WWW を丸ごとデータにすると何がわかるか?
格フレーム辞書を言語研究に利用する

黒田 航　　李 在鎬

1　はじめに

自然言語処理 (NLP) の技術の進歩により，Web データをコーパスとして扱う可能性が現実化し，その流れは一層堅固になって来ている．Web データをコーパスとして使うことへの批判もないわけではないが，使い方次第では利点の方が多いように思う．(1) に Web コーパスの長所を，(2) に短所を幾つか挙げた:[1)]

(1)　a. 電子化 (e.g.,「書き起こし」) の手間が不要である．
　　b. データの規模が (おそらく十分に) 大きい．
　　c. 内容/用例が (おそらく十分に) 多様である．
(2)　a. 量と内容が共に常に変化し，動的である．
　　b.「ノイズ」が多い．
　　c. 代表性に欠ける．

(1b) の利点と (1c) の利点は相関している (これは (1c) が (1b) を前提にするためである)．これは既存のコーパス (例えば新聞コーパス) に較べて明らか優位である．Web コーパスとの比較を通じて新聞記事に使われる言語は

[1)] より詳細な検討は前川 [6] などを参照のこと．

文体の面 (cf. 伊藤 [5]) でも語彙や語義の分布の面 (cf. 竹内 [10]) でも偏っているという事実が明らかになりつつある．

当然，Web データに欠点はある．(2a) が理由で，例えば正確な統計量 (語数) を測定できない．(2b) の問題は，一定の割合で誤用，逸脱用法 (e.g., 2ch の言語)，道徳的に好ましくない内容を表わした文章が含まれる点にある (これらを自動的に除外することは難しい)．(2c) の問題は，Web コーパスは新聞コーパスに較べ比較的多様なデータを含んでいるとは言え (話しコトバを含まないので) 均衡コーパスではないし，コンピュータで文章を書く習慣をもつ人々の言語実態しか対象にならない．

(1) の利点と (2) の弱点は表裏一体であり，(2) の弱点を容認することなしに (1) の利点を手に入れることは原理的に不可能である．となれば，Web データをコーパスとして利用することの是非は，研究目的に合っているか否かで評価するしかない．研究の目的が日本語の正確な統計的実態の調査であれば，(2a) は本質的な難点となるだろうが，用例研究とその先にある意味研究の見地からは (1c) の効果は絶大である．

以上の問題意識の下，本稿では次の二つのことをする: i) 長谷部 [22] による Wikipedia 日本語版をコーパス化する試みを §2 で紹介し，ii) それに続く §3 以下で河原・黒橋 [19, 20] によって開発され，`http://nlp.kuee.kyoto-u.ac.jp/nl-resource/caseframe.html` で一般公開されている「Web から自動構築された格フレーム辞書」(以下，単に「格フレーム辞書」) を使った言語研究を紹介する．§4 は李ほか [24] の格フレーム辞書の元になった Web コーパスを使った実証研究の紹介，§5 は黒田ほか [11] の現行の格フレーム辞書の精度評価の研究である．

2　Wikipedia 日本語版のコーパス化

日本語には著作権で保護されていない規模の大きなコーパスがない．これは日本語のコーパス基盤言語学がなかなか進展しない原因の一つであ

る．この現状を改善する試みの一つとして，長谷部 [22] は，Wikipedia の日本語版が比較的規模の大きな，著作権上の制約のない日本語コーパスとしての利用を可能にする WP2TXT と Mconc というツールを開発し，`http://yohasebe.com/` で公開している．

WP2TXT は，Wikipedia のオリジナルデータを bz2 を展開しつつ，xml データをスキャンして title 要素およびテキスト要素内のデータを抽出する．これは生データから不要な wiki タグと html タグを除去しテキストに変換する処理である (一般に構造化された Web データの平テキストへの整形の自動化は重要な処理である)．Mconc は WP2TXT で処理したテキストデータを形態素解析し，ユーザが指定する文字列や形態素のパターンを抽出する機能をもつ．形態素解析は MeCab[2] を用いて行っている．具体的には次の処理を行う．まずテキストデータを読み込んで文の単位に分解し，形態素解析を行う．そして出力結果を設定ファイルにあらかじめ記述された正規表現パターンに照らし合わせ，データの取捨選択を行い，抽出された文字列と詳細な形態素情報とを csv (comma separated value) ファイルとして出力する．

3 「格フレーム辞書」とは?

Web から自動構築した大規模格フレーム辞書とは，(i) 5 億文の Web データの一文一文を形態素解析し，(ii) それに「係り受け解析」と呼ばれる構文解析を行なって名詞句と述語の間の述語/項関係=依存関係を同定し，(iii) こうして得られた項と述語の対の集合を，項の意味の類似度に従って意味グループ化したデータ[3]である．この処理の際に具体的に行われることを (3) を例にして解説しておく．

---

[2] `http://mecab.sourceforge.net/`

[3] 現状では類似度の計算にシソーラスの情報を使っているが，それを使わない方法への移行を進めているとのことである．現行の構築法でグループ化に使っている類似度の指標は，述語の直前にある格要素 (直前格) の名詞の類似度である．

(3) 次郎は太郎を頼って東京に出た.

(3) の文を形態素解析プログラム JUMAN[4] で形態素解析し，その結果を係り受け解析プログラム KNP[5] に渡して係り受け解析して得られる解析結果は次の (4) である:

(4) a. 次郎は──┐
b. 太郎を──┐ │
c. 頼って──┤
d. 東京に──┤
e. 出た.

この出力は (i)「次郎は」は「出た」に係り，(ii)「太郎を」は「頼って」に係り，(iii)「頼って」が「出た」に係り，(iv)「東京に」が「出た」に係っていることを表わしている．多くの例に対して同様の処理を行なうことで，s_1 = { [次郎は, 出た], [東京に, 出た] }[6], s_2 = { [太郎を, 頼って] } のような項と述語の対 [Arg, Pred] の集合の集合 $S = \{ s_1, s_2, \ldots \}$ が得られる．S の部分集合のグループ化の結果が格フレーム辞書である.

3.1 格フレーム辞書でどんな検索ができるか

(5) a. 基本的な検索: 述語 (i.e., 動詞，形容詞，形容動詞) を入力することで，それを主要部とする格フレームが一覧で表示される．それぞれの格フレームには格要素として現れる語の一覧が表示される (頻度 2 以上，上位 20 位まで).

[4] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html
[5] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html)
[6] 依存構造の定義を尊重すれば厳密には s_1 は [頼って, 出た] も含むべきだが，現行の格フレーム辞書では “V の連用形 + テ” を項/述語対の集合に含めていない.

b. 原文の参照: 格フレームの検索結果で格要素をクリックすると原文が一覧で表示される.
c. 格要素との組み合わせでの検索: 「裏目に出る」のように格要素と述語の組み合わせを (スペースを空けて) 入力すれば，特定の格要素を含む格フレームが検索可能.
d. 受身形/使役形での検索: 検索語の後に “:P” をつけることで受身形の検索が，“:C” をつけることで使役形の検索が可能.
e. 名詞の共起関係の検索: 「名詞から検索」チェックボックスをオンにして名詞を項に取る述語が検索可能 (表示されるのは頻度 2 以上現れる格フレームの一覧). リンクをクリックすることで，その格フレームの詳細が表示される.

3.2 係り受け解析に関する幾つかの注意

係り受け解析は広い意味での依存構造解析 (dependency (structure) parsing) の一種であり，その解析の出力となる係り受け関係の集合は基本的には依存関係=述語/項関係の集合であるが，次の点には注意が必要である:[7)]

(6) 助詞 (例えばニやデ) の曖昧性は解消されていない (このため，述語のニ格とデ格の分類精度はヲ格に較べて非常に低い (§5.1.2 で後述).
(7) [次郎は, 上京した] のような主題マーカーのハで格助詞が非明示化されている係り受け関係は，格フレーム辞書構築の候補集合から除かれている (ガ格の生起数は一般に期待より少なくなる傾向があり，§4 で後述する事実の一部の (13b) に影響が出ていると思われる)[8)].
(8) 格フレーム辞書で名詞と述語との共起も調べることができるが (複合

7) 係り受け解析と依存構造解析の同一視には問題がある．詳細は黒田・飯田 [12] を参照.
8) 「次郎は」が「頼って」の (ガ格の) 項であることは係り受け解析では明示的には表現されず，(“V_1 て V_2” のような) 並列構造の処理などを経て復元する必要がある.

名詞の分析基準に問題があって)，検索には些か不合理な制約がある(例えば「犯人」や「防人」は検索できるが「料理人」や「使用人」は検索できない．同じ理由で「御者」や「患者」は検索できるが「逮捕者」や「追跡者」は検索できない[9]．

4　格フレーム辞書を使った連体節の使用実態調査

李ほか [24] は，日本語の連体節の実態調査を行なうため，11 個の述語(「恐れる」「積む」「冷やす」「寄与する」「固まる」「散る」「心掛ける」「賛成する」「逃げる」「逃れる」「躍進する」) の連体表現 5705 例を格フレーム辞書の元コーパスから機械的に抽出し，人手で構造を分析した．これから Keenan & Comerie [3] や 井上 [23] の提唱した関係節化の階層が，全体の傾向としては記述的に妥当であることが示唆されたが，その一方で語彙項目(かそのタイプ) によって異なる分布を示していることも明らかになった．これは文法レベルでの一般化とは別に語彙項目の特性を踏まえた，構文的なレベルでの一般化も必要なことを示唆する結果である．

4.1　先行研究と問題の所在

従来，日本語の連体 (修飾) 表現 (広義の関係節) をめぐっては記述的・理論的観点から様々な考察がなされてきた (例えば久野 [4]，奥津 [7]，井上 [23]，寺村 [15]，三原 [9]，加藤 [16, 17] など)．本節では，先行研究の指摘を踏まえながら日本語の連体 (修飾) 表現を記述する上で重要な視点について述べると同時に，本研究の問題提起を行う．

日本語の連体 (修飾) 表現の初めての包括的研究として寺村 [15] (元の論文は 1975–1978) が挙げられる．その記述的一般化の中心軸として (9a) の「内

[9] これは現在，開発者の河原大輔 (NICT) 氏が対応策を実装している最中である．

表 1　格関係に基づく装定化の可否

区分	装定化の可否
ガ格	可能
ヲ格	可能
ニ格	一部不可能 (原因 はやや難，変化の結果や判断基準は不可)
ヘ格	可能
デ格	ほとんど可能 (範囲 (クラスで一番背が高い) は不可)
カラ格	部分的に可能 (出発点，原因，起点は無理)
ト格	部分的に可能 (動詞の項であれば可能だが，非項は無理)
ノ	可能

の関係」と (9b) の「外の関係」による対比はよく知られている:

(9)　a. 秋刀魚を焼く男
　　b. 秋刀魚を焼くにおい

(9a) は連体先の名詞が連体節内の用言に対して補語の関係にあったと考えられることから「内の関係」と呼び，一方の (9b) はそのような関係が成立しないことから「外の関係」と呼ばれている.

内の関係に関する先行研究の観察として，寺村 [14] では連体 (修飾) 節の可否をめぐり，表 1 の観察を示した (以下では特定の名詞を連体 (修飾) 節化することを装定化と呼び，その逆を述定化と呼ぶ).

表 1 の観察をより精緻的に捉えた研究として Keenan & Comerie [3] と井上 [23] が挙げられる．K&C [3] は 50 の言語の類型的調査から装定化の容易さの順序を表わす階層として (10a) を提案している[10]．井上 [23] は日本語の形態格の曖昧性を考慮した階層として (10b) を提案している.

[10] この階層性を形成する要因として認知的な理解しやすさなどが挙げられており，使用頻度もこの階層に沿うと考えられている

(10) a. 主語 > 直接目的語 > 間接目的語 > 前置詞の目的語 > 所有格 > 比較級の目的語

b. 主格 > 直接目的格 > 間接目的格 > 位置格ニ > 位置格ヲ > 目標格ニまたはヘ > 位置格デ > 助格デ > 基準格デ > 奪格 > 所有格 > 基点格 > 随格 > 理由格 > 比較格

(10) の階層性をめぐっては英語を中心とした一部の言語では使用頻度の調査 (例えば Keenan [2] や Fox [1]) が行われていたが，日本語に関しては丸元・乾 [18] による内の関係に対する計量的調査はあるが，外の関係も含めた網羅的実態調査は行われていない．

寺村 [14] では外の関係に関する分類を試み，構造的相違が明確なタイプとして (11) と (12) の対比を指摘している:

(11) a. 選挙に出る考え

b. 赤ちゃんが笑っている写真

(12) a. タバコを買ったおつり

b. 文子が座ったうしろ

(11) は被修飾名詞の内容を説明するものであることから「内容節」と呼ばれている．(12) は「逆補充」と呼ばれ，被修飾名詞が持つ相対的概念を利用した連体 (修飾) 表現の一つとされ，日本語特有の構造だと指摘されている．

これまでの日本語学・言語学の知見はいずれも日本語の構造を理解する上では重要であり，多くの研究成果が上がっているが，問題がないわけではない．このような問題が意識される一方で，個々の一般化が日本語の使用例に対してどの程度まで妥当かを評価した研究はない．その理由の一つは，評価のための有用な資料が存在しなかったからである．この難点は格フレーム辞書を使うことで解消されると考えられる．

4.2　データの収集法と解析法

李ほか [24] は次の方法でデータを収集した．1) 格フレーム辞書の元コーパスから，11 個の述語の連体 (修飾) 表現を抽出し，その中から「～した A の B」のように係り先に曖昧性がある表現を除くことによって，分析対象とするデータを作成した[11)]．2) 人手で修飾構造の特定を行った．分析の際には，先行研究の指摘に従って構造的タイプとして，内の関係，内容節，逆補充の 3 タイプに分類すると同時に，連体節内での格関係を特定した．本研究では，加藤 [16, 17] の示唆に従い，格関係間の排他性は仮定していない．

4.3　結果と考察

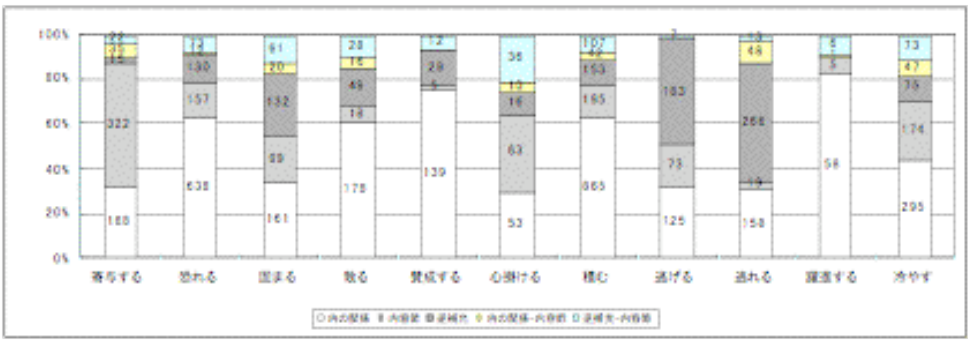

図 1　連体 (修飾) のタイプ分布

調査の結果，連体節のタイプとして図 1 の分布が得られた．全体の傾向としては，内の関係がもっとも多く，それに内容節と逆補充のタイプが続い

11) この抽出作業には格フレーム辞書の検索の web インターフェイス `http://reed.kuee.kyoto-u.ac.jp/cf-search/` は使っていない．ただし，格フレーム辞書の検索結果から原文を参照できるので，同じことを web インターフェイスを使って行なうことは (手間はかかるが) 可能である．

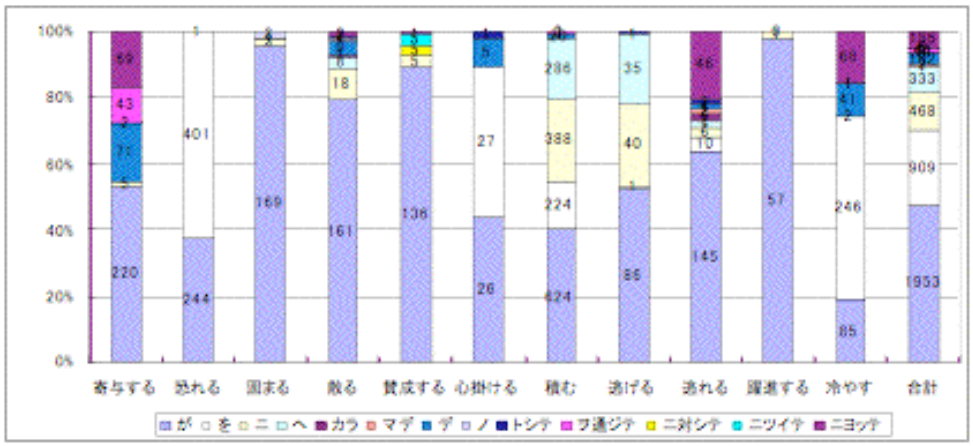

図 2　内の関係の格助詞別の分布

ている．その一方で非常に興味深いのは，語彙項目によって分布の内訳が異なっていることである．「寄与する」や「心がける」では内容節が主に用いられている．「逃げる」や「逃れる」では逆補充が主に用いられている．すべての語彙項目ごとに均質に分布しているわけではない (この分布の差は分散分析で有意差も出ているので，誤差ではない)．これは説明に値する事実の発見だと言える．

続いて内の関係における格関係の集計を行なった．格関係と動詞のクロス集計を行った結果を図 2 に示す．これから次の 4 点が示唆される:

(13)　a. 全体における使用頻度および生産性の面から，その傾向を捉えた場合，[ガ格 > ヲ格 > ニ格 > ヘ格 > デ格 > ニヨッテ格 > ヲ通ジテ格] という傾向が観察される．

b. (「積む」を除く) 他動詞「恐れる」「心がける」「冷やす」の場合，ガ格よりヲ格の方がより生産的である．

c.「散る」「積む」「逃げる」のように移動を表す動詞群においては，ガ格に次いで生産的なものとしてニ格がある．

d.「固まる」「賛成する」「躍進する」のような変化を表す動詞群に

おいては，9 割近い用例でガ格が使用されている．

(13b) では係り受け解析の段階でハ句に隠れたガ格を取りこぼしていることが原因になっている可能性がある[12)]．

4.4　まとめ

本研究の調査によって全体の傾向についての先行研究の指摘の妥当性が示唆される一方，語彙項目によって一般的な傾向から外れた分布が認められる可能性が示唆された．これはある集団の平均的な挙動とその部分の挙動は必ずしも同じにならないということ，統計解析では一般に「データの層別化の問題」として知られる問題の一種が生じている可能性が示唆される．これは文法レベルでの一般化とは別に語彙項目の特性を踏まえた，構文的なレベルでの一般化も必要なことを示唆する結果だと言える．

本研究の妥当性で問題となるのは二点である: 1) 11 個の述語の選択基準が明確ではなく，得られた結果の代表性が明らかではない．2) それに問題がないとしても人為的分類ミスが排除できていない．この種の問題が残っているものの，調査結果の妥当性以上に私たちが重要だと考えているのは，本研究が Web コーパスを基盤として構築された大規模格フレーム辞書が言語研究用の有益な研究資源になる可能性を提示している点である．

5　格フレームの精度の人手評価と改良のための提案

§4 から格フレーム辞書の言語研究への利用可能性が示せたと思う．ただ，現時点で格フレーム辞書がどれほど信頼できる資源なのか不明な点がある．黒田ほか [11] はこの問題意識から，格フレーム辞書の精度を人手で評価し

[12)] ただ，これが「積む」「固まる」「賛成する」「躍進する」の用法クラスターとの対比で説明として成立するためには，ハがガを隠しやすい述語 (か環境) とそうでない述語 (か環境) の区別があることを示す必要があり，そのための独立の調査が必要である．

た．この節ではその結果を簡単に報告し，言語学の観点から精度向上のための提案を行なう．これは言語学から NLP へのフィードバックである．そうするのは，現行のデータベースの精度が不十分であるという事実の根本的原因の一部が，言語学がそれらについて正しい理論を用意していないことが原因になっている可能性は高いと考えるからである．

5.1　評価の結果と個別の傾向の分析

5.1.1　格フレーム辞書が表現すべき情報の定義

格フレーム辞書は有用なデータベースだが，用法クラスターの一つ一つが表現している情報が何なのかは明確ではない．[8] の示唆に従うならば，「うまく行った」述語の用例クラスターとは，$[C_1,\ldots,C_n]$ を総合的に判断して，[8] の言う意味での**特定の状況**を表わすと判断できるクラスターである．

格フレーム辞書の用例クラスター C が理想的な状態にあるのは，C の格要素のどれにも，C が状況 S を表わすなら入っているべきではない要素の混じっていない状態のことである．どの述語を見ても格フレーム辞書の現状はこの理想からはほど遠い．

述語の用例クラスターは格要素集合ごとに適当な得点を与え，その合計点から次のように分類した: 極上品 (>150), 美品 (150∼125), 美良品 (125∼100), 良品 (100∼70), 並品 (70∼50), 難あり (70∼25), 評価不可 (25∼0).

16 個の述語の精度評価の結果を図 3 と図 4 に個数グラフと割合グラフにして示す (個数グラフは述語の使用頻度を表わすわけではないことに注意).

5.1.2　精度低下の主要因: 二格，デ格の曖昧性

一見して述語ごとの精度のばらつきが大きい．細かく見てゆくと次のようなことがわかった:

(14)　a. 述語の頻度のクラスター精度への影響は顕著ではない．
　　b. ヲ格が強い述語はうまく用例がクラスター化されているが，ヲ格

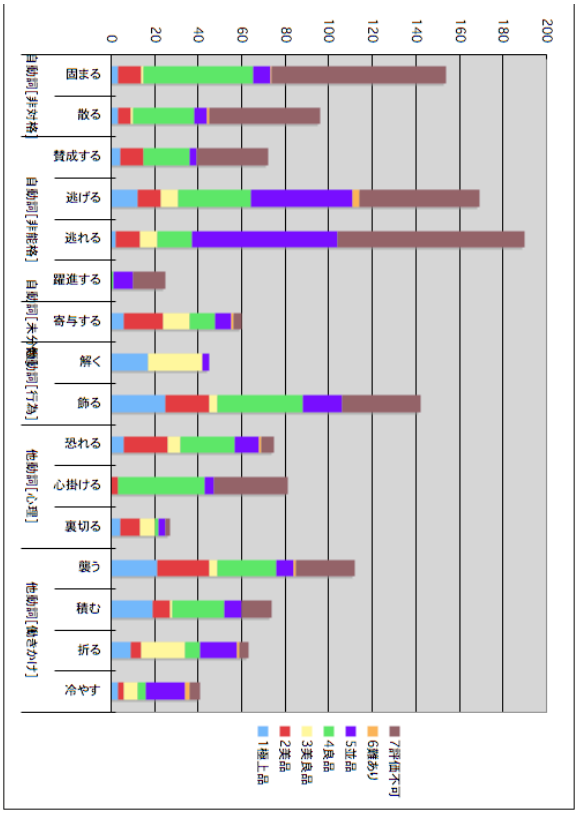

図３　横軸=述語ごとのクラスターの個数 (品質別)

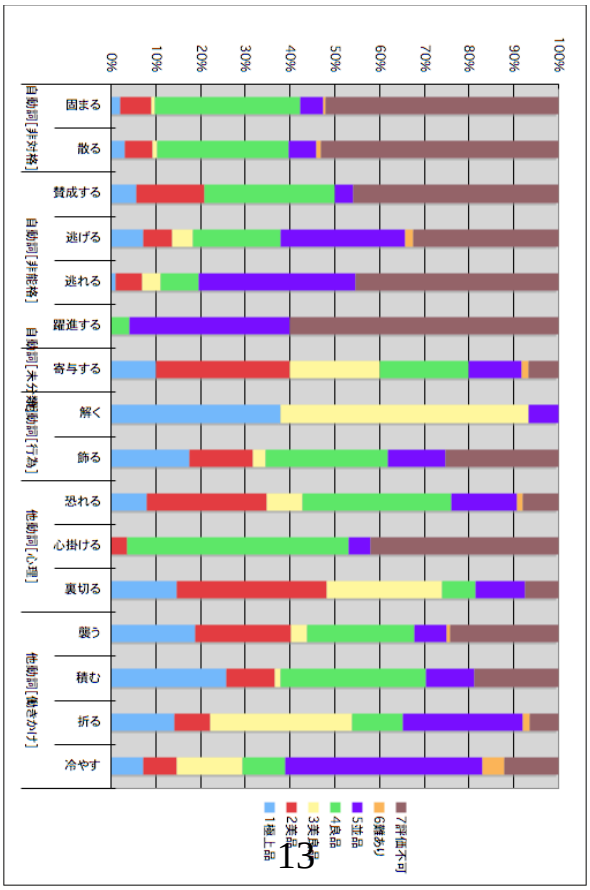

図４　横軸=述語ごとのクラスターの相対度数 (品質別)

を取らない述語のクラスター化精度が低い.

(14a) を考慮に入れて (14b) の結果を見ると，**クラスター化精度はおそらく (格) 助詞の意味の広がりに反比例している**と言える[13]．曖昧性が著しいのはニ格，デ格，ノ格であり，比較的に曖昧性が低いのがヲ格，ガ格である．ガ格は曖昧性は低いが，おそらくハへの吸収率が高いために獲得される実例数が少なく，用例クラスターの弁別に貢献していない．

ニ格が関係する述語の精度が低い最大の理由はニの用法の多様性にある (用法によってニ格が項に相当する要素だったり，しなかったりする)[14]．改善策は (17) で示す．ニ格要素と同じことがデ格要素の集合に関しても言えるが，デ格はそれでもまだニ格よりは曖昧性が少なく，また，ニ格と違って (準) 項をマークする率も低いので，デで終わっている文節の下位分類が効果的ではないことは，深刻な精度の低下には繋がっていないように思える．

5.2 言語学者による対策の提案

以上の問題は次の対応を取ることで改善できるのではないだろうか?

(15) 用例結合のための閾値は現在は一律に設定されているが，述語の意味クラスごとに別の閾値を最適化したり，統計量を反映させることには精度向上の効果があるのではないかと思う．図 3 の結果を見る限り，結合の閾値をヲ格をもつ述語ともたない述語に区別するだけでも，それなりの効果があるだろう．

(16) 用例収集・結合の活用形ごとの層別化: 大きな効果は期待できないだろうが，李ほか [13] の結果を考えると述語をタ/テイタ/ル/テイル形ご

[13] この可能性は黒田ほか [11] の研究に先立って藤井ほか [21] で示唆されている．

[14] ヲ格を項に取らず，ニ格を項に取る述語の中では「寄与する」が例外的に精度がよい．これはこの動詞と共起するニ格の曖昧性が例外的に低く，程度や様態を表わす副詞類が共起しないことだ理由だった．

とに採集し，活用形ごとに用例クラスター化した後に一つに統合することで，より良い結果が得られる可能性もある．一部の述語のクラスタリングはこれで精度が向上するかも知れない．

(17) ニ格要素と非ニ格要素の識別: ニで終わっている文節 (=名詞句あるいは後置詞句) からニ格を識別する必要については §5.1.2 ですでに述べた．これには次の方法に即効性のある効果が期待できると思う:[15)]

a.「X に」の要素 "X" について他の格 (e.g, X が, X を, X から, X の, ...) の要素になるかを判定し，

b. 他の格に現れる (か期待される格分布に従っていると判断できる) なら，ニ格要素と，

c. 他の格に現れない (か期待される格分布から明らかに偏っていると判断できる) なら，非ニ格要素とする．

これによって統計量を使ってニ格と非ニ格を効果的に区別できる可能性は高いと思われる (名詞句に期待される格分布の正確な姿は一般にはわかっていないので，これは独立に調べておく必要がある).

話を更に一般化すると，次の可能性に繋がる:

(18) 格分布クラスを使った名詞の下位分類の推定: 標準化された格列 A = [__ ガ, __ ガ 2, __ ヲ, __ ニ, __ デ, __ カラ , __ マデ, ...] がある時[16)]，$A(N) = [p_1, \ldots, p_m]$ ($p_i = \frac{n_i}{\sum_1^m n_i}$) によって名詞 N の格頻度分布を定義し，$A(N_1), \ldots, A(N_n)$ を同値類に分類 (例えばクラスタリング) して $N_1, \ldots, N_n$ の下位分類を得て，それで N_i のクラスを推定する．

15) この方法は加藤鉱三 (信州大学) の提案による．

16) ただし，自動詞のガ格と他動詞のガ格は区別する必要があるかも知れない．

5.3 現行の辞書構築の主な問題点

評価を通じて明らかになった現行の格フレーム辞書の (言語学者から見る限り) 問題だと思われる点は次の通りである:

(19) 共起情報の損失: 現行の格フレーム辞書では名詞 (句) 同士の共起情報が利用されていない．このため「x が y を襲った」からニ格名詞として y が採られた時のガ格名詞が何であったかが明示されない．

(20) 同表記異語の扱い: 「避ける」には「さける」と「よける」が混在していた．読みの推定と語義の推定は表裏一体なので，語義推定の手法の進歩が問題の根本的解決には不可欠だろう．

(21) より大きな単位での検索の必要性: 意味を単位とした場合，形態素よりも大きな単位の分割が必要である．この意味での過分割の発生頻度は高く，影響は無視できない．「者」のような接尾辞，「上」のような接尾辞化した形式名詞が —見かけの被覆率を向上させつつ— クラスター結合の精度の低下を招いている可能性は高い．これは係り受けの前に形態素の結合処理を入れることで対処できるだろう．

(22) 複合動詞の扱い: 複合動詞の格支配がうまく処理されていない．V_1 (e.g, 「襲う」「解く」) が $V_1 + V_2$ (e.g., 「襲い + かかる」「解き + 放つ」) という形で複合動詞をなす場合に，V_2 の格要素となっている形が V_1 の格要素として抽出されている．

(23) ハに吸収されている格の扱いの改善: 係助詞ハの要素は収集から外されている．これはハに化けて現れやすい格要素の事例が十分に収拾できないという結果を生んでいる．評価の結果を見ると，影響を受けやすいのはガ格である．ガ格は一般に省略されやすいと考えられているが，ハに吸収されているために取りこぼし率も高いと思われる．ガ格の被覆率を上げるにはハに隠された格の推定が必要かつ有効だろう．

6　終わりに代えて: 加工されたデータを使う理由

Web データを言語研究に使えるようにするには加工=前処理が必要である．その加工は長谷部 [22] の実装したマークアップ除去処理には限らない．表層情報よりも先に進み，意味情報にアクセスしようと思えば，形態素解析，構文解析の自動化の重要性が増す．格フレーム辞書は，人手で解析することは不可能なほど大規模なデータから抽出され，それなりの精度で一般化された述語/項関係の大規模データベースである．その被覆率の高さは圧倒的である．これをうまく利用することで意味研究のための前処理を省ける．

一方，NLP 研究者は言語学者からのフィードバックに期待している．係り受けの精度向上は時間が自然に解決するタイプのものではない．言語学者が格フレーム辞書を使い，その改良のために NLP にフィードバックをすることは，言語学の将来のための投資だとは考えられないだろうか?

参考文献

[1] B. A. Fox. The noun phrase accessibility hierarchy reinterpreted: Subject primacy or the absolutive hypothesis? *Language*, 63:856–870, 1987.

[2] E. Keenan. Variation in universal grammar. In R. Fasold & R. Shuy, editors, *Analyzing Variation in Language*, pp. 136–148. Georgetown University Press, Washington, D. C., 1975.

[3] E. Keenan & B. Comrie. Noun phrase accessibility & Universal Grammar. *Linguistic Inquiry*, 8:63–99, 1977.

[4] 久野 ススム. **日本文法研究**. 大修館書店, 東京, 1973.

[5] 伊藤 雅光. 計量言語学とコーパス言語学. **計量国語学**, 25(2):89–97, 2005.

[6] 前川 喜久雄. 大規模均衡コーパスが開く可能性. In **日本語言語学会第** *134* **回大会予稿集** *(***公開シンポジウム***)*, pp. 24–29. 日本言語学会, 2007.

[7] 奥津 敬一郎. **生成日本文法論**. 大修館書店, 東京, 1974.

[8] 中本 敬子 & 黒田 航. 「逃れる」の階層的意味フレーム分析とその意義: 「言語学・心理学からの理論的，実証的裏づけ」のある言語資源開発の可能性. In **言語**

処理学会第 *12* 回大会発表論文集, pp. 592–595, 2006. 発表 P4-1.

[9] 三原 健一. **日本語の統語構造**. 松柏社, 東京, 1994.

[10] 竹内 孔一. グラフ構造に基づく同時クラスタリングを利用した動詞の属性クラスの抽出. **情報処理学会研究報告**, 2007-NL-182(6), 2007.

[11] 黒田 航, 李 在鎬, 渋谷 良方, 河原 大輔, & 井佐原 均. 自動獲得された大規模格フレーム辞書の精度向上を見込んだ人手評価. In **言語処理学会第 *13* 回年次大会発表論文集**, pp. 1054–1057. 2007.

[12] 黒田 航 & 飯田 龍. 文中の複数の語の (共) 項構造の同時的, 並列的表現法: Pattern Matching Analysis (Simplified) の観点からの「係り受け」概念の拡張. **信学技法**, 106(191):1–5, 2006.

[13] 李 在鎬, 鈴木 幸平, 永田 由香, 黒田 航, & 井佐原 均. 動詞「流れる」の語形と意味の問題をめぐって. **計量国語学**, 26(2):64–74, 2007.

[14] 寺村 秀夫. **日本語の文法 *(下)***. 国立国語研究所, 東京, 1981.

[15] 寺村 秀夫. **寺村秀夫論文集 *I:* 日本語文法編**. くろしお出版, 東京, 1993.

[16] 加藤 重広. 日本語関係節の成立要件 (1): 先行研究の整理とその問題点. **富山大学人文学部紀要**, 30:65–111, 1999.

[17] 加藤 重広. 日本語関係節の成立要件 (2): 文法論的要因と語用論的要因. **富山大学人文学部紀要**, 31:71–156, 2000.

[18] 丸元 聡子 & 乾 裕子. 連体修飾を受ける体言の格構造の復元: コーパスに基づく「内の関係」の分析. In **言語処理学会第 *6* 回年次大会発表論文集**, pp. 16–19. 言語処理学会, 2000.

[19] 河原 大輔 & 黒橋 禎夫. 格フレーム辞書の漸次的自動構築. **自然言語処理**, 12(2):109–131, 2005.

[20] 河原 大輔 & 黒橋 禎夫. 高性能計算環境を用いた Web からの大規模格フレーム構築. **情報処理学会研究報告**, 2006-NL-171:67–73, 2006.

[21] 藤井 敦, 秋山 典丈, 徳永 健伸, & 田中 穂積. 動詞の多義性解消における格の弁別能力と集中度の有効性について. In **言語処理学会第一回年次大会発表論文集**, pp. 117–120. 1995.

[22] 長谷部 陽一郎. Wikipedia 日本語版を利用した言語研究の手法. **言語文化**, 9(2):374–403, 2006.

[23] 井上 和子. **変形文法と日本語 *(上)***. 大修館書店, 東京, 1976.

[24] 李在 鎬, 黒田 航, 渋谷 良方, 河原 大輔, & 井佐原 均. コーパス分析に基づく連体表現の使用調査. In **第 *134* 回日本言語学会大会予稿集**, pp. 422–427, 2007.