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

平成　３０　年　　５　月　３０　日現在

機関番号：　１１３０１
研究種目：　研究活動スタート支援
研究期間：　2016～2017
課題番号：　１６Ｈ０６６２３

研究課題名（和文）母語獲得における分かりやすい学習事例の解明

研究課題名（英文）Investigating clear learning instances in first language acquisition

研究代表者
折田　奈甫（Orita, Naho）

東北大学・情報科学研究科・特任助教

研究者番号：７０７８１４５９

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：言葉の学習を可能にするインプットの性質を探るべく、日本語の動詞に焦点を当て、絵本テキストに対して述語項構造や非言語情報をアノテーションした絵本述語項構造コーパスの設計と構築を行なった。このような言語資源はこれまでに例がなく、今後様々な言語獲得研究に応用できる可能性がある。また、このコーパスを用いて、日本語の動詞の意味の学習の手がかりになると提案されている項や格が、絵本でどのように分布するか調査を開始した。子ども向け自然発話と比較して、絵本はより明示的に項や格を使用する傾向があることが明らかになった。

研究成果の概要（英文）：We designed and built Japanese Picturebook Predicate-Argument Structure Corpus to explore properties of clear learning instances in the learning of verb meaning in Japanese. Our preliminary study that investigates the distribution of cues for the learning of verb meaning in Japanese found that there are less argument and case omissions in picture books compared to child-directed speech.

研究分野：　言語学

キーワード：　言語獲得

１．研究開始当初の背景

言語獲得研究においてインプットの詳細な分析は不可欠である。言語経験から何をどの程度学びうるのか、あるいは学びえないのか、どのような先行知識を仮定すれば学習が可能なのか、というような言語獲得の本質的問題と関係するからである（Chomsky 1986）。

近年の母語獲得研究では、子どもは多くの学習事例から言葉と意味の関係を学習するのではなく、起こるのは稀でもわかりやすい事例から言葉の学習を行うという即時マッピング仮説（Dollaghan 1985 他多数）が提案されている。しかし、この「わかりやすい学習事例」がどのような性質・特徴を持つのかについては、過去数年英語を中心に限られた言語知識について検証されているのみであり（Cartmill et al. 2013, Trueswell et al. 2016 など）、さらなる調査が必要である。このような背景から、日本語の非言語情報を含んだインプットの特徴を定量的に抽出し、「わかりやすい学習事例」についての知見を深めることを目指し研究を開始した。

２．研究の目的

本研究の目的は、即時マッピングを可能にする学習事例の特徴・性質を明らかにするため、日本語のインプットの言語・非言語情報の特徴を抽出・記号化し、学習事例の「わかりやすさ」を定量化して検証することである。

本研究では特に、日本語の動詞の意味の学習におけるインプットに焦点を当てる。子どもが新しい動詞を学習するには、その動詞が示す行為や状態などの抽象的な事象を特定する必要があり、単純な観察のみによる動詞と意味のマッピングは原理的に困難である（Gleitman 2005）。そこで、学習者は動詞の項の数をヒントに動詞の大まかな意味（causativity）を推測するという統語的ブートストラッピング仮説が提案された（Landau and Gleitman 1985）。

しかし、動詞の項の数が言語普遍的手がかりであることを支持する他言語を対象とした先行研究の一方で、日本語の子どもは項の数ではなく言語固有の格助詞という手がかりから動詞の意味を推測すると提案されている（Suzuki & Kobayashi 2016）。日本語は、文脈があれば「太郎がほめた」「花子をほめた」のように項を省略でき、自然発話ではこのような省略が高頻度で起こるため（Rispoli1995, Matsuo et al. 2012）、日本語の子どもが項の数ではなく格助詞を手がかりに動詞の意味を学習すると考えるのは妥当に思える。しかし、省略された項は本当に手がかりにならないのだろうか。省略された項であっても、指示対象が目前に存在したり談話において顕著であれば項として理解され、日本語においても項の数が動詞の意味の推測をする時の手がかりとして用いられる可能性はないだろうか。

本研究は、この仮説を検証するための第一段階として、言語資源の設計と構築を主に行う。以下で詳細を記述する絵本述語項構造コーパスを構築することにより、省略項と格助詞はインプットでどのように分布しどのような特徴や情報を持つのか、学習者はどのようなインプットを得て動詞の意味を推測しうるのか、動詞の意味の学習における「わかりやすい学習事例」とはどのような特徴を持つのかなどを検証し、今後の学習シミュレーションや実験の土台にする。

３．研究の方法

経緯　研究開始当初は養育者と子供の日常生活の様子を録画しデータにする予定だったが、事前に承諾を得ていた実験協力者からキャンセルの申し出があり、以降協力者が見つからなかった。代替として、言語情報とそれに対応した非言語情報（視覚情報）の抽出が可能な絵本からコーパスを構築する。

日本語のインプットを分析した言語獲得研究では、主に子ども向け発話データベース CHILDES（MacWhinney 2000）が用いられてきた。子ども向け自然発話から多くの知見が得られてきた一方で、自然会話の一部を切り取ったデータであるがために、共有する知識や文脈が推定しにくいことや、視覚情報が伴わないなどの制限があり、省略の解釈や談話情報の推定が難しい。

このような日常生活の一部を切り取った子ども向け自然発話と比較して、絵本は談話や視覚情報（絵本の絵）が限定されている。そのため、省略された項の解釈が容易になると考えられ、視覚情報を含んだ意味関係や談話情報をうまくとらえられる可能性がある。

絵本述語項構造コーパス　本研究は、このような利点のある絵本テキストに対して、図 1 のような述語項構造タグを付与したコーパスの構築を行う（exo1, exo2 はそれぞれ 1 人称 2 人称、null は項省略を表す）。

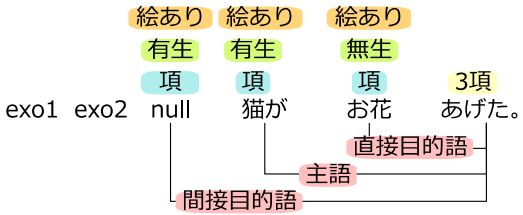

図 1 アノテーション見本

述語項構造は、文の構造のみならず、動詞の意味、項省略、格など、言語獲得の主要な問題と密接に関係する。将来的には共参照タグを付与し、談話レベルの問題にも対応できる。絵本テキストに対して述語項構造をアノテーションすることで、これまでの子ども向け自然発話データでは扱えなかった言語獲得の問題にも対応できる言語資源の構築を目指す。

コーパスの設計・構築では、自然言語処理分野で培われてきた日本語の述語項構造アノテーションの知識や技術（笹野ら 2017）を応用する。

コーパス作成手順 日本語のミリオンセラーの絵本（2017 年 3 月 31 日時点で 85 冊）を選定し、図 2 の手順でコーパスを作成する。

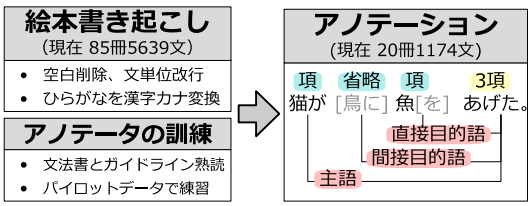

図 2 コーパス作成手順概要

絵本の読み聞かせ対象年齢は 0 歳から 5 歳の範囲で、3 歳から 4 歳を対象とする絵本が最も多く（47 冊）、次に 0 歳から 1 歳が多い（27 冊）。

絵本の本文テキストの人手による書き起こし作業では、アノテーション作業効率化のために、絵本に元々ある、主に文節区切りを表す空白やレイアウトのための改行を削除し、文単位で改行した。この文単位による改行後の行数の合計は 5639 行となった。

絵本はひらがなが多く一般的な解析器ではうまく解析できないため（藤田ら 2014）、文単位改行後にテキストのひらがなをできるだけ常用漢字やカタカナに変換し、これをアノテーション対象のデータとする。

アノテータには、言語学の基礎的訓練を受けた言語学専攻の大学院生二名を雇用した。アノテータは、基本的な日本語文法の復習とガイドラインの熟読をし、研究代表者からアノテーション方法について指導を受けた後、パイロット用データにアノテーションし、研究代表者がその結果を確認した後にアノテーション作業を開始した。アノテーションには brat（Stenetorp et al.2012）を用いる。

アノテーション仕様 図 1 のように、述語とその項（省略含む）、述語と項の文法関係、項の指示対象の有生性、項の指示対象が絵本の絵に出ているかどうかの 5 種類の情報に対しタグ付けを行う。タグ付与範囲は IPA 品詞体系の形態素区切りに基づく。

本アノテーションは、先行する述語項構造アノテーションと比較して以下の特徴を持つ点で新しい。（1）自動詞、他動詞など動詞の種類のタグを付与する。（2）述語の主語と補部のみに項タグを付与し、意味役割を担う付加部にはタグを付与しない。（3）項に相当する名詞句の指示対象の有生性のタグを付与する。（4）視覚情報を取り入れた分析を進めるため、絵に指示対象が出ているかどうかのタグを付与する。

述語項構造アノテーションの後には、共参照・照応のアノテーションを予定している。これにより、顕著性、話題性、情報構造など、様々な言語現象と関連する談話レベルの情報を定量化できるようにする。

4．研究成果

コーパスの設計と構築 2018 年 3 月 31 日時点で、39 冊の絵本、2408 文に対するアノテーション作業が完了している。

今後の研究に耐えうる言語資源が構築できているかを確認するため、絵本 20 冊 1174 文を対象に、アノテーション作業者間一致を計算した。一名の結果を正解、もう一名の結果をシステムの推定と仮定し、適合率、再現率、F 値を計算した結果を以下に示す。一致率は、述語/項の種類とアノテーション範囲の両方が合えば一致として計算している。

表 1 作業者間一致

タグ	適合率	再現率	F1
述語	88.4% (1330/1504)	87.8% (1330/1515)	88.1%
項	94.3% (1519/1611)	93.4% (1519/1626)	93.9%

述語の一致率は、動詞をカテゴリカルに分類することの難しさを反映している。他動性は連続的であるため（角田 2009）、動詞によって作業者間で揺れが観察される。「に」格の付く名詞句を項とするか付加部とするかで判断が分かれる場合が多い。項の不一致は、イベントを表す名詞句や態により項が追加されている場合に不一致が多い。

項の範囲が一致しているものを対象に、有生性と絵の有無のタグの一致率を算出した結果、有生性は 93.3%、絵の有無は 92.1%となった。どちらのタグにおいても、省略項の解釈で作業者間の揺れがあった。

これら作業者間の揺れは、自然言語の持つ本質的な難しさから生じるものであり、一意にタグを決められない事例が作業者間不一致に反映されている。これらの結果から、提案するタグ付け仕様は概ね妥当であり、今後の研究に耐えうる言語資源であると考える。

インプットの分析 –日本語の動詞の意味の学習の手がかり– 作成したコーパス（39 冊の絵本 2408 文）を用いて、日本語の動詞の意味を学習する時に、絵本ではどのような手がかりが利用可能か調査を開始した。

まず、コーパスに出現する述語と項の種類ごとの頻度を表 2 と表 3 にそれぞれ示す。項は省略、1 人称、2 人称を含む。表 2 から述語は自動詞と他動詞が大半を占めることがわかる。表 3 の補文は、発話・思考・感情を表す動詞が補部としてとるものが大半である。

次に、動詞のおおまかな意味を推測する手がかりになると提案されている項と格が、子ども向け自然発話と比較して絵本ではどの程度出現/省略されるかを調べた。

表 2 述語の種類ごとの頻度

述語の種類	トークン	タイプ
自動詞	1162	381
他動詞	1053	319
3 項動詞	71	30
形容詞	175	71
合計	2461	801

表 3 項の種類ごとの頻度

項の種類	トークン数
項（実体あり）	3024
補文	240
イベント性名詞句	226
追加項	35
合計	3525

表 4 自動詞の分布

項と格（主語）	% （トークン数）
項あり+ガ格	31% (361)
項あり+ハ	38% (445)
項あり+その他	1% (10)
項あり+格省略	4% (49)
項省略	26% (297)
項あり合計	74% (865)

	0 項有 +ヲ/ニ	0 項有 +ハ	0 項有 +他	0 項有 +格省略	0 項省略	合計
S 項有 +ガ	7.4% (78)	0.1% (1)	0%	5.0% (53)	3.5% (37)	16% (169)
S 項有 +ハ	26.4% (278)	0.6% (6)	0.4% (4)	7.7% (81)	5.7% (60)	41% (429)
S 項有 +他	0%	0.5% (5)	0.1% (1)	0%	0.4% (4)	1% (10)
S 項有 格省略	0.9% (9)	0.4% (4)	0%	0.3% (3)	1.2% (13)	3% (29)
S 項省略	16.7% (176)	2.4% (25)	1.5% (16)	6.9% (73)	12.0% (126)	40% (416)
合計	51.4% (541)	3.9% (41)	2.0% (21)	19.9% (210)	22.8% (240)	1053

表 5 他動詞の分布

子ども向け発話における項と格の省略の頻度を調べた先行研究（Rispoli 1995, Matsuo et al. 2012）を要約すると、自動詞文では約半数の主語が省略され、他動詞文においては主語・目的語が格の有無に関わらず共に発話される割合は他動詞文全体の 1〜2 割程度、主語・目的語にガ格ヲ格が両方とも付く文は他動詞文全体を母数にすると 1%程度、主語・目的語の項がある他動詞文を母数にすると 20%程度となっている。子ども向け自然発話では、項と格どちらの省略頻度も一貫して高い。

一方、本研究の絵本述語項構造コーパスから、絵本における自動詞と他動詞の項・格省略の割合は、子ども向け自然発話と比較して低いことがわかった。絵本における自動詞と他動詞の分布を表 4 と表 5 に示す。

絵本における自動詞の主語省略率は 26%で、子ども向け自然発話を分析した Matsuo et al. (2012)の主語省略率 54%よりも低い結果となった。副助詞が最も多く、格の省略自体は低頻度である。

他動詞で典型的なガ格主語とヲ/ニ格目的語を両方伴うものは全体の 1 割に満たないが、主語が何であれ目的語がヲ/ニ格と共起する他動詞は全体の約半数となり、子ども向け自然発話の 6%強と比較するとかなり高い。また、主語・目的語が格の有無に関わらず共に発話される割合は約 50%（523）であり、子ども向け発話の 10〜20%と比較すると高頻度である。

これらの結果から、絵本は子ども向け自然発話と比較するとより明示的に項や格を使用する傾向があることが明らかになった。

まとめ 本研究では、絵本テキストに対して述語項構造、有生性、絵の有無などをタグ付けした絵本述語項構造コーパスの設計と構築を行った。このような言語資源は過去例がなく、今後様々な言語獲得研究に応用できる可能性がある。また、このコーパスを用いて、日本語の動詞の意味の学習の手がかりとなりうる項や格の情報が絵本にどの程度あるのか調査を開始し、絵本は子ども向け発話と比べて項と格の省略頻度が低いことを明らかにした。

今後の展望 絵本述語項構造コーパスのアノテーションを継続し、共参照タグの付与も行う。このコーパスをさらに分析し、手がかりの種類・組み合わせごとの動詞の分布や非言語情報の特徴や影響など、インプットの分布を詳らかにする予定である。また、手がかりの有無や種類にばらつきがある中で様々な動詞の意味をどのように推測し学習しうるのかを検証するため、このコーパスを入力情報にした計算機モデルよるシミュレーションを行う予定である。

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　　1 件）

(1)折田奈甫，石井啓太，鈴木あすみ，松林優一郎．絵本述語項構造コーパスの設計と構築.言語処理学会第 24 回年次大会. 2018 年 3 月.pp432-435.[査読なし]

〔学会発表〕（計　　2 件）

(1)折田奈甫，石井啓太，鈴木あすみ，松林優一郎．絵本述語項構造コーパスの設計と構築.言語処理学会第 24 回年次大会. 2018 年 3 月.
(2)折田奈甫.コーパスを用いた談話情報の定量化．言語資源活用ワークショップ 2017 年.国語研究所.

〔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 なし

６．研究組織
(1)研究代表者
折田　奈甫（Naho Orita）
東北大学・情報科学研究科・特任助教
研究者番号：70781459
(2)研究分担者 なし
(3)連携研究者 なし