様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月 １０ 日現在

機関番号：３２６６０
研究種目：若手研究
研究期間：2018～2019
課題番号：１８Ｋ１２３５７

研究課題名（和文）項省略を考慮した日本語の統語的ブートストラッピング

研究課題名（英文）Syntactic bootstrapping with null arguments in Japanese

研究代表者
折田　奈甫（Orita, Naho）

東京理科大学・理工学部教養・講師

研究者番号：７０７８１４５９

交付決定額（研究期間全体）：（直接経費）　　1,900,000円

研究成果の概要（和文）：本研究では、項や格助詞の省略が頻繁な日本語における、統語的手がかりを用いた動詞の意味の学習に焦点を当て、日本語絵本述語項構造コーパスの構築と分析、そして計算機モデルを用いたシミュレーションを行った。主要な成果として次の3点がある。①これまでにないコーパスを構築し、絵本は子ども向け発話と比較して項と格助詞の省略が少ないことを示した。②子ども向け自然発話の分析が中心だったこれまでの第一言語獲得研究に対して、異なる種類のインプットを観察/分析することの重要性を示唆した。③日本語の動詞の意味クラスの学習においてどのような情報が役立つかを定量的に検証した。

研究成果の学術的意義や社会的意義
これまでにない言語資源を構築し、このコーパスの分析とシミュレーションを通して、子どもの動詞の意味の学習メカニズムの解明に与する新たな知見が得られた。特に、子ども向け発話のテキスト情報のみを分析してきたこれまでの言語獲得研究に対して異なる種類のインプットを分析する重要性を示唆した点と、日本語の動詞の意味クラスの獲得を定量的に検証するためのたたき台を築いた点において学術的意義が高い。
　動詞の意味の学習メカニズムに関する知見は、効率的な語彙学習方法を研究する第二言語習得分野のみならず、自然言語処理や人工知能など、計算機に自然言語を学習させる応用分野にとっても重要な知見である。

研究成果の概要（英文）：This project focused on the learning of Japanese verb meanings using syntactic cues, where these cues, arguments and case markers, are very frequently absent in the input. We built Japanese Picture book Predicate-Argument Structure corpus, analyzed this corpus to investigate the distribution of syntactic cues in picture books, and conducted simulations to explore how and to what extent Japanese verb classes could be learned using the cues in picture books. The corpus is unprecedented in the field. The analysis demonstrated that a different form of input may contain different kinds and amounts of evidence, suggesting the importance of examining a variety of input available to learners. The simulations built a basis to further explore the relation between input and the learning of verb meanings.

研究分野：言語学

キーワード：心理言語学　言語獲得　統語的ブートストラッピング　絵本　項省略　日本語　動詞　格助詞

１．研究開始当初の背景

言語獲得研究では、子どもが動詞の項の数を手がかりにその動詞の大まかな意味を推測し学習するという統語的ブートストラッピング仮説（Landau & Gleitman 1985; Gleitman 1990）が様々な言語で検証されてきた。日本語の子どもを対象にした実験でも、項の数と文法関係を示す格助詞が未知の動詞の意味の推測に用いられると報告されている（Matsuo et al. 2012; Suzuki & Kobayashi 2017）。一方で、日本語の子ども向け発話では、動詞の項や格助詞が高頻度で省略される（Rispoli 1995; Matsuo et al. 2012）。日本語のインプットでは統語的ブートストラッピング仮説による学習がうまくいくための情報が不十分であるにもかかわらず、日本語の子どもは他言語の子どもと同じように多様な動詞の意味を獲得する。このインプットと実際に獲得される知識との間の隔たりは、これまで様々な言語において検証・支持されてきた統語的ブートストラッピング仮説の妥当性に疑問を呈している。本研究は、動詞の意味の学習に役立つ手がかりがインプットでどのように分布するかという問題に焦点を当て、日本語における統語的ブートストラッピング仮説の妥当性の検証を行う。

２．研究の目的

日本語の動詞、その項、格助詞がインプットでどのように分布し、どのような情報を持つかを明らかにする。高頻度で省略される動詞の項や格助詞が、動詞の意味の学習の手がかりとなるかを検証する。これらの結果から統語的ブートストラッピング仮説の言語普遍性を問う。具体的には次の３段階で研究を行う。①日本語絵本述語項構造コーパスを構築する。②このコーパスを分析し、動詞、項、格助詞などがインプットでどのような特徴を持ち分布するかを検証する。③このコーパスを用いた動詞の意味の学習の計算機シミュレーションを行う。

３．研究の方法

(1) 日本語絵本述語項構造コーパスの構築

日本語のベストセラー絵本に対して、述語項構造、有生性、絵との対応などの情報をアノテーションしたコーパスを構築する。言語学専攻の大学院生をアノテーターとして雇用し、アノテーションツール brat（Stenetorp et al. 2012）を用いてタグ付けを行う。

(2) 日本語絵本述語項構造コーパスの分析

絵本を自然発話以外の重要なインプットであると仮定し、絵本に、動詞の意味の学習の手がかりとなる項や格助詞がどのように分布するのか、また、これらの手がかりが動詞の意味の学習にどの程度寄与し得るかを、(1)で構築したコーパスを用いて定量的に分析する。

(3) 動詞の意味の学習のシミュレーション

動詞の意味クラスの学習のシミュレーションを行う。日本語絵本述語項構造コーパスをインプットとし、どのような手がかりがどの程度あればどのような動詞クラスを学習できるかを検証する。計算機モデルは、英語の動詞の意味クラスの学習のシミュレーションを行った Pearl and Sprouse(2019)のベイジアンモデルを日本語に応用する。

４．研究成果

(1) 日本語絵本述語項構造コーパスの構築

日本語のベストセラー絵本 50 冊に対して、述語項構造、有生性、絵との対応などの情報をアノテーションしたコーパスを構築した。合計 3,359 の述語に対してタグ付けを行った。図１に見本を示す。また、表１に述語カテゴリごとの頻度と動詞の例を示す。

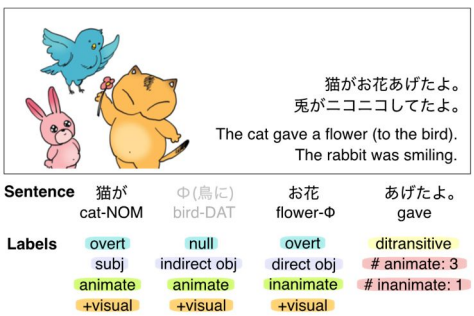

図 1 絵本述語項構造コーパスの例

(2) 日本語絵本述語項構造コーパスの分析

絵本では、動詞、項、格助詞がどのように分布するのか、また、項や格助詞は動詞の意味を学習する上で手がかりとなり得るかを定量的に分析した。まず、自動詞と他動詞に焦点を当て、自然発話では高頻度で起こる項と格助詞の省略が、絵本ではどの程度起こるのかを分析した。表２に結果をまとめる。子ども向け自然発話を分析した先行研究（Rispoli 1995; Matsuo et al.2012）と比較して、絵本では動詞の項と目的語を示す格助詞ヲの省略が少ない。

表 1 絵本述語項構造コーパスの動詞の分布

動詞カテゴリ	頻度	頻度の高い動詞（上位 5 動詞）
自動詞	1,339	行く、来る、ある/いる、出る、やって来る
他動詞	1,184	言う、見る、食べる、引く、思う
3 項動詞	79	やる、教える、見せる、着せる、載せる
形容詞	204	おいしい、いい、面白い、ひどい、楽しい
コピュラ	553	
Total	3,359	コピュラ、言う、行く、来る、見る、食べる

表 2 自動詞と他動詞の分布

動詞フレーム		絵本（本研究）	子ども向け発話 Rispoli (1995)	子ども向け発話 Matsuo et al. (2012)
自動詞	NP_{subj} V	**71.7%** (897)	37.5% (84)	54% (486)
	Φ_{subj} V	28.3% (354)	62.5% (140)	46% (419)
	合計	1251	224	905
他動詞	NP_{subj} NP_{obj} V	**47.3%** (387)	9.9% (22)	16.5% (144)
	Φ_{subj} NP_{obj} V	28.6% (234)		63% (551)
	NP_{subj} Φ_{obj} V	9.3% (76)	57.4% (128)	4.5% (39)
	Φ_{subj} Φ_{obj} V	14.9% (122)	32.7% (73)	16% (140)
	合計	819	223	874
	格助詞 ヲ	**55.8%** (457)	6.6% (19)	8.5% (58)

次に、これらの分布が、他動詞かどうかを予測する上でどの程度役立つのかを、線形混合モデルを用いて分析した。結果を表 3 に示す。格助詞と項の数のどちらもが、他動詞を予測する上で有意に影響した。また、係数$\hat{\beta}$の値から、格助詞ヲと 2 項の情報は他動詞を予測する上で強い影響があり、1 項の情報は自動詞を予測する上で強い影響があることがわかった。

表 3 他動詞を予測する線形混合モデル

予測子		$\hat{\beta}$	$SE(\hat{\beta})$	z	p	χ^2(df)	p_χ
格助詞 ヲ		5.6	0.6	9.3	<.001	973.8(1)	<.001
項の数	1項	-0.49	0.14	-3.4	<.001	250.7(2)	<.001
	2項	18.7	39.5	0.47	0.64		

まとめると、子ども向け自然発話と比較して、日本語の絵本は項と格助詞ヲの省略が少ないことがわかった。また、これら項の数と格助詞ヲの情報は、動詞の自他を予測する上で有意に影響することがわかった。これまでの多くの第一言語獲得研究では、子ども向け発話の分析だけをもとにして言語獲得とインプットの関係を議論してきた。しかし、絵本のような異なる種類のインプットでは、子ども向け発話では得られない質と量の情報が存在することを本研究は示した。

(3) 動詞の意味の学習のシミュレーション

上記(2)では、日本語の絵本は動詞の自他を予測する上で十分な項の数とヲ格の情報を含んでいる可能性を示した。しかし、実際の学習者は、省略されたりされなかったりする項、様々な格、有生性などの多様な情報を用いて、自他のみならず様々な動詞の意味を学習すると考えられる。省略や曖昧性を含むこれらの情報から、どのような日本語の動詞の意味クラスがどの程度学習可能であるかはわかっていない。そこで、英語の動詞の意味クラスの学習をシミュレーションした Pearl and Sprouse(2019)のベイジアンモデルを日本語に応用し、日本語絵本述語項構造コー

パスをインプットに、絵本からどのような動詞クラスが学習され得るかを検証した。以下に、予備実験（シミュレーション）の結果を報告する。

動詞の意味クラスは、自動詞と他動詞の２クラスの他、影山(1996)などを参考に、統語および/または有生性と対応がある８つの意味クラス（非対格、非能格、非対格と非能格どちらも可能、３項動詞、心理動詞、知覚動詞、発話動詞、その他）を学習対象とした。

学習のために利用する情報は、意味的情報として、ガ格/ヲ格/ニ格の項の有生性を、統語的な情報として、「格フレーム全て」、「ガ格/ヲ格/ニ格のみ」、「名詞句の数」の 3 種類を用意し、どのような情報があればどのような意味クラスの学習が可能かを探る。具体的には以下の表 4 の組み合わせを検証した。統語的情報のうち、「格フレーム全て」は、主語や目的語を示すガ格/ヲ格/ニ格だけでなく、述語と関係する項に付く全ての格助詞と副助詞を含む。これらは未知の動詞のより詳細な意味を推測する上で役立つ可能性がある一方、格助詞そのものが持つ曖昧性や述語と共起する格助詞の多様性が、意味クラスの学習を困難にさせる可能性もある。ガ格/ヲ格/ニ格のみは、学習者が主語と目的語という文法関係のみに焦点を当てた場合を仮定している。「名詞句の数」は、学習者が項の数のみに焦点を当てた場合を仮定している。

日本語絵本述語項構造コーパスから、頻度５以上の自動詞と他動詞（３項動詞含む）合計 120 種類（1684 動詞）を抽出し分類対象とした。モデルの分類結果は、ランド指標（1 に近づくほど良い）により評価した。ランド指標の結果を表 4 に、表 5 はモデル#1 が分類したクラスの一部とその内訳を示す。

表 4 動詞の意味クラスの学習 ランド指標

モデル#	統語的情報	有生性	自他２クラス	8 クラス
1	格フレーム全て	あり	**0.568**	**0.821**
2	格フレーム全て	なし	**0.568**	**0.821**
3	NP の数	あり	0.524	0.654
4	NP の数	なし	0.537	0.634
5	ガ/ヲ/ニのみ	あり	0.582	0.794

表 5 モデル#1 が分類した意味クラスの例（合計 13 クラスを学習；そのうちの 7 クラスの動詞）

意味クラス	動詞
1	やってくる、帰る、来る
2	下る、通り抜ける、知る
3	転ぶ、聞こえる、びっくりする、泣く、起きる
4	言う、怒鳴る、考える
5	乗る、入る、住む
6	作る、拾う
7	開ける、歌う、捕まえる、探す

表 4 の結果から、格フレーム全ての情報を利用するモデル（#1, #2）が、有生性の情報の有無にかかわらず、自他のみ 2 クラスと 8 クラスのどちらの分類においても最も学習目標に近い。表 5 のモデル#1 が学習した意味クラスを観察すると、クラス 1 と 2 は有方向移動の自動詞であり非能格、クラス 3 は非対格、クラス 4 は態度動詞、クラス 5,6,7 は他動詞と対応している。学習した 13 クラスの内の残りの 6 クラスについては学習目標と対応するようなパターンは見られなかった。これらの結果は、無関係な情報や曖昧な情報を含んでいても、一見雑多な情報がある方がうまく学習できる可能性を示唆している。NP の数やガ/ヲ/ニ格のみの方が自他 2 クラスの学習には優位に思えるが、項や格助詞の省略が起こるインプットでは、一見してわかりやすい情報は、雑多な情報と比較してあまり役に立たないと考えられる。また、この実験設定では有生性の情報の有無は統語情報を越えて役立たないこともわかった。自他のみ 2 クラスと 8 クラスの分類を比較すると、全てのモデルにおいて自他のみの分類の方が難しいことも定量的に示された。今後はさらに詳細に結果を分析し、動詞の意味クラスの学習とインプットとの関係をより明確に特定したい。

(4) 結果まとめ

本研究課題では、項や格助詞の省略が頻繁な日本語における統語的手がかりを用いた動詞の

意味の学習に焦点を当て、日本語絵本述語項構造コーパスの構築、このコーパスの分析、そして計算機モデルを用いたシミュレーションを行った。2 年間の研究期間で主要な目標は概ね達成できた。特に、(i)絵本テキストに対して丁寧に述語項構造をアノテーションしたこれまでにないコーパスを構築したこと、(ii)絵本は子ども向け発話とは異なる量と質の情報を含むことを定量的に示したこと、(iii)日本語の動詞の意味クラスの学習においてどのような情報が役立つかを定量的に検証したこと、の三点が顕著な成果であると考える。(i)と(ii)については、分野では最高レベルの国際学会（Annual meeting of the Cognitive Science Society, Boston University Conference on Language Development）で発表し、(iii)については情報処理学会自然言語処理研究会にて招待公演の一部として発表をした。認知科学、発達言語学、そして自然言語処理という様々な分野で認知される機会が得られた。(i)と(ii)の内容については、2020 年の Proceedings of 44th Boston University Conference on Language Development に掲載されている。(iii)については現在追加の分析とシミュレーションを行なっており、論文誌投稿を目指している。

＜参考文献＞

Gleitman, Lila. (1990). The structural sources of verb meanings. Language acquisition, 1(1), 3-55.

Landau, Barbara, and Lila Gleitman. (1985). Language and experience: Evidence from the blind child. Harvard University Press.

Matsuo, Ayumi, Sotaro Kita, Yuri Shinya, Gary C. Wood, and Letitia Naigles. (2012). Japanese two-year-olds use morphosyntax to learn novel verb meanings. Journal of child language, 39(3), 637-663.

Rispoli, Matthew. (1995). Missing arguments and the acquisition of predicate meanings. In Michael Tomasello & William E. Merriman (Eds.), Beyond names for things: Young children ' s acquisition of verbs, 331-352.

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Junichi Tsujii. (2012). BRAT: a web-based tool for NLP-assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 102-107.

Suzuki, Takaaki, and Tessei Kobayashi. (2017). Syntactic cues for inferences about causality in language acquisition: Evidence from an argument-drop language. Language Learning and Development, 13(1), 24-37.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Naho Orita, Asumi Suzuki, Yuichiro Matsubayashi	４．巻 2
２．論文標題 The Input to Verb Learning in Japanese: Picture Books for Syntactic Bootstrapping	５．発行年 2020年
３．雑誌名 Proceedings of the 44th annual Boston University Conference on Language Development	６．最初と最後の頁 457-464
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計4件（うち招待講演　1件／うち国際学会　2件）

１．発表者名 Naho Orita, Asumi Suzuki, Yuichiro Matsubayashi
２．発表標題 Verb arguments in Japanese picture books
３．学会等名 41th Annual Meeting of the Cognitive Science Society（国際学会）
４．発表年 2019年

１．発表者名 折田奈甫
２．発表標題 日本語絵本における動詞の項と格助詞の省略について
３．学会等名 Computational Psycholinguistics Tokyo
４．発表年 2019年

１．発表者名 Naho Orita, Asumi Suzuki, Yuichiro Matsubayashi
２．発表標題 The Input to Verb Learning in Japanese: Picture Books for Syntactic Bootstrapping
３．学会等名 44th Boston University Conference on Language Development（国際学会）
４．発表年 2019年

１．発表者名 折田奈甫
２．発表標題 計算モデルを用いた第一言語獲得研究 －統語的・意味的言語知識の獲得を例に－
３．学会等名 情報処理学会 自然言語処理研究会（招待講演）
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
連携研究者	松林　優一郎 (Matsubayashi Yuichiro) (20582901)	東北大学・教育学研究科・准教授 (11301)	