日本語ツリーバンクからの動詞格フレームの抽出

吉本 啓　　アラステア・バトラー　　プラシャント・パルデシ
東北大学　　弘前大学　　国立国語研究所
kei@compling.jp　　ajb129@hirosaki-u.ac.jp　　prashant@ninjal.ac.jp

1 はじめに

外国人に対する日本語教育が日本社会において重要性を増しているのに対して，中級以上の非母語学習者用辞書が存在しないことが日本語の学習と普及にとって大きな障害となっている．発表者たちは日本語としては初めての本格的な統語・意味解析情報を付加したコーパス（ツリーバンク）である NINJAL Parsed Corpus of Modern Japanese (NPCMJ) の構築を行ってきたが，学習者用辞書の根幹をなすコロケーションの情報をこれから抽出し，日本語教育体制の不備を補うのに役立てることができる．本発表では，12 個の基本動詞を取り上げ，それらの格フレームのパターンを NPCMJ の特性を生かして抽出，分析し，現実のテクストにどのような特徴があらわれているかを考察する．続いて，日本語の基本動詞全体の格フレーム分析のためにどのような方法が有効であるかについて検討を行う．

2 NPCMJ と基本動詞ハンドブック

NPCMJ は，国立国語研究所共同研究プロジェクトで開発中の，日本語として初めての本格的な文統語・意味解析情報をタグ付けしたコーパスである (Butler and Horn 2017, 吉本・パルデシ 2020)．従来，日本語のコーパスは文を文節へと区切った上で，形態素情報を中心とするアノテーションおよびせいぜいそれに文節間の係り受け情報を加えたものしか存在しなかった．そのため，文法研究者が興味を持つ用例を検索したり，文法情報を抽出するのには不十分であった．また，外国人に対する日本語教育に役立てるために日本語使用の実態を知りたいと思っても，得られる知識は限定的であった．これに対し，NPCMJ では文を句構造に分析した上で，さらに自動意味解析によって得られる述語論理式を利用して，関係節による修飾等の非境界依存 (unbounded dependency) の関係にある語句間の関連づけも行える．これにより，例えば関係節中の動詞と修飾される主名詞との間の格フレーム関係までもが抽出可能となる．

従来の言語学において，統語論や意味論上の精密な理論的研究は少量のデータや作例にもとづいて行われ，これに対して，コーパスを利用した研究は語彙や形態論に限定される傾向が強かった．NPCMJ の登場により，一定量の実例に裏づけられた理論研究，言い換えれば言語研究における質と量の統合が可能になる．これにより，日本語教育等の応用面においても新しい展開が期待される．

格名詞句，すなわち必須文法役割 (主語，直接目的語および間接目的語) を果たす名詞句/助詞句は，NPCMJ においては，格助詞による同定以外にも，文法役割を当該の句に直接タグ付けすることによって同定する．日本語において必須文法役割を担う句には必ずしも格助詞が明示的に付加されるわけではなく，助詞が脱落した裸の名詞句としてあらわれることがある．また係助詞 (とりたて助詞) が付加された場合，格助詞「が」と「を」は出現せず，「に」については任意である．NPCMJ では，このような場合を含む全てのデータに対して文法役割がタグ付けされている．また，日本語のテクスト中には主語や目的語の省略が頻出するが，NPCMJ ではそれらをゼロ代名詞としてタグ付けするので，この場合も格フレームは正確に把握されることになる．さらに，上記のように関係節を伴う文等の非境界依存において，また主節中の主語や目的語に従属節中の非明示主語が一致するコントロール構文においても，当該の文の自動意味解析が行われることにより，必須文法役割を担う句と述語との関係を突き止めることができる．このように，NPCMJ は，データにあらわれるすべての述語の格フレームを提供するという点で画期的なコーパスである．

NPCMJ は，2021 年 1 月の時点で 40,831 文 (560,098 語) の現代日本語テクストのアノテーションが終了し，ウェブ上で公開を行っている(注 1)．2022 年 3 月のプロジェクト終了までに約 6 万文のアノテーションを完成させる予定である．

『基本動詞ハンドブック』は，外国人による日本語の学習やその教育に役立てることを目的として，使用頻度の高い日本語基本動詞 125 語の重要な統語的，意味的，語彙的特徴を集積したインターネット上の知識ベースである(注2)．基本動詞には多様な異なる用法を持つ，いわゆる「多義語」と呼ばれるものが多いが，『基本動詞ハンドブック』はそれらを，項の数や格助詞による表示等の統語論的振舞いや意義特徴を手掛かりとして抽出した格フレームごとに分類していることに特色がある．『基本動詞ハンドブック』は格フレームの他にも，受動形の有無やアスペクト等の文法的特徴，意味拡張，自他の対をなす場合にはパートナーである自動詞/他動詞が何であるか，また類義語との対比等，基本動詞について体系的に学ぶための便宜を提供している．

『基本動詞ハンドブック』は,『現代日本語書き言葉均衡コーパス (BCCWJ) 』(Maekawa et al. 2014)を利用してデータ収集を行っている．しかし，そのため先に述べた理由によって，各基本動詞の有する統語論的特徴について，精度の高い情報を参照するには至っていない．特に，上記のように『基本動詞ハンドブック』は統語・意味的なアプローチで格フレームを規定することにより多義の問題に対処しようとしているが，このことは，全ての述語に格フレームを網羅的にタグ付けした NPCMJ とは相性が良い．

『基本動詞ハンドブック』の記述を NPCMJ を用いて検証することにより，前者がどれだけ現実の言語使用の実態をカバーしているかが解明でき，そのことはより精緻な情報を日本語の学習者や教師に提供することにつながると考えられる．

3 NPCMJ の用例の分析

3.1 概要

NPCMJ を利用した『基本動詞ハンドブック』所収の動詞の本格的調査の前提として，一部の動詞を選んでパイロット・スタディを行った．その対象としたのは，以下の 12 個の動詞である．NPCMJ（2021 年 1 月現在） の検索によって得られた用例文の数を（ ）の中に示す．

あたる（71），あてる（25），いう（1,562），おさまる（6），おちる（67），きこえる（66），さわる(5)，でる（320），ぶつかる（14），みえる（280），もつ（414），わかる（395）

まず目につくのは，同じく基本動詞と言っても，その用例数に大きな開きが存在することである．用例が数百から千を超える動詞がある一方で,「おさまる」は 6 例，また「さわる」については 5 例しかない．後者の 2 つから何か一般的な結論を抽き出すことは無理である．この傾向は，NPCMJ プロジェクトにおいて 6 万文のアノテーションが完成してもあまり変わらないと思われる．これは，NPCMJ がツリーバンクとして，比較的少量のテクストに対し集約的にアノテーションを行う方針を取っているためである．しかしながら，かなり大規模なコーパスにおいてすら，十分な検索数の得られない単語の存在することが知られている．

また,『基本動詞ハンドブック』では多義性の問題に対処するために各動詞の語義を重層的に分類しているが，それらの実例がすべて NPCMJ により提供されるわけではない．例えば,「出る」については 6 段階をなして分類が行われているが，そのうち 3 層目である「5. 身体部位の突出」や「41. 出土・発見」については，検索の結果用例は得られなかった．また，検索して得られた語義の用例が 1〜3 個程度の少数にとどまるものはさらに多い．

3.2 視覚動詞の対象の表示

本節では，視覚的認知活動を表す動詞「見える」を取り上げて，NPCMJ からどのような格フレーム情報が抽出できるかについて考察する．

動詞「見える」は視覚に関わる出来事を表すことから，それが成り立つ場合には「見る」行為の主体と「見られる」対象とが存在するはずである．このうち，主体についてはほとんどの場合明示的に表現されない．これについては後で述べる．対象の方は明示されることが多いが，格助詞としては主として「が」が用いられる．これが主題化されると，助詞は「は」となる．通常の文において「が」による主語表示の文と「は」表示の文とでは頻度にきわめて大きな差はない. これに対し, NPCMJ の検索の結果,「見える」を伴う文の主語表示は，特異なものであることが分った．これは『基本動詞ハンドブック』における「見える」の多義分類のうち，用例のもっとも多い「1. 視覚による認識 (非意図的) 」に限定した調査である. NPCMJ の全データ中で, 文の主語で「が」により表示されているものは 32,530 例，主語で「は」により表示されているものは 21,741 例であり, 両者に圧倒的差は無い. これに対し,「見える」

の用例について見ると,「が」表示の主語が 73 例で,「は」表示の 13 例に比べて突出している.

「が」表示の主語が多くなる理由は,下の例に典型的に見られるように,語られている状況の中で新しい出来事が生じたことを表すのに使われることが多いからである.

(1) この装置で,尖端放電の研究をするつもりだったところが,膨張させてみると,白い線が 見えた.

この文では,「白い線が見えた」全体が新情報であり,伝統的な日本語学で言う「現象文」に相当する.

格フレームは格助詞を中心として語られることが多い.しかし,ここでの「が」の働きは単なる格表示にとどまらない.それは,(「は」の不在を表示することによって) 情報構造に関わる積極的な役割を果たしている.格フレームの用法について正確な知識を伝えるためには,文脈的な要因も考慮に入れることが重要であることをこの例は示している.

なお,外国人に対する日本語教育の教科書では,以下のように視覚的認識行為の主体を「に」格により明示した例文が使用されることがある.

(2) 私には細かい字が見えません.

しかし, NPCMJ の検索結果によると,このような構文の出現頻度はきわめて低い.「見える」の全用例 278 例のうち,主体が「に」より表示されたものは 4 つあるが,そのうち 2 例は欧文からの翻訳で,オリジナルの日本語文は 2 例にすぎない.このことは,日本語教育では対象のみが明示された構文を中心に教えるべきことを示唆している.

3.3 理解動詞の格フレーム

理解を表す動詞「分かる」の格フレーム表現は,さらに複雑なものである.以下では, NPCMJ 中の動詞「分かる」の全用例 387 例中で最大の 280 例を占める「1. 不明瞭な事柄の明確化」に的を絞って論じる.

動詞「分かる」の目的語 (理解の内容) を表す名詞句の内部構成—普通名詞をヘッドとするか,それとも節か,後者の場合はどのような助詞や形式名詞により導かれるか—および格助詞・係助詞による表示の内訳を表 1 に示す.

表 1: 「分かる」の目的語

節+こと		**40**
	節+こと+が	35
	節+こと+その他	5
節+か		**84**
	節+か+Φ	71
	節+か+その他	13
節+の		**7**
節+と		**14**
普通名詞・代名詞		**72**
合計		**217**

形式名詞「こと」が導く目的語節は 40 例,「か」が導く節は 84 例ある.助詞「の」および「と」が導く節も含めると 145 例となり,全体の約 66%を占める.「分かる」の目的語は大半が節によって表されていることが分かる.また,その格表示に目を向けると,内部構成の種類にかかわらず,格助詞としては「が」が圧倒的に多い.「を」の使用は 1 例あるが,この比率の差は母語話者としての直観に一致している.

理解という行為は理解を行う主体と理解の対象である内容とを伴い,それぞれ主語と目的語によって表現される.ところが,両者の出現頻度には著しい差がある.表 2 に示すように,ここで扱う 280 例のうち,主語が出現する用例は 63 例,目的語が出現するのは 217 例である.さらに,主語と目的語の両方を伴う文についても興味深い事柄が観察される.一般に,日本語としては有標な語順であるはずの,目的語が主語に先行する文 (表 2 では「目的語 < 主語」で表す) が,主語が先行する文 (「主語 < 目的語」) と比べて,出現数に大差が無い (16 例対 29 例).また,主語の助詞による表示は,「目的

表 2: 「分かる」の主語と目的語

目的語 < 主語		**16**
	目的語 < 主語-に	14
	目的語 < 主語-その他	2
主語 < 目的語		**29**
主語のみ		**18**
目的語のみ		**172**
主・目両方無し		**45**
合計		**280**

語 < 主語」の語順の 16 例中の 14 例を「に」が占めている (後に係助詞が後続するものを含む).

以上をまとめると,「分かる」の格フレーム表現の特徴として，以下が挙げられる.

i. 目的語のみがあらわれる文が多く，主語の出現は限定的である.
ii. 目的語は節の形式を持つものが多い.
iii. 主語と目的語の両方が出現する文においては,「主語 < 目的語」と「目的語 < 主語」の語順の頻度に大差はなく，主語はほとんどが「に」により表示される.

理解を表す「分かる」は主観表現の一種であり，一人称や感情移入の対象となる登場人物を主語とすることが多く，その場合にしばしば省略されるのは他の主観表現と同様と考えれらる．目的語と共起する主語に「が」でなく「に」が付加される傾向が強いのは，目的語との混同を避けるためであろう．しかし，なぜ「目的語 < 主語」の有標な語順となるのかは不明である．目的語が節で長くなるため，記憶の負担を避ける処理上の理由を検討すべきかも知れない.

4 検索結果の階層的表示

以上で行った調査から，現実のテクストにおいて頻出する格フレーム表現を把握するためには，格助詞による表示のみならず,必須格を担う句の省略や,係助詞が表示する文脈情報，さらには語順や名詞句の内部構成まで考慮に入れなければならないことが分かった．そこで，このように複雑な要因を考慮しながら頻度の高い格フレーム表現の抽出作業を容易にするために，ツリーバンクの検索結果を要因ごとに階層的に自動的にグループ化するシステムを提案する．暫定的な案として，システムによる用例の分類は 3 つの軸に沿って行われ，各々の軸において階層が深まるにつれて文法上の制約がより強くなり，担当する用例の数は少なくなる.

I 格表示の軸

i. 動詞 → ii. 文法役割の格フレーム → iii. 付加句 (adjuncts) の格フレーム → iv. ゼロ代名詞の有無 → v. 格助詞による表示 → vi. 係助詞による表示

II 語順の軸

i. 語順未指定 → ii. それぞれの語順

III 名詞句内部構成の軸

i. 名詞句 → ii. 名詞をヘッドとする句，または節 → iii.(節の場合) こと - 節，の - 節，または か - 節

I〜III の軸に共通して, i は，それぞれの動詞が出現する全ての用例に対応する．動詞「会う」を例に取ると,「I-ii. 文法役割の格フレーム」のレベルでは,<主語> または <主語+目的語> の格フレームが選択され，それぞれ 1 項述語および 2 項述語としての用例に対応する．後者の格フレームは「v. 格助詞による表示」で <主語+目的語-と> および<主語+目的語-に> に分かれる．これら第 I 軸の各階層は，語順に関する第 II 軸および名詞句の内部構成に関する第 III 軸と自由に組み合わせることができる．これにより，3 つの軸に分類された要因のどの組み合わせについてもその頻度を即座に知ることができ，また他の組み合わせとの比較も容易である.その結果,頻度の高い格フレームの抽出が簡単に行える.

5 おわりに

本稿では，日本語教育への応用を目的として，日本語ツリーバンクからの格フレーム情報抽出のパイロット・スタディを行い，抽出を効率的に行うための方法について考察した．今後，格フレーム以外のコロケーションについても検討を広げる予定である.

注

1. http://npcmj.ninjal.ac.jp/
2. https://verbhandbook.ninjal.ac.jp/

謝辞

本研究は日本学術振興会科研費基盤 (B) 15H03210，基盤 (C) 16K02654， および国立国語研究所共同研究プロジェクト「統語・意味解析コーパスの開発と言語研究」の助成を受けた。

参考文献

Butler, A. and S. W. Horn. (2017) Annotating syntax and lexical semantics with(out) indexing. *Proceedings of Logic and Engineering of Natural Language Semantics*.

Maekawa, K., et al. (2014) Balanced corpus of contemporary written Japanese, *Language*

Resources and Evaluation 48(2).
吉本啓・プラシャント-パルデシ（2020）「統語・意味解析情報付き日本語コーパスの構築」*KLS Selected Papers 2: Selected Papers from the 44th Meeting of The Kansai Linguistic Society*, pp. 196-211.