても慣用表現としても用いられるものである。

コロケーションの重複は慣用表現とだけではない。コロケーションは、区分ごとに意味素性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パターンの形で重複している可能性がある。実際、いずれのパターンの重複例も見つかった。以下に重複パターンを例とともに示す(数字は区分番号、英文字３字は素性を表す）。

a．違う区分内で同じ素性と
　01:ACT （敵軍の）攻撃ガ始まる
　02:ACT （大統領への）攻撃ガ始まる
b．違う区分内で別の素性と
　01:MEA （大豆の）収穫ガ多い
　02:GRA （会議の）収穫ガ多い
c．同じ区分内で別の素性と
　01:CON （シャワーデ）汗ヲ流す
　01:LIQ （額ニ）汗ヲ流す
d．同じ区分内で同じ素性と
　01:CON 自転車（のサドル）ガ高い
　01:CON 自転車（の価格）ガ高い

尚、名詞と述語のペアが重複していても、実際の辞書記述では、上記で（）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あるものが多い。表層的な違いのパターンには、以下のものがある。これらは組合わさって現れる場合もある。

1．格形式が違う
　01:AML ～ヲえびガ食べる
　02:EDI ～ガえびヲ食べる
2．見出し語が立つ格以外の格に立つ名詞句が違う
　02:HUM （控え室）ニ家族ガいる
　02:ROL （彼女）ニ家族ガいる
3．見出し語の先行句が違う
　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
4．見出し語の後行句が違う
　01:POT のど（の調子）ガいい
　03:FOR のどガいい（＝歌う声がいい）

3　コロケーションの重複記載の要因

なぜ、重複するコロケーションが4,000以上もあるのであろうか。重複する要因として、次の４点が考えられる。

A．その区分間の語義に大きな違いがないため
B．その素性に特有の述語であるため
C．その述語の使用頻度が高いため
D．そのコロケーション自体が比喩的意味も持つため

Towards an IPAL Hypertext version : Analyzing polysemous collocations

†Wakako KUWAHATA, et al.,
Software Technology Center,
Information-technology Promotion Agency, Japan
3-1-38 Shiba-koen, Minato-ku, Tokyo, 105 JAPAN
[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

要因A：例えば、見出し語「しゅじい【主治医】」は、区分０１は、「［文脈の中で示される］ある患者の治療にあたる医師たちのうち、中心となる医師」であり、区分０２は、「［文脈の中で示される］ある人や団体のかかりつけの医者」である。このように、区分間に微妙な違いしかないと、それぞれに収録されるコロケーションはほぼ同じものになる。

要因B：「その素性に特有の述語」であれば、コーパスにも現れやすく、また、コーパスにたまたま出現していない場合にも、結びつく可能性の高い述語として積極的に収録してあるので、各区分に同じ意味素性がある場合、それぞれに同じ述語が重複する可能性は高い。例えば、〈LOC〉であれば、「行く，戻る，広い，狭い」、〈ACT〉であれば、「始まる，終わる，止まる，早い，遅い」といったものが、素性に特有の述語である。冒頭であげた「（敵軍の）攻撃ガ始まる」、「（大統領への）攻撃ガ始まる」の重複例が、その一例である。

また、中には二つ以上の素性に特有の述語もある。例えば、〈MEA〉は計れるもの、〈GRA〉は計れないものに、という違いはあるものの、どちらもその量の多少に焦点をあてる意味素性である。冒頭の「（大豆の）収穫ガ多い」「（会議の）収穫ガ多い」の例で示した「多い」という述語は、〈MEA〉と〈GRA〉のどちらの素性にも特有の述語であるため、重複が生じている。

要因C：例えば、「ある」「ない」「する」「なる」といった述語は多義であるので、あらゆる見出し語と結びつきやすく、よって、重複する数も多くなる。これらの述語の重複は、延べで「ある:815，ない:772，する:408，なる: 323，計2,318」、異なりで「ある:309，ない:299，する:164，なる:129，計901」であった。先にも述べた通り、重複した異なりは全体で4,619であるので、以上の４語だけでその約５分の１を占めていることになる。

要因D：見出し語「あせ【汗】」の場合、区分０１「暑い時やひどく緊張したときなどに、体内から皮膚上に出てくる水分」と、区分０２「物の表面にできた結露」との二つに区分される。この時、区分０２は、区分０１のメタファーによって意味が拡張したものだと捉えられる[4][1]。このように区分間にメタファーの関係があると、コロケーションが重複しやすい。区分０２に記載したコロケーションは、「（ハム）ニ汗ガ出る」と「（ミカン）ガ汗ヲかく」の２例であるが、「汗ガ出る」、「汗ヲかく」のいずれも区分０１その記載がある。さらに、「汗をかく」全体で「尽力する」という慣用的な意味もあるため、慣用表現の例とも重複する。

区分間にメタファーの関係があるものの例をさらに挙げると、見出し語「あめ【雨】」は区分０１「大気中の水蒸気が上空で冷やされ、水滴となって落ちてくるもの。また、それが降り続いている状況」と区分０２「上から絶え間なくたくさん降り注ぐもの」に分かれ、区分０２の「～ニ火の／血の雨ガ降る，降り注ぐ」が区分０１の「雨ガ降る」「～カラ～ヘ／ニ雨ガ降り注ぐ」と重複する。また、見出し語「まど【窓】」は、区分０１「建物の内部から外をみたり明かりを採ったりするためのもの」と区分０２「内と外の境界にあって、内部のものを外部に見せるところ」に分かれ、区分０２の「～ガ～ニ心の／話し合いの／窓ヲ開く，開ける，閉ざす」が、区分０１の「～ガ窓ヲ開く，開ける，閉ざす」と重複している。

しかしながら、メタファーの関係があっても、コロケーションが重複しないものもある。その一例を示そう。見出し語「ふね【船，舟】」は、区分０１「人や荷物をのせて海や川などの水面を移動する乗り物」と区分０２「料理屋や旅館などで刺身や貝などをもりつけて食膳にだす、船の形をかたどった入れ物」の二つに区分されるが、メタファーである区分０２のコロケーション「～ガ（刺身の）舟ヲ注文する」や「～ガ（刺身）ヲ舟ニ盛る」などは、区分０１と重なるものではない。ただ、「舟ガ大きい，小さい」は区分０１と０２とに重複しているが、これは、〈CON〉に特有の述語であるので、先に述べた「要因B」で説明される例に相当すると考える。

4 おわりに

ＩＰＡＬハイパーテキスト化の自動リンク付けの失敗を減らすために、ＩＰＡＬのコロケーションの重複記載について分析をした。重複がどこと生じているか、重複しているものの表層的な違いは何か、また、なぜ重複が生じるのかについて明らかにした。今後も分析を続け、さらに詳細な分析結果を作成する予定でいる。そして、自然言語処理全般においても問題となる、コロケーションの多義性解消方法まで探る予定でいる。

謝辞

コロケーションの記述を担当した山下智弥氏、木田敦子氏、佐藤幸子氏をはじめ、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。

参考文献

[1] 青山文啓.「素性に基づく名詞記述のための枠組」.『ＩＰＡＬシンポジウム '95 論文集』, pp. 1–9, 1995.

[2] 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情報科学研究科情報処理学専攻修士論文, 1996.

[3] 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」.『計算機用日本語基本名詞辞書ＩＰＡＬ(Basic Nouns)解説編』, pp. 86–103, 1996.

[4] 桑畑和佳子, 本多啓.「区分間の意味的関係」.『計算機用日本語基本名詞辞書ＩＰＡＬ(Basic Nouns)解説編』, pp. 211–227, 1996.

[1] 比喩的多義には、メタファーの他、「メトニミー」と「シネクドキー」があるが、メトニミーにはあまり重複はなく、シネクドキーにも、メタファーほど目立った重複が見られなかった。比喩的多義と重複の関係については今後さらに細かく分析する予定でいる。