様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年　 ５ 月 ２６ 日現在

機関番号：１１３０１
研究種目：基盤研究(C)（一般）
研究期間：2013～2015
課題番号：２５３７０４５７

研究課題名（和文）次世代日本語コーパスプロトタイプの構築とその脳認知言語学実験への応用

研究課題名（英文）Development of New-type Corpus and Its Application to Neurolinguistic Experiments

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）　　2,900,000円

研究成果の概要（和文）：　日本語テクストに統辞・意味情報をタグ付けしたコーパスの開発を目指して、現代日本語の多様なデータに適合し、文自動意味解析を行っていくのに十分な統辞情報のタグ付け方法を確立した。実際に、開発した方法に従って 10,000 以上の文に対して統辞・意味解析情報のタグ付けを行った。また、開発した統辞情報タグ付け法を日英両語のマニュアルとして編集した。
　統辞情報のアノテーションにおいて問題となる事柄のうち、特に重要な、1個の助詞相当の連語、1個のモーダル助動詞相当の連語、および空要素の扱いについて検討を行った。

研究成果の概要（英文）：We have established a method to build up a corpus (treebank) which has sufficient information to automatically obtain semantic representations with syntactic annotations from diverse linguistic texts in Modern Japanese. Following the method we have established, we have annotated more than 10,000 sentences. In particular, we have solved problems concerning syntactic annotation, i.e., collocations which function as single P's, collocations functioning as modal auxiliaries, and null elements.

研究分野：コーパス言語学

キーワード：コーパス言語学　統辞論　意味論

１．研究開始当初の背景

現代日本語についてこれまでに作成、利用されているコーパスは形態素情報を付加したものが中心であり、統辞構造や意味に関する情報を得るためには限界がある。これに対して、世界の言語研究では、文の統辞解析情報を持つツリーバンク (treebank) が先端的なものとして使用されるようになってきている。これによってようやく、文の中での語句と語句との関係をとらえる可能性が開ける。しかし、日本語について、十分な量を備え通常のスタイルの書き言葉をカバーするツリーバンクは存在しない。そこで、その構築が急務となっている。

しかし、実はツリーバンクはあくまで表層の統辞構造をタギングしたものに過ぎないので、言語研究の目的には限界がある。文中での語句と語句との関係を調べたい場合、たとえば特定の動詞とその目的語となる名詞とが共起する頻度を知りたい場合を考える。これらは同一の節の中にあらわれるとは限らず、関係節や主題化など、いわゆる長距離依存によって隔てられている可能性がある。ところが、これらの構文についての直接の情報はツリーバンクには存在しないため、それだけでは求めたい頻度を知ることは出来ない。研究代表者らは、バトラーの提唱する意味解析理論 Scope Control Theory (SCT; Butler 2010) に従い、ツリーバンクに相当する文の表層的統辞構造を当該理論をインプリメントしたシステムに入力して、高精度の意味表示 (述語論理式) を出力する研究を行ってきた。そのためのプログラムはほぼ完成し、またコーパス開発の新機軸として関連学会での評価も得ている。

提案する日本語コーパスが言語処理分野において占める重要性は説明するまでもない。また、研究代表者は、脳認知言語学者の横山と脳機能画像法を用いて脳の言語処理過程を研究してきた。脳の反応は使用頻度の影響を受けやすく、文理解に関する脳機能データが文処理の違いによるのか、使用頻度の違いによるのか、特定できないことが多い。構文レベルでの頻度データを提供してくれるコーパスがここでも求められている。

２．研究の目的

本研究では、第一に、現代日本語の書き言葉の文に対して、ジャンルや文体を問わず、十分な統辞論的情報 (句構造) を均質的にタギングするための方法を確立し、実際に相当数の文に対して統辞情報を付加したプロトタイプ日本語ツリーバンクを構築する。さらに、これらの文に付加された統辞情報にもとづいて、文の意味情報 (述語論理式による意味表示) を SCT を利用して自動的にタグ付けするための手法を開発する。

本研究では、日本語コーパスに関して、これまでにない質の高さの達成を目標とする。しかし、単に机上の理論にとどまらず、相当量の語データを実際にタグ付けすることを通じて、近い将来における大規模コーパスへの道を開く。この経験を通じて高精度なコーパス作成のためのノウハウを得、作業マニュアルとして公開する。

また、東北大学医学系研究者とのこれまでの脳認知言語学共同研究を、コーパスを利用してさらに発展させる。本研究では、今までに頻度効果の可能性が指摘されている構文に対して、報告されている認知的処理の困難度の差が、使用頻度の差によるものであるかどうかを明らかにする。

３．研究の方法

提案する日本語ツリーバンクは、原則的に Annotation Manual for the Penn Historical Corpora and the PCEEC（以下、AMPHC; Santorini 2010）の規約に従って作成する。この方式は、極力フラットな統辞構造を採用してノードの数を減らすことと、名詞句、動詞句、節等に必要に応じて機能情報（主語、目的語、時間副詞句、節の様々な機能等）をタグ付けすることを特色としている。構造的曖昧性が問題になる場合の多くで統辞的埋め込みをフラットなままに未指定とすることが出来るので記述しやすく、また有用な文法情報に富んでいる。また多くの言語のコーパス開発に使われていることから、それらにおける多様な文法事象の取り扱いが日本語ツリーバンク作成に当たって参考になり、さらに外国人研究者にも利用しやすいという利点が生じる。

統辞解析は、フリー解析ソフトである MeCab および CaboCha をまず使用し、さらに句構造解析器により自動構文解析を行った上で行うが、人手による修正が主たる課題となる。修正には、主として木構造編集用フリーソフトウェアを用いる。日本語の日常的に使われる文法語や構文を網羅的に取り上げ、基本的な例文の統辞解析を行う。この経験を集積して、作業従事者の差を超えて均質なコーパスを構築するノウハウ確立の基礎とする。これを受けて、現実に使われた日本語の書き言葉文、約１万文をデータとして日本語ツリーバンクを作成する。この日本語コーパスプロトタイプの構築を通じて、統辞解析情報をタグ付けするための客観的基準をマニュアルとして確立し、一般に公開する。

研究の最終目標は論理意味表示の自動生成にあるので、意味解析ソフトウェアの入力として適切な形に整える必要がある。例えば、複数の語が格助詞やモダリティ助動詞等の１つの機能語に相応する場合（「に対して」「なければならない」等）は、全体の機能を明示する。また、省略された必須格は、ゼロ代名詞として表記する。さらに、並列構文は意味表示生成に際して大きな問題を生じるので、

統辞解析のための明確な指針を入念に作成する必要がある。

以上、タグ付け基準の客観性・明確性・一貫性、日本語使用の実情、意味表示からの要請、という時には互いに矛盾も生じるそれぞれの条件を最大限満たし、バランスの取れたコーパス開発のための方法を確立する。脳認知言語科学実験は、学内の機能的磁気イメージング計測機および脳磁図計測機を利用して行う。

４．研究成果

本研究によって、現代日本語の多様なデータに適合し、文自動意味解析を行っていくのに十分な統辞情報のタグ付け方法を確立した。実際に、開発した方法に従って 10,000 以上の文に対して統辞・意味解析情報のタグ付けを行った。また、開発した統辞情報タグ付け法を日英両語のマニュアルとして編集した。

統辞情報のアノテーション法について行った検討のうちで主要なものは、1 個の助詞相当の連語、1 個のモーダル助動詞相当の連語、および空要素の扱いについてである。

１個の助詞相当の連語

日本語にはしばしば、助詞や動詞等からなる連語が固定表現となって１個の助詞相当の機能を果たすことが少なくない。意味表示からの必要により、これらは各々1 つの助詞として扱うのが望ましい。しかし、これらの単語の構成素の独立性に関しては場合ごとに事情が異なるので、慎重な検討が必要である。データ中の用例にもとづき、構成素が自立的に使われる可能性および全体のまとまり性を考慮して､以下のものを１個の助詞（P）として認めることにした。

うえで、うえに、からすると、代わりに、際に、だけではなく、たって、たところ、ために、ための、っていう、で言う、という、というより、といえば、といった、として、としても、とすれば、とともに、とはいえ、ともに、ながらの、ならでは、に相次いで、に当たって、にあたり、にあたる、において、における、にかかわらず、に限らず、にかけて、に関して、に関する、に比べて、に際し、に際して、に従い、にして、にしろ、に対し、に対して、に対する、について、に次いで、につき、にとって、に伴う、に反して、に比して、にまつわる、に向けて、にも関わらず、によって、により、によりて、による、によると、によれば、にわたって、にわたり、にわたる、の代わりに、の方で、ほどなく、ように、よりも、わりに、を介して、を通じ、を通じて、を通して、を始め、をめぐって、をめぐる、をもって

１個のモーダル助動詞相当の連語

上の場合と同様に、日本語には連語が１つにまとまってモーダル的な機能を果たす場合が少なくない。これらについても、その多くを１つの助動詞（MD）としてタグ付けすることにした。その判断基準は以上と同じである。形態にバリエーションのあるものが多いが（「ない」と「ん」、丁寧の助動詞「ます」の有無等)、それらすべてを 1 個のモーダル助動詞としている。1 個の MD として扱うのは以下の連語である。

う、かも、かもしれない、かもしれません、こと、ことだ、ことだろう、ざるを得ない、そう、だろう、っちゃ、であろう、でしょう、てはいけない、てはだめ、てはならない、てもいい、ても良い、てよい、ないといけない、ないといけません、なくてはならない、なければいけない、なければならない、に違いない、ねば、ねばならない、ばいい、ば良い、必要がある、必要はない、べき、べく、べし、まじ、までもなく、もの、みたい、よう、(よ)うとする、らしい、らしく、わけ、わけがない、わけにはいかない

空要素

日本語の話し言葉では空範疇(ゼロ代名詞)がきわめて多くあらわれ、その検出と解析は話し言葉の文のより深い理解にとってきわめて重要である。本研究では、ツリーバンクにおけるゼロ代名詞の正確で有用なタグ付けをもっとも重視してアノテーションを行った。また、こうしたアノテーションの有用性を検証するために、ツリーバンクを用いて、日本語の空範疇検出を実現する方法について検討した。

日本語において空範疇は文脈指示上重要な役割を果たし、事実上代名詞の役割を果たしている。このため、ゼロ代名詞と呼ばれることも多い。Penn Treebank では‘ pro-drop ’と呼ばれるカテゴリーがそれに相当するが、英語をはじめとする西ヨーロッパ諸語ではその使用は限定されていることから、このカテゴリーをそのまま日本語に応用するには問題がある。そこで本研究では、日本語における使用の実情に適ったタグ付け法を開発した。

まず、通常「空範疇」と呼ばれるものをゼロ代名詞とコントロールとに分けた。ゼロ代名詞は概ね代名詞として文脈上の機能を有するものである。コントロールとは、複文において埋め込まれた節(従属節)がより上位の節(主節)の主題や主語を継承するものである。コントロールについては、次節でより詳しく述べる。また、統辞論では通常、関係節の中には主名詞に相当する格名詞句が表現されないまま存在すると考えるが、これは「トレース」と呼び、ゼロ代名詞にもコント

ロールにも含めない。
　ゼロ代名詞はさらに、以下のように細分される。

(1) *expletive*
　天候現象等のように、主語を持たないと考えられる述語に対してはこれがタグ付けされる。従属節の内部にこのような述語があらわる場合、主語が無いと主節の主語を誤って継承してしまう。そのためにこのタグ付けが必要である。

(2) *pro*
　いわゆる「小さい pro」や `pro-drop ' と呼ばれるもので、ほぼ英語の人称代名詞に相当する。本ツリーバンクでは、この種のゼロ代名詞の個々の指示対象までは扱わない。

(3) 話し手や聞き手を指すか、またはそれらを含む代名詞
　この種のゼロ代名詞は(2)とは区別する。その理由は、話し手や聞き手が主語であるということはきわめて重要な情報であり、話し言葉ツリーバンクとしてはタグ付けすることが望ましいからである。指示対象がそれぞれ話し手、聞き手である場合は *speaker* および *hearer*、両者である場合は *speaker+hearer*、話し手や聞き手を含む複数の指示対象である場合は *speaker+pro* や *hearer+pro* と記す。

(4) *arb*
　一般的非人称主語の場合。

(1)～(4)は人手によるアノテーションを必要とする。
　コントロール構文はチョムスキー派の統辞論では PRO の素性を与えられるが、本ツリーバンクではこれに相当するタグ付けを行わない。従属節が IP-TE（テ従属節）、IP-ADV（副詞節）、IP-INF（不定詞節）または IP-EMB（「外の関係」の関係節）で主語が明示されない時は、それよりも上位の節の名詞句をディフォールトの主語として継承する。この名詞句は、NP-SBJ（主語）＞ NP-OB1（直接目的語）＞ NP-LGS（受動文の論理主語）＞ NP-OB2（間接目的語）の順で優先される。ただし、前節で述べたように、従属節主語が *expletive* となっている時はこの限りではない。このディフォールトとしての主語情報の継承は意味解析(評価)において行われるため、特別なアノテーションは不要である。コントロール現象をディフォール的な継承として捉えることは日本語としての実情に適っていると考えられる。また、コーパス開発においても、ゼロ代名詞に関わるアノテーションの負担を大きく軽減できるという利点がある。
　以下のコントロール構文を持つ例文

(1)a．果汁を凍らせてデザートを作った。

の統辞解析木さえ与えられれば、意味解析を行うことによって論理意味表示（述語論理式）

(1)b．∃x1x2 x3e3e4e5(pro
　　= x3 ∧ 果汁(x1) ∧ デザート(x2)
　　∧て(せ(e4, x3, x1,
　　凍ら(e3, x1)),
　　作っ_た(e5, x3, x2)))

が自動的に出力される。ここで、「凍らせ」の主語が「作った」の主語と共通であることは、変項 x3 が両方の述語の第 2 項として出現することによって示されている。

　脳認知言語学研究に関連しては、コーパスにおける頻度に基づく語句リスト作成を行い、脳認知機能実験に応用して効果を挙げた。

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．横山悟、「入学試験区分による経時的データに基づいた大学初年次学生の英語力の分析、千葉科学大学紀要、査読無、9 巻、2016 年、9-16 ページ
2.横山悟、「Generation effect と Testing effect を援用した大学英語教育における補講の効果：項目反応理論による分析」、千葉科学大学紀要、査読無、9 巻、2016 年、17-22 ページ
3．周振・吉本啓、「中国人日本語学習者の VN 型二字漢語動詞の習得に関する研究: VN 型二字漢語動詞の一体性の視点から」、国際文化研究 21 号、99-112 ページ、2015 年 3 月 31 日、東北大学国際文化学会
4.横山悟、「大学における初年次英語教育の効果に関する多角的分析」、千葉科学大学紀要、査読無、8 巻、2015 年、17-21 ページ
5．Yosuke Hashimoto, Satoru Yokoyama, Ryuta Kawashima, “Neural Differences in Processing of Case Particles in Japanese: An fMRI Study”, Brain and Behavior, 査読有、4 巻、2014 年、180-186 ページ
6．Satoru Yokoyama, Kei Takahashi, Ryuta Kawashima, “Animacy or Case Marker Order?: Priority Information for Online Sentence Comprehension in a Head-Final Language”, PLoS ONE, 査読有、9 巻、2014 年

〔学会発表〕（計 23 件）
1．周振・Alastair Butler・吉本啓、「中国語連体修飾節構文の解析」、言語処理学会 22 回年次大会発表論文集、809-812 ページ、2016 年 3 月 10 日、東北大学（宮城県仙台市）

2．アラステア・バトラー・吉本啓・岸本秀樹・プラシャント・パルデシ、「統辞・意味解析情報付き日本語コーパスのアノテーション」、言語処理学会第 22 回年次大会発表論文集』、589-592 ページ、2016 年 3 月 9 日、東北大学(宮城県仙台市)
3.周振・Alastair Butler・吉本啓、「中国語結果構文の解析」、言語科学会第 17 回年次国際大会ハンドブック、56-59 ページ、2015 年 7 月 18 日、別府国際コンベンションセンタービーコンプラザ（大分県別府市)
4．伊藤克将・森 芳樹、「日本語のガ・ノ交替の統辞論と意味論―ドイツ語との対照を交えて」、日本言語学会、2015 年 6 月 20 日、大東文化大学 (東京都板橋区)
5．吉本啓・プラシャント・パルデシ、「文の統辞・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」、関西言語学会ワークショップ、2015 年 6 月 13 日、神戸大学（兵庫県神戸市)
6．Kei Yoshimoto and Alastair Butler, “Development of Japanese Corpus Tagged with Syntactic and Semantic Information”, The 18th Joint Workshop on Linguistics and Language Processing. Korean Society for Language and Information, 2015 年 5 月 22 日、Kyung Hee University, Seoul（韓国）
7．Alastair Butler and Kei Yoshimoto, “Large scale semantic representation with flame graphs”, 言語処理学会第 21 回年次大会発表論文集、301-304 ページ、 2015 年 3 月 17 日、京都大学（京都府京都市)
8.プラシャント・パルデシ・Alastair Butler・吉本啓・岸本秀樹、「統辞・意味解析情報付き日本語コーパスの開発」、言語処理学会第 21 回年次大会発表論文集、20-23 ページ、2015 年 3 月 17 日、京都大学(京都府京都市)
9．周振・Alastair Butler・吉本啓、「中国語意味解析コーパス構築のための句レベルのスコープアノテーション ―文の構成要素の間のコントロール関係の同定および否定の作用域の制御を中心に―」、言語処理学会第 21 回年次大会発表論文集、856-859 ページ、2015 年 3 月 19 日、京都大学（京都府京都市)
10．Alastair Butler, Shota Hiyama and Kei Yoshimoto、“Coindexed null elements for a Japanese parsed corpus”, 言語処理学会第 21 回年次大会発表論文集、708-711 ページ、2015 年 3 月 18 日、京都大学(京都府京都市)
11．Alastair Butler and Kei Yoshimoto, “Semantic Visualisation with Flame Graphs”, Proceedings of the Eleventh International Workshop of Logic and Engineering of Natural Language Semantics, 205-215 ページ, JSAI International Symposia on AI, the Japanese Society for Artificial Intelligence, 2014 年 11 月 22 日～24 日、お茶の水大学、慶応大学（東京都文京区、神奈川県横浜市)
12．Yoshiki Mori, “Where is the verum focus in Japanese?”, Satellite Forum around the European Association of Japanese Studies EAJS 2014, 2014 年 8 月 31 日, University of Ljubljana（スロベニア）
13．Yoshiki Mori and Shinya Okano, “On abductive uses of Japanese ‘hazu’”, Chronos 11, 2014 年 6 月 18 日, Scuola Normale Superiore Pisa（イタリア）
14．Alastair Butler・吉本啓、“Meaning representations from treebank annotation”, 言語処理学会第 20 回年次大会発表論文集、1023-1026 ページ、2014 年 3 月 20 日、北海道大学（北海道札幌市)
15．Alastair Butler・方采薇・檜山祥太・周振・小菅智也・吉本啓、「統辞・意味情報を付加した日本語コーパスの構築 欅ツリーバンクプロトタイプについて」、言語処理学会第 20 回年次大会発表論文集、904-907 ページ、2014 年 3 月 20 日、北海道大学(北海道札幌市)
16．檜山祥太・吉本啓・ Alastair Butler、「連体修飾節における曖昧性とその解決策の提案」、言語処理学会第 20 回年次大会発表論文集、674-677 ページ、2014 年 3 月 19 日、北海道大学(北海道札幌市)
17．周振・Alastair Butler・吉本啓、「中国語コントロール構文の解析」、言語処理学会第 20 回年次大会発表論文集、670-673 ページ、2014 年 3 月 19 日、北海道大学（北海道札幌市)
18．方采薇・Alastair Butler・吉本啓、“Parsing Japanese with a PCFG treebank grammar”, 言語処理学会第 20 回年次大会発表論文集、432-435 ページ、2014 年 3 月 18 日、北海道大学(北海道札幌市)
19．Yoshiki Mori and Hitomi Hirayama、“Nominal Semantics in the Left Periphery in German and Italian”, Incontro di Grammatica Generativa 40. Workshop on Specificity in the Grammar: Form and Interpretation, 2014 年 2 月 12 日、University of Trento（イタリア）
20．吉本啓・周振・小菅智也・大友瑠璃子・Alastair Butler、「日本語ツリーバンクのアノテーション方針」、言語処理学会第 19 回年次大会発表論文集、924-927 ページ、2013 年 3 月 15 日、名古屋大学（愛知県名古屋市)
21．周振・Alastair Butler・吉本啓、「中国語統辞解析木の形式変換及びその応用に関する研究−Penn Chinese Treebank (3.0) を対象として−」、言語処理学会第 19 回年次大会発表論文集、920-923 ページ、2013 年 3 月 15 日、名古屋大学（愛知県名古屋市)

〔図書〕（計 12 件）
1.吉本啓・中村裕昭『現代意味論入門』，くろしお出版，2016 年，259 ページ
2．Shinya Okano and Yoshiki Mori, “On CG management of Japanese weak necessity

modal 'hazu' ", In: New Frontiers in Artificial Intelligence (LNAI 9067: Post-proceeding Publication of Logic and Engineering of Natural Languages 11 (LENLS 11)), Springer, 2015年, 160-171ページ
3.横山悟、「実証的方法論により検証された最新の科学的知見「産出効果とテスト効果」に基づく効率的学習法」、総説出版、2015年、17ページ
4.横山悟、「Q&A 心理学入門」、ナカニシヤ出版、2015年、179-182ページ
5.横山悟、「英語教育学と認知心理学のクロスポイント: 小学校から大学までの英語学習を考える」、北大路書房、2015 年、84-98ページ
6. Eric McCready, Katsuhiko Yabushita and Kei Yoshimoto (eds.), Formal Approaches to Semantics and Pragmatics: Japanese and Beyond, Springer, September 2014, 374 ページ
7. Yoshiki Mori, "Die interne Struktur von subordinierten Sätzen mit ga/no Wechsel im Japanischen", In: Beiträge zur generativen Linguistik, iudicium , 2014年、32-53ページ
8. Yoshiki Mori and Chunhong Park, "On Some Aspects of the Deictic/Evidential Component in Korean -(u)l kesita and Japanese hazuda", In: Proceedings of 15th International symposium on Korean Linguistics, Harvard University, 2014年, 119-133ページ
9. Kei Yoshimoto and Masahiro Kobayashi, "Floating Quantifiers in Japanese as Adverbial Anaphora", In: Eric McCready, Katsuhiko Yabushita and Kei Yoshimoto (eds.), Formal Approaches to Semantics and Pragmatics: Japanese and Beyond, 2014年9月, 343-374ページ
10. Yoshiki Mori, "Was öffnet unsere Zukunft?" In: Modalität und Evidentialität im Deutschen, iudicium, 2013 年、68-87ページ
11. Yoshiki Mori and Shinya Okano, "Evidentialität als Inferentialität", In: Funktion(-en) von Modalität, de Gruyter, 2013年, 189-217ページ
12. Alastair Butler, Ruriko Otomo, Zhen Zhou and Kei Yoshimoto, "Treebank Annotation for Formal Semantics Research", In Y. Motomura, A. Butler and D. Bekki, eds., New Frontiers in Artificial Intelligence, Lecture Notes in Computer Science, Volume 7856, Springer , 2013年, 25-40ページ

〔産業財産権〕
○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)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

(2)研究分担者
森　芳樹（MORI, YOSHIKI）
東京大学・総合文化研究科・教授
研究者番号：30306831

横山　悟（YOKOYAMA, SATORU）
千葉科学大学・薬学部・准教授
研究者番号：20451627

(3)連携研究者

研究者番号：