二格深層格の定量的分析

松田　真希子*　　森　篤嗣**　　川村　よし子***
庵 功雄****　　山本　和英*****　　山口　昌也******

*金沢大学, **帝塚山大学, ***東京国際大学,
****一橋大学 *****長岡技術科学大学 ******国立国語研究所
mts@staff.kanazawa-u.ac.jp, moria24@gmail.com, kawamura@tiu.ac.jp,
isaoiori@courante.plala.or.jp, yamamoto@jnlp.org, masaya@ninjal.ac.jp

1　はじめに

日本語では名詞と述語の関係のあり方は格関係によって規定される．格の意味的な規定を指す深層格の内訳は多くの研究者によって提案されているが，共通見解は得られていない．

今回深層格の自動推定技術の開発のため最も深層格の類型の多い助詞である二格を対象に先行研究を整理した上で妥当性の高い深層格リストを提案し，3種類のコーパス（合計30,000句）に人手でアノテーションを行った．その後，ナイーブベイズ法を用いて3種のコーパス別に深層格推定の精度評価を行った[8].

本論文ではその作業によって得られた二格深層格コーパスに対して言語学的見地から定量的分析を行った結果について報告する.

2　先行研究

言語研究における二格深層格リストの提案に関するものは数多く，代表的なものでは[2][3][4][5][6][7]等がある. そのうち[4]が最も多くの語例に基づいた詳細な分類がなされており，二格の深層格推定研究にも応用されている[9]．しかし，全ての深層格リストの提案は人手で収集された限られた小規模コーパスの分類から導かれたもので，定量的分析によって導かれたものは管見の限りない.

3　研究手法

本研究では先行研究を基にタグリストを再検討し，コーパスにアノテーションし，定量的に分析を行った．深層格リストの作成にあたっては，まずEDRの関係子及び[2]-[7]等の文献を参照し，言語学の専門家メンバーが最終的に決定した．提案深層格リストを表1に示す. 設計にあたっては，(1)他の助詞との置き換え可否や二格に前接・後接する語の品詞等，客観的基準によって分類が可能なものを優先的に分類，(2)意味上の隔たりが小さいものは一つにまとめる,という方針で行った.

次にコーパスを選定し，人手でアノテーションを行った. コーパスは(1) Web日本語Nグラム（以下Web），(2) 京都大学テキストコーパス（以下京大），(3)BCCWJの三種類のコーパスを選定した. (1)の選定理由はインターネット上にある膨大なコーパスに基づく情報で汎用性が高いため，(2)は京都大学テキストコーパスのアノテーション情報を今後深層格推定に利用するため，(3)は学術的に公開された日本語の均衡コーパスとして最大のものであるためである．最終的な深層格リストと先行研究で提示された深層格の対応を表1に示す.

表1　提案深層格リスト

	例文	先行研究
時間	8時に起きる	[2][3][4][5][6][7]
場所	公園に現れる	[4][5]
	ハワイにいる	[2][3][4][5][6][7]
	東京に行く/着く	[2][3][4][5][6][7]

結果	息子を医者にする	[5][7]
	どろどろに溶ける	[5][6][7]
	コの字型にならべる	[5]
	医者になる	[3][5][6]
対象	実験に成功する	[2][6]
	A は B にまさる	[2][5]
	対応に怒る	[2][4][5][6]
	説明に困る,彼にほれる	[2][4][5][6]
	駅に近い,父に似る	[2][3][6]
	父に手紙をあげる	[2][3][4][5][7]
	服にくっつく	[2][3][4][5][6]
	お母さんに甘える	[2][3][4]
動作主	太郎に殴られる 私にできること	[2][5][6]
目的	映画を見に行く	[3][4][5][6]
役割	貿易を外交の手段に用いる	[6]
頻度*	三年に一回	[7]
副詞化*	元気に歩く，お気軽に申し付け下さい	[4]
複合辞*	環境について語る 法律に基づく表示	[3][6]
起点*	太郎にもらう	[7]
その他	真犯人にちがいない,口に出す,役にたつ	[4][5]

*は論文[8]の後に設定した深層格

4 結果と考察

4.1 コーパスによる深層格の差について

深層格を付与したリストをフリーのテキストマイニングツール KH-Coder で分析した結果を表 2 に示す. データ数の合計は, 約 30,000 句, 560,000 語である. 分析の結果, 出現頻度の低い「起点」を除き, 全ての深層格で有意な差となった. 特に差が顕著となったのは,「場所」(BCCWJ＞京大>Web),「複合辞」(京大＞BCCWJ＞Web),「目的」(Web>BCCWJ>京大) であった. 名詞や動詞といった自立語の使用傾向がコーパスによって異なるということは一般的な共通認識となっているが, 本研究で深層格の使用傾向もコーパスによって異なることが明らかになった.

表 2： 3 種のコーパスに対する深層格付与の結果

	Web	京大	BCCWJ	χ2 値
時間	842	1023	673	84.5**
場所	111	307	578	326.7**
結果	1235	1182	1090	16.5**
対象	4136	3646	4293	91.0**
動作主	64	105	207	82.7**
目的	533	168	329	211.2**
副詞	1183	1116	740	141.0**
頻度	8	47	10	44.5**
役割	1249	1101	1145	16.1**
起点	16	12	12	0.95
複合辞	434	1052	867	265.4**
その他	1249	1101	1145	16.1**
句数	9827	10001	10154	

4.2 深層格の出現傾向の異なりについて

3 コーパスの深層格の出現頻度を図 1 に示す. コーパス全体, またいずれのコーパスでも「対象」が最も多く 4 割程度となった. 次に「結果」(12%),「副詞化」(10%)が続いた.

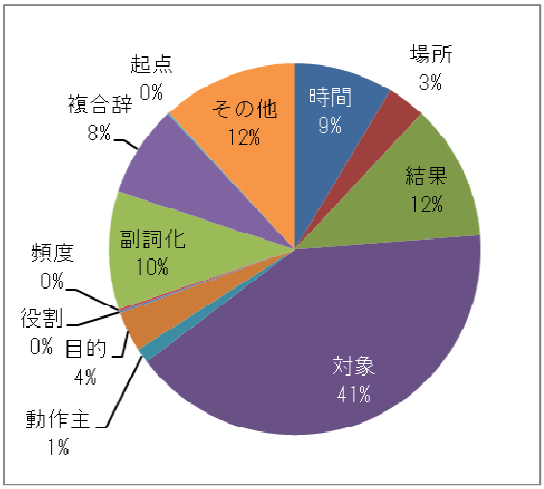

図 1 深層格の出現割合（全体）

すべての先行研究で挙げられている「時間」「場所」については「時間」(9%),「場所」(3%) と, それほど大きな割合は占めていなかった. 一方, 一

部の先行研究（[5] [6]）しか言及のない「その他」は上記二つの深層格以上の比率（12%）を占めた．

次に，深層格間の関係性を分析するため，二格との共起語と深層格との関係について主成分分析を行った．分析にあたっては，二格前後の単語数が最大 10 語に上る BCCWJ は除き，単語数が最大 3 語ずつの Web と京大で行った．総句数は 19,603 句，総抽出語数は 200,552 語である．そして，二格の前後に出現する単語の品詞のうち名詞と動詞だけを入力データとして与え，語の最小出現数 80 以上で分析を行った．結果を図 2 に示す．第二成分までの累積寄与率は 62%となった．図中の語の配置より，縦軸は「なる」に関わる語の軸，横軸は新聞記事に関わる語の軸によって構築されたと解釈した．この二つの軸に対し「結果」「その他」「複合辞」は他の深層格と比べ共起する語に偏りがあり，異なりも大きいことが明らかになった．

実際に例文検索したところ，「結果」の深層格が付与された 2418 句中 1330 句に「なる」が含まれていた．「複合辞」も「によって」(431 句），「について」(355 句）などが大半であった．

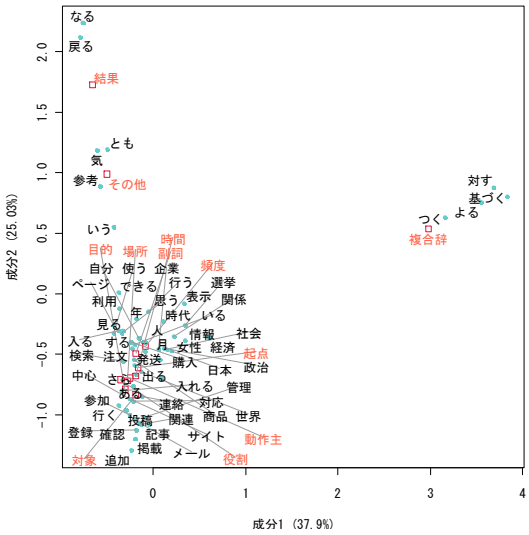

図２　全深層格の主成分分析結果

次に他と隔たりの大きかった三つの深層格を除いて同様の条件で主成分分析を行った結果，「副詞」「目的」が他との大きな隔たりを示したため除外した．残った 7 つの深層格に対し名詞と動詞（サ変名詞含む）で再度主成分分析を行った．結果を図 3 （名詞／語の最小出現数 40 以上，累積寄与率 88%），図 4 （動詞／語の最小出現数 60 以上，累積寄与率は 93%）に示す．名詞の場合，縦軸は「事前」「時代」「政府」「記事」と新聞記事に関わる語によって軸が構成され，横軸は「ログイン」「リスト」等 Web に関わる語によって軸が構成された．

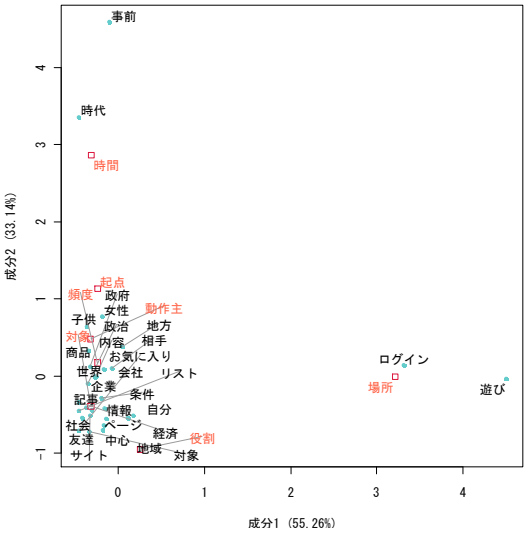

図３深層格と名詞との主成分分析結果

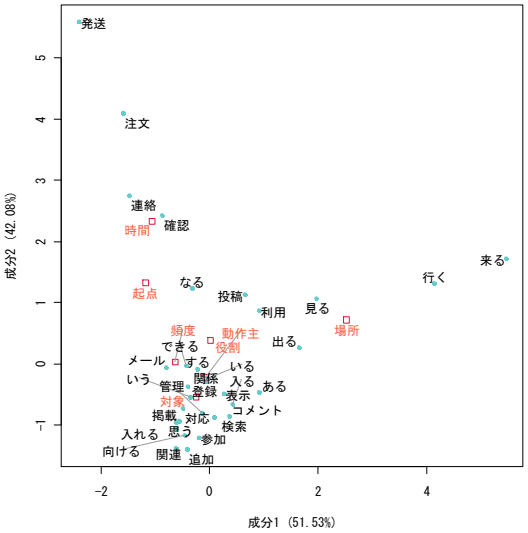

図４　深層格と動詞の主成分分析結果

「その他」の中の高頻度語句を調べたところ，最も多かったのは慣用表現の「気になる」（123）であったが，続いて高頻度だったのは「ご覧にな

る」(49)「ご利用になる」(40) といった敬語表現であった．その他「役に立つ」(37)「気分になる」(16)などの慣用表現が続いた．

図４については縦軸に「発送」「注文」「確認」といったWeb関連の語が並び，横軸に「行く」「来る」「出る」「する」といった基本動詞が並んだ．「場所」「時間」は名詞の場合同様，中心から離れた位置に出現した．このことからこの二つの深層格は前接名詞にも後接動詞にも特徴があると言える．そうしたことが，先行研究においても定性的に深層格が切り出された理由ではないだろうか．

5　まとめと今後の課題

本研究では３種のコーパスに対して二格深層格情報を付与したデータに基づき定量分析を行った．

その結果以下のことが明らかになった．

(1)二格深層格の出現比率はコーパスによって有意な差がある．特に「場所」，「複合辞」，「目的」における差が顕著である．

(2)定性的に分類された深層格を定量的に見ると頻度において差が顕著であった．特に「対象」(間接目的語) の頻度が高い．一方，「場所」や「時間」の頻度は全体の割合から見るとさほど高くない．

(3)二格と共起する名詞と動詞と深層格との関係について主成分分析を行った結果，「結果」「複合辞」「その他」＞「副詞」「目的」＞「時間」「場所」の順に他の深層格との隔たりが確認された．これに対して，「対象」「役割」「動作主」「頻度」は共起語においては特性が薄いため，深層格推定にも困難が予想される．

定性的には分類が可能な「動作主」「役割」「頻度」「起点」「対象」といった深層格については，共起語彙上では大きな差が見出しにくいことが明らかになった．こうした傾向の表す意味についてはさらにアノテーションの適切さや例を詳細に検討していく必要がある．

また，今回はコーパス設計段階でBCCWJの二格からの単語数を他のコーパスと揃えて設定しなかったため，主成分分析では除外せざるを得なかった．併せて今後の課題にしたい．

謝辞

本研究は科学研究費補助金基盤研究（B）[課題番号23320105]の助成を受けて行われた．

利用した言語資源およびツール

[1]国立国語研究所現代日本語書き言葉均衡コーパス（BCCWJ）．国立国語研究所, 2011. http://www.ninjal.ac.jp/corpus center/bccwj.

[2]工藤拓，賀沢秀人．Web日本語Nグラム第1版，言語資源協会, 2007. http://www.gsk.or.jp/catalog/gsk2007-c/.

[3]黒橋禎夫，河原大輔．京都大学テキストコーパス・プロジェクト．言語処理学会 第 3 回年次大会, pp.115-118,1997.

参考文献

[1] 松田真希子，森篤嗣，川村よし子，庵功雄，山口昌也，山本和英「日本語深層格の自動抽出のためのコーパス開発」『言語処理学会第18回年次大会発表論文集』205-208, 2012

[2] 城田俊『日本語形態論』ひつじ書房, 2002

[3] 鈴木重幸『日本語文法・形態論』むぎ書房, 1978

[4] 奥田靖雄「二格の名詞と動詞のくみあわせ」言語学研究会編『日本語文法・連語論（資料編）』281-323,むぎ書房, 1983

[5] 石綿敏雄『現代言語理論と格』ひつじ書房, 1999

[6] 高橋太郎『日本語の文法』ひつじ書房, 2005

[7] 庵 功雄, 中西 久実子, 山田 敏弘, 高梨 信乃『初級を教える人のための日本語文法ハンドブック』スリーエーネットワーク,2000

[8] 竹野峻輔，松田真希子，梶原智之，山本和英「機械学習を用いたニ格深層格の自動付与の検討」『言語処理学会第 20 回年次大会発表論文集』（印刷中）,2014

[9] 田辺利文，吉村賢治，首藤公昭．格格助詞「に」の深層格推定 -格助詞の意味再考-．情報処理学会研究報告, No.113,　65-72, 2009.