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　６ 年　 ６ 月 １９ 日現在

機関番号：１２１０２
研究種目：研究活動スタート支援
研究期間：2019～2023
課題番号：１９Ｋ２３０６８

研究課題名（和文）大規模児童作文コーパスにおける埋め込み節の発達の計量的分析

研究課題名（英文）A quantitative analysis of the development of embedded clauses in a large corpus of children's writing

研究代表者
今田　水穂（Imada, Mizuho）

筑波大学・人文社会系・助教

研究者番号：１０５７９０５６

交付決定額（研究期間全体）：（直接経費）　　2,200,000円

研究成果の概要（和文）：児童作文研究のための言語資源の整備と、それを利用した構文の複雑性に関する研究を行った。言語資源については、既存の作文コーパスの整備と節境界ラベルの付与を行った。 統語的複雑性については、自然言語における係り受け距離の増大が抑制的であることを確認する一方で、ごく短い文においてはランダムに生成した構造よりも係り受け距離が長くなることを確認した。これは児童の統語能力の発達が、複雑化と合理化の混合によって複雑に進行していることを示唆する。また節の種類の分析からは、等位構造から従位構造へ、話し言葉的な文体から書き言葉的な文体へという変化が学齢の上昇に伴って観察されることを実証的に確認した。

研究成果の学術的意義や社会的意義
データを基盤とした言語の実証的研究においてはデータと分析手法の両方が必要となるが、児童の統語能力の発達は、データの入手困難性と、語彙情報に比べて構文情報の分析の技術的要求の高さから、比較的活発に研究されてこなかった領域と言える。本研究の成果は、児童の統語的能力の発達という言語学的な課題に対して、データを用いた実証的な知見を提示する。また、この研究は言語学における言語資源の活用領域を語彙研究から構文研究への拡大を促進すること、および自然言語における統語構造の数学的特性についての理解を深めることに貢献する。

研究成果の概要（英文）：We developed a linguistic resource for the study of children's compositions, and used it to study syntactic complexity. First, we modified an existing composition corpus and further annotated clause boundary labels. We then analysed syntactic complexity and observed that while dependency distance is suppressed in natural language, rather long dependency relations are often observed in short sentences. This result suggests that the development of children's syntactic competence has two aspects: complexity and rationalisation. The analysis of clause types confirms that the change from coordinate to subordinate structures and from spoken to written language occurs with increasing school age.

研究分野：日本語学

キーワード：児童作文　コーパス言語学　統語的複雑性　節境界

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

2000 年代以降、特に国立国語研究所のコーパス整備計画を中心として、日本語の多様な位相における大規模コーパスの整備が進み、日本語研究においてデータを基盤とした実証的研究の手法は一般的なものになりつつある。一方で、オープンな言語資源の構築には常に著作権や許諾の問題が伴い、どのようなデータも自由に構築・共有できるわけではない。ほとんど共有が進んでいないジャンルの一つは子供の書き言葉や話し言葉のデータである。児童作文の資料は日本語を母語とする子供の言語発達を研究する上で非常に高い有用性を持ちながら、許諾処理の困難さから、ほとんどオープンデータとして公開・共有されることがない。そのため、この分野の研究者は独自に資料を収集し、クローズドなデータセットを構築して研究に利用することが一般的である。

このような状況を背景として、我々の研究グループは 2010 年代以降、小中学生の児童・生徒の作文資料を収集し、いくつかのコーパスを構築した。その中でも「児童・生徒作文コーパス」は特定の小中学校の全児童・生徒の作文を 2014 年度から 2016 年度の 3 年間にわたって悉皆的に収集・電子化した 164 万形態素規模のコーパスである。特定の学校のみを調査対象とした点、作文のテーマ「夢」「頑張ったこと」の 2 種に限定される点において代表性を有する子供の書き言葉コーパスとは言えないが、データの均質性、悉皆性、大規模性において学年を横断した統計的分析に適した稀有なコーパスである。このデータを利用して、これまで語彙多様性の分析など、いくつかの研究課題を展開してきた。

電子化されたコーパスの多くは、形態素解析や構文解析などの機械処理によって言語学的情報が付与される。このうち形態論情報は、その扱いの容易さから広く日本語学や言語教育の研究に利用されている。語彙多様性の研究も、形態論情報を利用した研究の一つである。一方で構文情報については、技術的な難しさから比較的利用が進んでいない。例えば文章の読みやすさを評価するリーダビリティ研究においても、文字、単語、文節などの言語単位の数、文の長さ、品詞や語種の使用率などは変数として使われることが多いが、文の統語的構造が評価指標として利用されることは少ない。

一方で、文の統語的構造については言語発達や話者の認知能力と関連するいくつかの特性があることが既に知られている。例えば、母語話者や第二言語学習者は発達や学習が進むに従って埋め込み構文をよく使うようになることが知られているし、また、自然言語では距離の長い係り受けが避けられる傾向があることが知られている。こうした自然言語の統語的特性は、一般的な言語資源や、比較的整備が進んでいる第二言語学習の資源をデータとして研究が行われてきたが、データの入手困難性から、母語話者のデータを用いた実証的研究はごく限定的にしか行われていない。

要するに、言語の実証的研究のためにはデータと分析技術の両方が必要だが、母語話者児童の統語的特性の量的な分析は、データの入手性、要求される分析技術の両面における敷居の高さにより、研究がそれほど活発に行われてこなかった領域だと言える。そのため、この領域の研究を推進することは、いくつかの面で重要性を持つ。第 1 に、児童の統語能力の発達という言語学的な課題について、データ基盤的およびデータ駆動的な実証研究を展開することができる。第 2 に、従来よく研究されてきた語彙の観点に注目するのではなく文法（特に統語）の観点に注目することは、言語研究におけるコーパスの活用領域を広げ、より多様な観点からの量的研究法の開発に貢献する。第 3 に、これらの研究を通じて、自然言語の数学的性質の理解を深めることができる。

２．研究の目的

日本語を母語とする児童の言語発達の過程を統語能力の観点から評価する手法を開発するために、児童作文コーパスを用いて統語構造の複雑性の数値化および統計的モデル化のための研究を行う。特に、学齢の進行に伴う埋め込み構造の使用傾向の変化と、依存構造（係り受け構造）の数学的モデル化とその特性の理解に焦点を置く。

３．研究の方法

主なデータとして「児童・生徒作文コーパス」を使用する。このコーパスは 164 万形態素規模の統計的研究に適した設計のコーパスであり、UniDic に依拠した形態論情報と文節単位の係り受け構造が付与されている。本研究は、文の統語的構造の分析が主目的であるため、このうち係り受け構造のデータを主要な分析対象とする。

第 1 に、係り受け構造の数学的特性、特に文の統語的複雑性を数値化する方法について検討する。係り受け構造は、文節をノード、係り受けをエッジとするグラフ構造とみなすことができる。ただし、ノードに線形順序（語順）が設定されていることが特異であり、特に係り受け距離の計算に関連する。言語学的な観点から見ると、まずノードの数（文の長さ）自体を統語的複雑性の数値化とみなすことができ、従来のリーダビリティ研究などでも利用されてきたところではあるが、同じノード数の文であっても、係り受けの長さや埋め込みの深さは構造によって異なり、

それによって文の複雑さには違いがあると考えることができる。そこで、この構造の複雑性を依存関係にあるノード間の線形距離(係り受け距離)やルートノードまでの距離(係り受けの深さ)を用いて数値化する方法について検討する。また、ランダムに生成された構造と自然言語との複雑性の差異や、学齢による複雑性の差異を統計的に分析する。統計的手法としては、主にフィッティングによる分布の特性の検討と、作文の著者をランダム変量とする一般化線形混合モデル分析を実施する。

第2に、コーパスに対して節境界ラベルを付与し、節の種類が文の複雑性とどのように関連するかについて検討する。母語話者や学習者は言語の習得・学習が進むに従って埋め込み構造をよく使うようになることが知られており、グラフ理論的に同等の構造であっても、等位構造と従位構造では統語的複雑性について異なる評価を与える必要が考えられる。2020 年代以降、依存構造 (Universal Dependencies) による構文解析が利用可能になり、ノードやエッジの種類も取得できるようになったが、本研究が開始された 2019 年当時は文節係り受け構造に基づく構文解析が主流であり、この構造はノードやエッジの種類に関する情報を含んでいなかった。そのため、これを補うために節境界情報を付与し、ノード（文節）の種類を表す情報として利用することを考えた。節境界については「現代日本語書き言葉均衡コーパス」に対して節境界ラベルを付与したデータ（BCCWJ-CBL）というモデルケースがあったが、このラベルを自動付与するツールは公開されていない。そのため、SVM によって系列ラベリングを行う汎用的なアノテーションツールの YamCha を用いて BCCWJ-CBL のデータを学習し、構築したモデルを用いて児童作文コーパスに対して節境界ラベルの自動付与を行った。このデータを利用して、学齢の違いによって節の選択がどのように変化するかを質的、量的に分析する。

４．研究成果

研究に先立って、言語資源の整備と構築を行った。まず「児童・生徒作文コーパス」の形態論情報の確認・修正作業を行い、併せて係り受け構造の再解析も実施して、パッケージのバージョンを更新した。研究期間を通じて数回の更新を行い、コーパスのバージョンを 1.2 から 1.6 まで更新した。次に YamCha で BCCWJ-CBL を学習して節境界モデルの構築を行った。BCCWJ を学習データと検証データに分割して行った実験では、90%後半程度の高い精度で節境界ラベルが付与できることを確認した。構築したモデルを使用して児童作文コーパスに対する節境界ラベル付与を実施し、児童作文の節境界ラベルデータを構築した。この言語資源を利用して、統語的複雑性の数値化や、節の種類の学齢による変化などに関する研究を実施した。

統語的複雑性については、文の長さ（文節数）に加えて係り受け距離平均（MDD）および階層係り受け距離（MHD）を複雑性の数値化として利用し、その妥当性と数学的特性について検討した。ランダムな文節列においては、文長の分布は幾何分布に従うことが理論的に明らかであるが、作文データにおいては文長の分布は対数正規分布ないしガンマ分布によく当てはまり、学齢が上がるほど対数正規分布からガンマ分布に接近する傾向が見られた。幾何分布とガンマ分布は事象が発生する間隔の分布、対数正規分布は乗算過程によって生成されるデータ分布として知られており、文の統語構造が成長する過程の理解について一定の示唆を与える結果と言える。MDD の分布は、ランダムに生成した構造では対数正規分布によく当てはまったが、作文データではガンマ分布によく当てはまった。また、MDD の平均は、ランダムデータでは文長に対して冪的に増加するのに対し、作文データでは対数的に増加していた。これは、自然言語では係り受け距離の増大は抑制的であるということを示しており、長い係り受けは避けられる傾向があるという従来の知見を実証的に支持する結果である。一方で、5 文節以下の短い文においては、ランダムデータよりも作文データの方が MDD が大きくなる傾向が見られた。その要因として自然言語はランダムデータと比べて単文の構成要素数が大きくなりやすいことが考えられ、児童の統語能力の発達が、構造の複雑化（要素数の増大）と合理化（係り受け距離の最適化）の混合による複雑な過程であることを示唆する結果と言える。MHD についてはまだ十分な分析結果が得られていないが、可能な構造のパターン数が MDD と共通の漸化式で得られる点、MDD と MHD の間にある程度の負の相関が確認できる点など、MDD と MHD の対称性を示唆するいくつかの知見が得られた。

節境界ラベルの分析からは、児童の言語発達における 2 つの傾向性が確認された。第 1 に、学齢の上昇は連体節、補足節などの埋め込み構造を形成する節の増加と、条件節、理由節、並列節などの等位構造を形成する節の減少を引き起こす。第 2 に、類似する機能を持つ節標識においても学齢によって選好の違いが見られ、さらに一部の節類型においては「が」と「けど」類の対立、「けど」と「けれど」の対立というような複数の節選択の混合による複雑な語彙選択の変化が観察された。これらの結果は、児童作文における語彙や文法の選択が話し言葉的なレジスタから書き言葉的なレジスタへと推移することを示しており、児童の書き言葉におけるレジスタの獲得過程を実証的に可視化する結果と言える。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 Imada Mizuho	４．巻 12:379
２．論文標題 Distribution of sentence length and dependency distance in children's compositions: Characteristics of natural language and variations in language development	５．発行年 2023年
３．雑誌名 F1000Research	６．最初と最後の頁 1-17
掲載論文のDOI（デジタルオブジェクト識別子） 10.12688/f1000research.132383.1	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 －

１．著者名 今田，水穂；田川，拓海；文，昶允；那須，昭夫	４．巻 -
２．論文標題 「児童・生徒作文コーパス」に対する節境界ラベル付与	５．発行年 2021年
３．雑誌名 F1000 Reseach	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.12688/f1000research.40669.1	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 －

１．著者名 今田，水穂	４．巻 84
２．論文標題 児童の言語発達と係り受け次数の増加	５．発行年 2023年
３．雑誌名 文藝言語研究	６．最初と最後の頁 21-35
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1件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 今田，水穂
２．発表標題 児童作文における係り受け距離と階層距離
３．学会等名 言語資源活用ワークショップ2021
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

「児童・生徒作文コーパス」に対する節境界ラベル付与
https://doi.org/10.17605/OSF.IO/5KCRB

「児童作文における文節数および係り受け距離の分布」
https://doi.org/10.17605/OSF.IO/U72NK

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

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関