科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１０　日現在

機関番号：　１３３０１
研究種目：　基盤研究(B)
研究期間：　2011～2013
課題番号：　２３３２０１０５

研究課題名（和文）文脈依存の意味情報を判別する機能表現抽出ＷＥＢシステムの開発と運用実験

研究課題名（英文）Research for the Development of an Web Grammatical-Form Extraction System

研究代表者
松田　真希子（Matsuda, Makiko）

金沢大学・国際機構・准教授

研究者番号：１０３６１９３２

交付決定額（研究期間全体）：（直接経費）　　11,000,000円、（間接経費）　　3,300,000円

研究成果の概要（和文）：本研究課題では、（１）５つの表層格に対する深層格タグ付きコーパスの開発（２）機能表現等を自動抽出する機能表現WEB辞書の開発、（３）複合動詞の多義性を解消するための知識源の開発、（４）Naive Bayes分類器による深層格自動推定技術の開発を行った。これらの成果により文脈依存の意味情報を自動推定する技術の進展が期待される。

研究成果の概要（英文）：In this project we developed a deep case-tagged corpora of 5 Japanese surface cases, a web dictionary which enables learners to extract grammatical words and phrases, a knowledge resource for word Sense Disambiguation of compound verbs and an automatic deep case estimation with Naive Bayse Classifier. It is expected that a series of results will develop into the automatic estimation technology for the meaning of context-sensitive grammatical words.

研究分野：　人文学

科研費の分科・細目：　言語学・日本語教育

キーワード：　機能表現　自動推定　深層格　定量分析　自動抽出　コーパス

１．研究開始当初の背景

過去の研究では機能表現を文字列等の表層情報から抽出することは可能だが、用法・意味等の機能表現が有する深層情報の抽出や推定は、文脈や出現条件に依存するため、十分な成果が出ていない。また、難易度情報を伴った抽出機能や、複数の文節にまたがった連続しない機能表現の抽出機能もまだ開発されていない。

２．研究の目的

そこで本課題では言語処理、日本語学、日本語教育の研究者が連携し、日本語の深層格・難易度等の深層情報を表示できる機能表現辞書やタグ付コーパスを開発するとともに、より精度の高い自動検出手法を考案し、高精度の機能表現抽出システムの提案を目指す。

３．研究の方法

主に以下の４つの領域で遂行した。

（１）深層格タグ付きコーパスの開発と自動推定技術の開発
（２）機能表現 WEB 辞書の開発
（３）複合動詞の多義性を解消するための知識源の開発
（４）コーパスを用いた機能表現の定量分析

（１）については表層格であるニ格、ガ格、ヲ格、デ格、ノ格の５つの格について深層格リストを作成した。リストの作成にあたっては、まず EDR の関係子及び先行研究の分類基準を参照し、言語学の専門家チームが最終的に決定した。設計方針としては、他の助詞との置き換え可否やニ格に前接・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し、意味上の隔たりが小さいものは一つにまとめるように設計した。

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であり、（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深層格推定に利用するためであり、（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であるためである。これら三種類のコーパスに対し、ニ格についてのみ各コーパス２万フレーズの深層格コーパスを構築した。ガ格、ヲ格、デ格、ノ格については Web コーパスに対して深層格タグ付きコーパスを構築した。

次に構築されたニ格のコーパスに対し、深層格の自動推定を行った。自動推定にあたっては、ナイーブベイズ分類手法を用いた。ナイーブベイズ手法は設計が単純でありながら、その分類性能はある程度の高さを持っていることから、汎用性が高く幅広く利用されている。

（２）については、まず『日本語文型辞典』等のリストに基づき機能表現リストを構築した後、パターンマッチで抽出するシステムを構築した。抽出は４つの流れ（①入力文を MeCab を用いて形態素解析、②表現文型リストと照合 、③機能表現・複合動詞・名詞修飾節等を自動抽出④本文とリンクさせて表現文型リストを表示）で行うよう開発した。

（３）については『複合動詞用例データベース』を構築した。用例 DB は、Web データをもとに半自動的に構築したデータベースで、複合動詞の用例、語構成情報に加えて、構成動詞の用例を収録している。収録対象の複合動詞は、「動詞(連用形)＋動詞」タイプの「語彙的複合動詞」(影山 1993)である。収録語数は、複合動詞が 3912 語、単一動詞が 1148 語である。収録語の用例数（中央値）は、複合動詞が 977 例、単一動詞が 5858 例である。

４．研究成果

4.1 ニ格深層格の自動付与（推定）

３種類のコーパスに対する深層格情報をもつニ格についてナイーブベイズ分類手法を用いて深層格の自動付与作業を行い、精度評価を行った。そのことにより、コーパスの妥当性の検討も行った。

コーパスの訓練セットとしての妥当性を検討したところ３つのコーパスのうち BCCWJ が最も訓練セットとして適していることがわかり、分類正答率は評価用コーパスに対して最大 62%の結果が得られた。

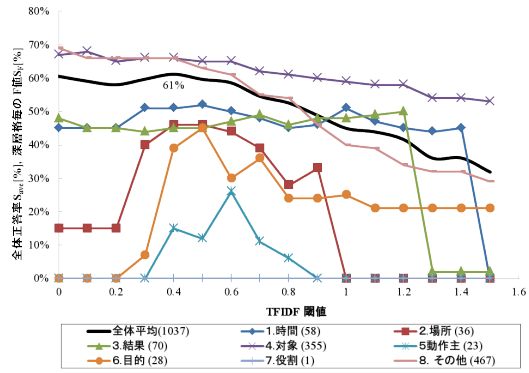

図１深層格の自動付与結果の正答率の変化

続けて日本語語彙体系を利用した汎化および TFIDF による素性選択を試みた.その結果として分類性能の大きな向上は見られなかったものの、図１の深層格推定において「8. その他」、「4. 対象」が深層格推定において大きな重みを占めていることが判明した。今後はこれらの深層格に対して改善を加えることで全体の分類性能向上を目指す。

4.2 表現文型抽出器の開発

表現文型抽出ツールのインターフェースを図２に示す。この抽出ツールにより、接続詞・接続助詞，機能表現の抽出が可能になった。同時に複合動詞や名詞修飾節を導く名詞の抽出も可能になった。

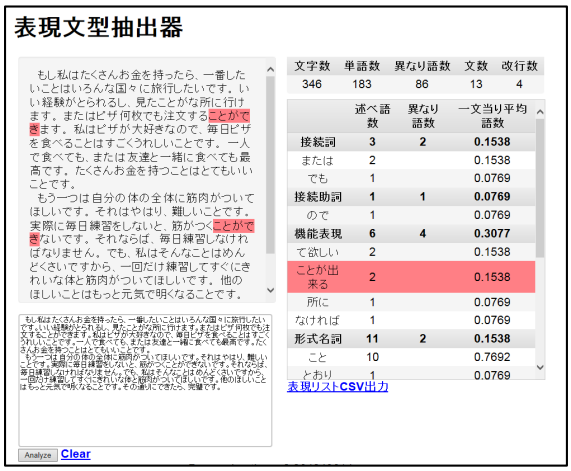

図２　表現文型抽出ツール

本抽出ツールにより数多くの機能表現を検索することが可能になり、一定の研究成果を挙げることができた。

4.3　複合動詞の多義性に関する研究

複合動詞の多義性が発生するメカニズムに構成動詞がどのように関わっているかを『複合動詞用例データベース』を用いて分析した。分析対象の多義性は、(a)比喩による多義性、(b)（複合動詞が）複数の項構造を持つことによる多義性である。(a)の例として「(記録などを）塗り替える」の語義に対する比喩的な語義「(壁などを）塗り替える」を挙げる。(b)に相当する複合動詞として「噴き出す」があり、非対格自動詞としての語義（例：「温泉が噴き出す」）と他動詞（例：「タバコの煙を噴き出す」）を持つ。
分析の結果、(a)には構成動詞の多義性に起因するもの、複合動詞自体の多義性に起因するものがあることがわかった。また、(b)に該当する 17 の複合動詞を示し、11 動詞で構成動詞の多重項構造が複合動詞の多義性につながっていることなどを明らかにした。

この研究は、（本科研で構築した）単一動詞の辞書記述を複合動詞の辞書記述に反映する手がかりとなるものである。

4.4 コーパスを用いた機能表現の定量分析

深層情報を付与したコーパスを用いた言語研究を行い、定性的研究と定量的研究の異なりについて分析した。一つはニ格に関する分析、もう一つは接続の機能表現であるナガラに関する分析である。

ニ格については本科研で構築した深層格タグつきコーパスを使用し、コーパスにおける深層格の出現比率の異なりや、出現頻度情報に基づく深層格間の関係性を主成分分析で分析した。その結果、同一の深層格であってもコーパスによって出現比率が大きく異なることが明らかになった（表１）。また、先行研究で一般的に挙げられていた場所の深層格は使用全体の中ではそれほど高頻度ではないことも明らかになった。しかし主成分分析の結果、場所は他の深層格より共起語に特徴があり、判別しやすいことも明らかになった。こうした助詞のような高頻度に出現する機能表現であっても、コーパスにおける出現頻度が異なる深層格があるということが明らかにできたことは、本科研の成果といえる。

表１　３種のコーパスに対する深層格付与

	Web	京大	BCCWJ	χ^2 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	

ナガラ節についても同様に BCCWJ から抽出したナガラ節を人手で意味解釈し、文法研究によって蓄積された規則がどの程度、

言語使用の予測に寄与するかを検証した。その結果、動詞＋ナガラ節の用法は９割が付帯状況であり、逆接は１割にすぎないこと、定性的に重用視されていた「テイル形＋逆接」の用法より「動詞＋ナガラモ」のほうが逆節を自動判定するための寄与率が高いこと等が明らかになった。

このように、本研究の推進により、機能表現の自動意味判別を行うための言語資源の構築、深層格の自動推定方法に関する研究の進展、機能表現抽出辞書開発における貢献、機能表現の定量分析に関する貢献を果たすことができたと言える。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計５件）
①宮永愛子、松田真希子、聞き手配慮要素からみた超級日本語話者の発話の特徴、日本語／日本語教育研究、査読有、Vol.5、2014、1-17
②松田真希子、宮永愛子、庵功雄、 超級日本語話者の談話特性―テキストマイニングを用いた分析―、 国立国語研究所論集、 査読有、Vol.5、2013、43-63
③森篤嗣、 日本語教育文法にどうアプローチするか―コーパス調査はどこから始まるか―、 日本語学、 査読なし、 2013、32-7
④川村よし子、インターネット上で利用可能な文章の難易度判定システムの開発、ヨーロッパ日本語教育、査読有、16号、2012、194-198
⑤Makiko Matsuda、 Yuta Takemoto and Kazuhide Yamamoto. Phrase-based Statistical Machine Translation via Chinese Characters with Small Parallel Corpora. International Journal of Intelligent Information Processing (IJIIP), 査読有, Vol.2, No.3, 2011, 52-61

〔学会発表〕（計 11 件）
①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、札幌、2014.
②松田真希子、 森篤嗣、 川村よし子、 庵功雄、 山本和英 、 山口昌也、 ニ格深層格の定量的分析、言語処理学会第 20 回年次大会、札幌、 2014.
③松田真希子、庵功雄、 限界性を有する事態に対する否定の応答形式をめぐって、 日本語学会 2013 年度春季大会、大阪、 2013.
④山口昌也、 多義複合動詞の語義構造の分析、 国語研究所 第 4 回 コーパス日本語学ワークショップ、 東京、2013.
⑤川村よし子、 機能表現および文型に着目した表現文型抽出ツールの開発、 AATJ Annual Spring Conference、フィラデルフィア、アメリカ、 2014.
⑥森篤嗣、中島明則、岩田一成、テキスト評価ツール「やさ日チェッカー」の開発と指標の有効性の検証、The Eighth International Conference on Practical Linguistics of Japanese（ICPLJ8)、東京、2014.
⑦森篤嗣、意味判別における文法記述の効果の計量化 ―ナガラ節の意味判別を例として―、国立国語研究所領域指定型共同研究プロジェクト「学習者コーパスから見た日本語習得の難易度に基づく語彙・文法シラバスの構築」第１１回共同研究会、東京、2014
⑧森篤嗣、文法記述はナガラ節の自動意味判別に資するか、 日本語文法学会第 14 回大会、東京、2013.
⑨杢真奈見、 山本和英、「やさしい日本語」変換システムの試作、 言語処理学会 19 回年次大会、名古屋、 2013.
⑩川村よし子、専門分野の辞書を組み入れた日本語学習者のための Web 辞書の開発、2012 年日本語教育国際研究大会（ICJLE2012)、名古屋、2012.
⑪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、広島、2012.

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）

○取得状況（計０件）

〔その他〕
ホームページ等
研究プロジェクトページ
https://sites.google.com/a/jnlp.org/matsuda2013/
機能表現抽出器
http://lab.chammarit.com/phrase_analyzer/index.rb

６．研究組織
(1)研究代表者
松田　真希子（MATSUDA Makiko）
金沢大学・国際機構・准教授
研究者番号：10361932

(2)研究分担者

森　篤嗣（MORI Atsushi）
帝塚山大学・現代生活学部・准教授
研究者番号：30407209
川村　よし子（KAWAMURA Yoshiko）
東京国際大学・言語コミュニケーション学部・教授
研究者番号：40214704
山本　和英（YAMAMOTO Kazuhide）
長岡技術科学大学・工学部・准教授
研究者番号：40359708
山口　昌也（YAMAGUCHI Masaya）
大学共同利用機関法人人間文化研究機構国立国語研究所・言語資源研究系・准教授
研究者番号：30302920
庵　功雄(IORI Isao)
一橋大学・国際教育センター・准教授
研究者番号：70283702