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　元 年　 ６ 月　 ５ 日現在

機関番号：１１３０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ０２６５４

研究課題名（和文）高度な統語・意味解析情報を持つコーパスの開発とその応用

研究課題名（英文）Development and application of an advanced corpus with syntactic and semantic information

研究代表者
吉本　啓（Yoshimoto, Kei）

東北大学・高度教養教育・学生支援機構・教授

研究者番号：５０２８２０１７

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：日本語テクストに対し統語・意味解析情報をアノテーションとして与えて開発中のコーパスに関して2つの課題がある。第一に、一般に利用可能なソフトウェアを利用した形態素解析は、正確さおよび情報量の点で問題がある。また論理意味表示において、意味役割はタグ付けされていない。さらに、コーパスの日本語研究への応用法も開拓する必要がある。形態素解析に関しては、幼児言語発達コーパスCHILDES Japanの方式を取り入れて全面的に改訂を行っており、これにより同コーパスデータのツリーバンク化にも道が開けた。意味役割に関しては述語項構造シソーラスとリンク付けを行い、自然言語処理やAIにも応用することが可能になった。

研究成果の学術的意義や社会的意義
日本語に関してこれまでに利用できなかった、句構造解析にもとづくコーパス NINJAL Parsed Corpus of Modern Japanese を開発することの意義は大きいが、本研究における形態素解析の改善により、それがさらに利用しやすくなった。また、幼児言語発達データのツリーバンク化は、これまでほぼ未開拓であった大量データにもとづく統語・意味能力発達の研究に道を開く。さらに、意味格情報のアノテーションは日本語の意味論的研究の強力なツールとなるだけでなく、言語処理や AI 研究へのインパクトも大きい。

研究成果の概要（英文）：We have solved two problems in the development of a modern Japanese corpus with syntactic and semantic annotations, NINJAL Parsed Corpus of Modern Japanese. First, we have improved its morphological analysis by adopting a morphological tagging system for the study of first language acquisition data. This has also paved a way for building up a treebank from Japanese speaking children's data. Second, we have developed a resonable method to give semantic roles as part of our annotations. This will also have an essential influence on the study of natural language processing and artficial intelligence.

研究分野：コーパス言語学

キーワード：コーパス　日本語　統語論　意味論

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

近年、言語情報処理技術の発達によって、大量の言語データに対し有用な言語情報を付加(タグ付け)したコーパスの構築が可能になり、開発も行われるようになった。しかし、形態素解析を中心として文節間の係り受け関係をタグ付けしたのみの現在の日本語コーパスでは、文法に興味を持つ研究者は限定された利用しか出来ない。例えば、格助詞「が」「を」が表示する格役割はそれぞれ曖昧であり、その区別には単なる形態素情報以上のアノテーションが必要である。さらに、すでに 1960 年代に三上章が指摘しているように、文をまず文節に区切って係り受け関係をまとめるやり方では、語や句の持っている情報を文の情報へと合理的に統合していくことが出来ない。現状の日本語コーパスを利用して行える研究は、共起(cooccurrence)関係を利用して言語データの範囲を限定することだけであり、そこから真に有用なデータを得るには人手によるチェックが必要である。

これに対して世界では、Penn Treebank を代表とする、文に対して統語解析情報をタグ付けしたツリーバンクが主流になろうとしている。これは句構造文法にもとづいたもので、文節係り受けを用いたアノテーションとは異なり、語や句が文全体の統語・意味情報にどう貢献しているかを明示したものである。しかし、現状のツリーバンクは決して完璧なものではなく、その利用にはやはり限界がある。例えば、関係節による修飾のような非有界依存構文(unbounded dependency)における依存関係は標準的な形式文法のやり方では 2 つの語句のインデクス付けにより関連づけられるが、この方式で人手によりアノテーションを行うのは、非常な手間が掛かるため、本格的なツリーバンクでは行われていない。かと言って、統語解析ソフトウエアを用いて非有界依存構文の解析までも自動的に行うことは、今に至るまで実用段階に達していない。

研究代表者である吉本と研究協力者のバトラーは、現行の Penn Treebank 方式の表層的な統語解析情報を利用し、これをバトラーの開発した文意味解析ソフトウェアに入力することによって、文の論理意味情報(述語論理式)を自動出力する研究を行ってきた。上記の非有界的な依存関係についての情報は論理意味表示に含まれているので、これを統語構造へとフィード・バックさせて、当該の語句と語句を関連づけることができる。この方法がほとんどの日本語構文に対応できる頑強なものに達したことから、本格的な規模の Penn 方式の日本語文統語解析情報アノテーションを行い、さらにこれに対し自動意味解析により得た論理意味表示も付加したコーパス開発を平成 16 年度より国立国語研究所共同研究プロジェクトとして開始した。

こうして、現代日本語の書き言葉文に対し、従来の作業量と同程度のアノテーションを加えるだけで、文統語・意味解析情報を付加したコーパスの開発が可能になった。この成果をより高い価値のあるものにするためにはいくつか課題がある。第一に、一般に利用可能な形態素解析ソフトウェアを利用した形態素解析は、言語研究において要求される正確さおよび情報量の点で問題がある。また、同じく現状の論理意味表示においては、格に関する情報は主語・目的語等、統語的レベルにとどまり、意味役割はタグ付けされていない。さらに、このような高度の情報を持つコーパスを利用した経験が我が国の言語研究者にはほとんど無く、日本語研究への応用法も自ら開拓する必要があり、またそのような応用を意識したコーパス開発が望ましい。

２．研究の目的

文統語・意味解析情報付きコーパスの構築とそれを利用した研究を深化させるために、2 つの課題について研究を行う。第一に、コーパス構築の基礎となる形態素解析をより正確で言語学的に意義あるものへと改訂する。第二に、文統語・意味解析情報付きコーパスを応用した、新しい研究領域を開拓する。具体的には、幼児言語発達データのツリーバンク化および自然言語処理・AI に応用可能な意味格データベース構築のための基礎作りを行う。

これまで、コーパス構築の基礎となる形態素解析は、UniDic 辞書を使用する形態素解析器 MeCab およびもう 1 つの形態素解析器 Comainu を使用し、その結果に人手で修正を行って使用してきた。これらの辞書および形態素解析器は自然言語処理用に開発されたもので、定評もあって安定的に使用できるものだが、問題が無いわけではない。第一に、自動形態素解析の常として、誤りが生じる。誤り自体は避けられないとしても、これらのソフトウェアは使用者にとってブラック・ボックスであるため、誤りを修正して精度を向上させることができない。また、これらは学校文法 (伝統文法) にもとづいており、言語学的な分析が不十分である。例えば、五段活用動詞の可能形 (読める、行ける、等) は独立した動詞として扱われているため、助動詞「られる」が付加されて作られる、他の活用型の動詞の可能形と一括して検索条件とすることができない。さらに、日本語に習熟していない使用者を考慮して各単語にグロス (訳) を付けることも重要な課題である。

課題の解決のために、幼児言語発達データのコーパス作りを行ってきた宮田スザンネ教授(愛知淑徳大学) とともに検討を行った。宮田教授は、言語発達コーパス CHILDES Japan の開発者であり、コーパス開発を目的として形態素解析器 JMOR を構築してきている。検討の結果、これまでに行ってきた形態素解析を完全に改め、JMOR にもとづく形態素解析を新たに行うことにした。これにより、上に指摘した問題点を解決することができる。

JMOR を形態素解析器として採用することは、宮田教授が開発してきた幼児言語発達コーパ

ス CHILDES Japan と NPCMJ とが共通の基盤を持つことを意味する。これにより、CHILDES Japan に対し、従来のような依存関係でなく、NPCMJ と同様に句構造をアノテートしたツリーバンクを開発して利用者の利便性を高めることが可能になる。そこで、幼児言語発達データに NPCMJ 方式のアノテーションを施すための基礎検討が必要になる。

高度の意味情報を持つ述語項構造データをまとめたシソーラスが日本語について作られれば、NPCMJ についても、より意味的側面に立ち入った利用が可能になる。このような目的から、かねてから述語項構造シソーラスの研究を行っている竹内孔一講師 (岡山大学) と共同研究を行っている。竹内講師が概念フレームと意味役割を付与した例文に対し、NPCMJ の意味解析により得られる、Arg0, Arg1 のように番号付けされた意味役割を新しく付け加える。これにより、一般的、抽象的な意味格の理解と、具体的、個別的な概念識別の両方の長所を兼ね備えた意味役割り情報の把握が可能になる。さらに、この方式で NPCMJ に対し概念・フレームと意味役割を与える作業を進める。

３．研究の方法

宮田教授が開発してきた JMOR は元来幼児の話し言葉を対象としたものであり、書き言葉を中心とする一般のテクストを扱ってきた NPCMJ とは対象が異なる。このため、両者が共存しうるような拡張的な枠組みを新しく作り出し必要がある。

また、幼児言語発達データのツリーバンク化に関しても、データのアノテーションに際して問題点が散見され、その基礎的な検討が必要である。

NPCMJ および CHILDES Japan の言語データの中から代表的なものを選び、試行錯誤的にアノテーションを行うことを通じて、問題のあぶり出しと解決を図る。

さらに、竹内講師のデータや NPCMJ に対し概念フレームおよび意味役割を与える課題についても、複数のアノテーター間の統一をいかに維持するか等、問題は多い。

竹内講師のシソーラスの中から代表的な例文を選んで NPCMJ 方式の意味格をアノテートし、2 種類の意味格のマッチングについて知見を蓄積する。

４．研究成果

以下に述べるように、コーパスの改善と応用について十分な基礎作りを行うことができた。

形態素解析

基本的に JMOR に従いながら、NPCMJ のデータをカバーするために必要な拡張を行った。これにより、一般的な日本語テクストに対し適用可能な段階に至ることができた。

現在、NPCMJ データ全体への適用（形態素解析の全面的改訂）を国立国語研究所チームで進めている。

基礎となる辞書のフォーマットについて検討を重ね、満足できる段階に達した。これについても、NPCMJ 処理のための拡張を今後進める。

幼児言語発達データのツリーバンク化

NPCMJ 方式の統語解析を与えるに際しての問題点を検討し、ほぼ解決を得た。これにもとづき、今後、CHILDES Japan のツリーバンク化を進める。

研究成果を 2019 年 7 月の言語科学会国際大会ワークショップで宮田教授が発表する予定である。

意味格のアノテーション

2 種類の意味格を付与していくために、十分な検討を行うことができた。今後は量の拡大が可能となる。

研究成果を 2018 年 5 月の International FrameNet Workshop 2018、2019 年 3 月の言語処理学会年次大会で発表した。さらに、2019 年 7 月の言語科学会国際大会ワークショップでも竹内講師が発表する予定である。

日本語に関してこれまでに利用できなかった、句構造解析にもとづくコーパスを開発することの意義は大きいが、本研究における形態素解析の改善により、それがさらに利用しやすくなった。また、幼児言語発達データのツリーバンク化は、これまでほぼ未開拓であった大量データにもとづく統語・意味能力発達の研究に道を開く。さらに、意味格情報のアノテーションは日本語の意味論的研究の強力なツールとなるだけでなく、言語処理や AI 研究へのインパクトも大きい。

５．主な発表論文等

〔雑誌論文〕（計　3　件）

吉本啓・プラシャント-パルデシ「文の統語・意味解析情報をタグ付けした日本語構造体コーパスの開発」，KLS Proceedings，関西言語学会，2016 年．

吉本啓「統語・意味解析情報を伴う日本語コーパスの開発とその日本語教育・学習への応用」，『日本言語文藝研究』第 18 号，pp．1-11，台湾日本語言文芸研究学会，2018 年 6 月．

周振・吉本啓「統語・意味情報付きコーパスの開発に関する研究：中国語名詞句の解析について」『国立国語研究所論集』、第 17 号、1-32．

〔学会発表〕（計　18　件）

Alastair Butler, Ai Kubota, Shota Hiyama and Kei Yoshimoto．“Treebank Annotaion of FraCaS and JSeM”, Logic and Engineering of Natural Language Semantics, the Japan Society of Artificial Intelligence, National Institute for Japanese Language and Linguistics, Tachikawa, 2016 年 11 月 13 日．

吉本啓「アノーテーション方式とコーパスの特色」，ワークショップ「統語・意味解析情報付き日本語コーパスの構築に向けて」，日本言語学会第 153 回大会，福岡大学，2016 年 12 月 4 日．

Kei Yoshimoto．“Tenses in Japanese Complex Sentences”, Workshop/Symposium Philosophy of Mental Time V: Time in Language, Nihon University, Tokyo, 2017 年 1 月 28 日．

周振・Alastair Butler・吉本啓「中国語名詞句の内部構造について」，『言語処理学会第 23 回年次大会発表論文集』pp．46-49，筑波大学，2017 年 3 月 14 日．

Stephen Wright Horn, Alastair Butler and 吉本啓 “Keyaki Treebank Segmentation and Part-of-Speech Labelling”, 『言語処理学会第 23 回年次大会発表論文集』, pp．414-417，筑波大学，2017 年 3 月 15 日．

周振・Alastair Butler・吉本啓「中国語助詞の解析」，言語科学会第 19 回国際年次大会 (JSLS2017)，ハンドブック pp．190-191，京都女子大学，2017 年 7 月 1 日．

吉本啓「NPCMJ の概要」，吉本　啓・Alastair Butler・Stephen Horn・長崎郁「統語・意味解析コーパスの開発と言語研究」，第 2 回 NPCMJ チュートリアル～NINJAL Parsed Corpus of Modern Japanese（NPCMJ）を利用するための講習会～」神戸大学，2017 年 11 月 4 日．

吉本啓「統語・意味解析コーパス NPCMJ のアノテーション」，第 85 回 NINJAL コロキウム，国立国語研究所，2017 年 11 月 14 日．

小林昌博「コネクショニストモデルを使った第二言語（英語）習得の実験」， Conference Handbook 35，pp．261-266，日本英語学会，東北大学，2017 年 11 月 19 日．

吉本啓「日本語複合的機能語中の名詞の『名詞性』について」，シンポジウム「言語の変化、言語の成長―複眼的視点から」，Conference Handbook 35，pp．267-272，日本英語学会，東北大学，2017 年 11 月 19 日．

吉本啓「イントロダクション」，パネルセッション「統語・意味解析情報をタグ付けした日本語コーパスの開発―アノテーションの方法と文法研究への応用―」，『日本語文法学会第 18 回大会発表予稿集』,pp．113-121，日本語文法学会，2017 年 12 月 3 日．

Kei Yoshimoto and Akiko Takahashi．“Exploiting Coreferential Information in NPCMJ for L2 Reading of Japanese Texts”, NINJAL International Symposium “Exploiting Parsed Corpora: Applications in Research, Pedagogy, and Processing”, National Institute for Japanese Language and Linguistics, Tachikawa, 2017 年 12 月 10 日．

Kobayashi, M and W. Takinami．“Simulation of First Language Interference on Acquisition of Adjectival Participles in English: A Connectionist Approach”. Bulletin of Tottori University Education Center 13, pp．43-60, 2017.

長崎郁・アラステア-バトラー・スティーブン-ライト-ホーン・プラシャント-パルデシ・吉本啓「統語解析情報付きコーパス検索用インタフェースの開発」，『言語処理学会第 24 回年次大会発表論文集』，pp. 1123-1126，岡山コンベンションセンター，2018 年 3 月 15 日.

Stephen Wright Horn, Alastair Butler, Iku Nagasaki, and Kei Yoshimoto. “Deriving Mappings for FrameNet Construction from a Parsed Corpus of Japanese”, Tiago Timponi Torrent, Lars Borin and Collin F. Baker (eds.). The International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons Proceedings. The International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons, collocated with the 11th edition of the Language Resources and Evaluation Conference, pp. 28-32, シーガイアコンベンションセンター, 2018 年 5 月 12 日.

周振・アラステア-バトラー・吉本啓「中国語存現文の解析」言語科学会第 20 回国際年次大会，文京学院大学，ふじみ野市，2018 年 8 月.

吉本啓「言語研究と統語・意味解析情報付きコーパス」日本英語学会第 36 回大会シンポジウム「ツリーバンク開発と言語理論」，Conference Handbook 36，pp. 242-247，横浜国立大学，横浜市，2018 年 11 月 25 日.

吉本啓「統語・意味解析情報付き日本語コーパスの構築とその文法研究への応用」韓国日本語学会特別講演，『韓国日本語学会第 39 回国際学術発表大会予稿集』，pp. 23-31，明知専門大学，ソウル，2019 年 3 月 23 日.

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

６．研究組織

(1)研究分担者
研究分担者氏名：　森　芳樹

ローマ字氏名：　Yoshiki Mori

所属研究機関名：　東京大学

部局名：　総合文化研究科

職名：　教授
研究者番号（8 桁）：　30306831

研究分担者氏名：　小林　昌博
ローマ字氏名：　Masahiro Kobayashi
所属研究機関名：　鳥取大学
部局名：　大学教育支援機構
職名：　准教授
研究者番号（8桁）：　50361150

(2)研究協力者
研究協力者氏名：　アラステア J. バトラー
ローマ字氏名：　Alastair J. Butler