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 １９ 日現在

機関番号：６２６１８
研究種目：基盤研究(B)（一般）
研究期間：2015～2019
課題番号：１５Ｈ０３２１０

研究課題名（和文）統語・意味解析情報タグ付きコーパス開発用アノテーション研究：複文を中心に

研究課題名（英文）Research on annotation for the development of a parsed corpus of Japanese with a special focus on complex sentences

研究代表者

ＰＡＲＤＥＳＨＩ　Ｐ．Ｖ．（Pardeshi, Prashant）

大学共同利用機関法人人間文化研究機構国立国語研究所・理論・対照研究領域・教授

研究者番号：００３７４９８４

交付決定額（研究期間全体）：（直接経費）　　12,800,000円

研究成果の概要（和文）：本研究は現代日本語の特徴の一つである関係節および従属節を中心とする複文について言語学的情報を検索、抽出するために必要なアノテーション方法を研究し、それに基づいてタグ付け作業を行い、複文に関するより高度の環境整備を整えることを目標とした。この作業は国立国語研究所の「統語・意味解析コーパスの開発と言語研究」プロジェクトと連携をしながら進め、2020年3月時点で、40,831万文規模のコーパス(560,098語)を検索ツールとともに公開した。

研究成果の学術的意義や社会的意義
本研究の成果として約4万文（56万語）規模の統語・意味解析付きコーパスが開発され、公開された。このコーパスでは現代日本語のテクストに対し文の統語・意味解析情報を付与し、複数な検索ツールが用意されており、多様な日本語の機能語や句構造、節の諸類型および複雑な構文を大量の言語データから検索・抽出して研究に活用することが可能である。また、初心用の検索ツール完備により、大学などで日本語の統語論の教育にも利用できる。さらに、このコーパスでは日英両言語表記で公開されており、日本語表記に精通してない海外の研究者も利用できるので、日本語研究の国際化に貢献できる。

研究成果の概要（英文）：The goal of this research project was to build a large-scale parsed corpus (treebank) of modern Japanese that would enable the users to search and retrieve complex sentences involving relative clauses and subordinate clauses, which are peculiar characteristic of Japanese, through the development of annotation method and corpus search tools. This task was jointly carried out with the collaborative research project “Development of and Linguistic Research with a Parsed Corpus of Japanese” at NINJAL. A corpus of 40,831 sentences (560,098 words) was build and is made available for free access at the following site: http://npcmj.ninjal.ac.jp/. The corpus can be searched with the following search tools: http://npcmj.ninjal.ac.jp/explorer/ http://npcmj.ninjal.ac.jp/interfaces/index_en.html

研究分野：コーパス言語学

キーワード：関係節　複文　アノテーション　統語・意味解析付きコーパス　コーパス研究

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

本研究の代表者は第二期中期計画において『現代日本語書き言葉均衡コーパス（BCCWJ）』のデータを利用して、内容語（動詞、名詞、形容詞、副詞）の振る舞い（主に共起関係）を検索・抽出するシステム NLB を開発した（http://nlb.ninjal.ac.jp/）。この検索システムでは、特に高度なコンピュータの知識がないユーザーも、簡単かつ瞬時に研究に必要な情報を検索・抽出・ダウンロードでき、ユーザー数は数万人となっている。この間、国内外の研究者・大学院生からコーパスに基づく機能語、句、節、複文といった様々なレベルでの構造体（constructions）を検索・抽出できるシステムを構築してほしいとの要望を多数受けた。また、学会などでも研究者らからその要請を受けた。

英語に関しては、90 年代初頭から Pennsylvania（以下、Penn）大学で統語解析情報タグ付きコーパスが開発され、20 年かけて意味情報も付加したコーパスが複数完成し、コーパスに基づく機能語、句、節、複文のような様々なレベルでの構造体を大量のデータから検索・抽出して研究を行うことが可能となり、目覚ましい成果をあげている。

英語の成功を受けて、世界各国では、言語研究に資するコーパスが開発されている。共通なフォーマットで構築される Penn 方式のコーパス間での対照研究も活発に行われている。一方、日本語に関しては、言語研究を目的とする汎用的統語・意味解析情報タグ付きコーパスは未だに存在せず、コーパスに基づく日本語の諸構造体の研究は英語などと比べて、立ち遅れている。

２．研究の目的

上記の問題意識を踏まえ、本研究は現代日本語における複文の一つである連体節を検索・抽出するために必要なアノテーション方法を研究し、それに基づいて約 5 万文（約 90 万語）に対してタグ付け作業を行い、連体節を簡単に検索・検出できる環境整備を整えることを目標とした。

３．研究の方法

統語解析情報タグ付きコーパス開発のためには日本語の諸構造体（constructions）を検索・検出できるためのタグ付け（アノテーション）作業が不可欠である。この基礎作業を行うためには日本語文法の専門的知識、特に統語的な側面の専門知識が必要であり、さらに、大量データに対してタグ付け作業を（半）自動的に行う必要があるためため、形式文法、コンピューター言語学および自然言語処理といった言語関連諸分野の共同研究・共同作業を行った。

具体的には、本研究は現代日本語の特徴の一つである関係節および従属節を中心とする複文について言語学的情報を検索、抽出するために必要なアノテーション方法を研究し、それに基づいて約４万文（約５５万語）に対してタグ付け作業を行い、複文に関するより高度の環境整備を整えた。なお、この作業は国立国語研究所の「統語・意味解析コーパスの開発と言語研究」プロジェクトと連携をしながら進め、国立国語研究所のホームページでコーパスを一般公開した（詳細は以下を参照）。

４．研究成果

統語解析情報タグ付きコーパス開発とそれを利用して研究を行い、以下のような、成果がありました。

● コーパスおよび検索ツールの構築と公開

NPCMJ（NINJAL Parsed Corpus of Modern Japanese）コーパス構築：研究期間を通して、毎年アノテーション作業を行い、では 2020 年 3 月現在、約 4 万文（ツリー）、語数にして約 56 万語のデータが収録されているコーパスを構築し、国立国語研究所のウェブサイトを通じて（http://npcmj.ninjal.ac.jp/）無償公開した。

データはすべて、漢字仮名混じりとローマ字の両方で表記されており、全データ、あるいは下で紹介する検索ツールを使った検索結果を自分のコンピューターにダウンロードすることが可能である。

NPCMJ コーパス検索用に専用の検索ツールの開発も行い、初中級者向けの NPCMJ Explorer および中上級者向けの NPCMJ Search を公開した。

NPCMJ Explorer は、単純な文字列検索と、特定の文法項目に関する用例の検索という 2 種類の検索手段を備えたツールであり、文法項目として、益岡隆志・田窪行則（1992）『基礎日本語文法』（くろしお出版）で解説されている 136 の項目から 73 項目を取り上げ、各項目のラベルをクリックするだけで、利用者が複雑な検索式を作成しなくても、用例が表示される仕組みになっている。ジャンル（表 1 における「出典」）ごとの頻度を調べるためのジャンル指定機能もある。

NPCMJ Search では、品詞や句・節に与えられたラベルや、ラベルとラベルの関係を検索式

によって指定しながら用例の検索が可能である。どちらのツールでも、検索結果を一覧表示させたり、木構造で表示させたりすることができる。本コーパス・検索ツールの開発によって複文に関するより高度の環境整備を整えることができた。

- 雑誌論文・学会発表

本研究の成果を国内外の学会や研究雑誌などを通じて公開した。

５．主な発表論文等

〔雑誌論文〕　計11件（うち査読付論文　0件／うち国際共著　0件／うちオープンアクセス　5件）

１．著者名 Alastair Butler	４．巻 13
２．論文標題 From meaning representations to syntactic trees	５．発行年 2016年
３．雑誌名 Proceedings of the Thirteenth International Workshop of Logic and Engineering of Natural Language Semantics13 (LENLS13)	６．最初と最後の頁 147-160
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 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Alastair Butler	４．巻 10
２．論文標題 DynamicPower at SemEval-2016 Task 8: Processing syntactic parse trees with a Dynamic Semantics core	５．発行年 2016年
３．雑誌名 Proceedings of SemEval-2016	６．最初と最後の頁 1148-1153
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 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Alastair Butler	４．巻 2
２．論文標題 Deterministic natural language generation from meaning representations for machine translation	５．発行年 2016年
３．雑誌名 Proceedings of the 2nd Workshop on Semantics-Driven Machine Translation	６．最初と最後の頁 1-9
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 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 周振・吉本啓	４．巻 21
２．論文標題 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	５．発行年 2015年
３．雑誌名 国際文化研究	６．最初と最後の頁 99-112
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 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 プラシャント・パルデシ・Alastair Butler・吉本啓 ・岸本秀樹	4．巻 21
2．論文標題 統語・意味解析情報付き日本語 コーパスの開発	5．発行年 2015年
3．雑誌名 言語処理学会第21回年次大会発表論文集	6．最初と最後の頁 20-23
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 無
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 Alastair Butler and Kei Yoshimoto	4．巻 21
2．論文標題 Large scale semantic represent ation with flame graphs	5．発行年 2015年
3．雑誌名 言語処理学会第21回年次大会発表論文集	6．最初と最後の頁 301-304
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 無
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 Alastair Butler, Shota Hiayama and Kei Yoshimoto	4．巻 21
2．論文標題 Coindexed null elements for a Japanese parsed corpus	5．発行年 2015年
3．雑誌名 言語処理学会第21回年次大会発表論文集	6．最初と最後の頁 708-711
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 無
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 周振・Alastair Butler・吉本啓	4．巻 21
2．論文標題 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－	5．発行年 2015年
3．雑誌名 言語処理学会第21回年次大会発表論文集	6．最初と最後の頁 856-859
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 無
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 周振・Alastair Butler・吉本啓	4．巻 17
2．論文標題 中国語結果構文の解析	5．発行年 2015年
3．雑誌名 言語科学会第17回年次国際大会，ハンドブック	6．最初と最後の頁 56-59
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 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 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	4．巻 22
2．論文標題 統語・意味解析情報付き日本語コーパスのアノテー ション	5．発行年 2016年
3．雑誌名 言語処理学会第22回年次大会発表論文集	6．最初と最後の頁 589-592
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 無
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 周振・Alastair Butler・吉本啓	4．巻 22
2．論文標題 中国語連体修飾節構文の解析	5．発行年 2016年
3．雑誌名 言語処理学会第22回年次大会発表論文集	6．最初と最後の頁 809-812
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 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計22件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 Yusuke Kubota
2．発表標題 Reconsidering the Coordinate Structure Constraint once again: Corpus-based evidence
3．学会等名 Conceptual and Methodological Alternatives in Theoretical Linguistics
4．発表年 2018年

１．発表者名
アラステア・バトラー，長崎郁，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

２．発表標題
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

３．学会等名
言語処理学会第24回年次大会

４．発表年
2018年

１．発表者名
Alastair Butler, Stephen Wright Horn, Iku Nagasaki

２．発表標題
Seeding lexical semantics: resources using parsed corpora

３．学会等名
NINJAL International Symposium "Exploiting Parsed Corpora: Apllication in Research, Pedagogy and Processing"

４．発表年
2017年

１．発表者名
Hideki Kishimoto, Prashant Pardeshi

２．発表標題
Parsed corpus as a source for testing generalizations in Japanese syntax.

３．学会等名
NINJAL International Symposium "Exploiting Parsed Corpora: Apllication in Research, Pedagogy and Processing"

４．発表年
2017年

１．発表者名
Yusuke Kubota, Ai Kubota

２．発表標題
A case study on the Coordinate Structure Constraint in Japanese

３．学会等名
NINJAL International Symposium "Exploiting Parsed Corpora: Apllication in Research, Pedagogy and Processing"

４．発表年
2017年

１．発表者名
Kei Yoshimoto, Akiko Takahashi

２．発表標題
Exploiting coreferential information in NPCMJ for L2 reading of Japanese texts

３．学会等名
NINJAL International Symposium "Exploiting Parsed Corpora: Apllication in Research, Pedagogy and Processing"

４．発表年
2017年

１．発表者名
Alastair Butler, Stephen Wright Horn

２．発表標題
Treebank Semantics parsed corpus series

３．学会等名
NINJAL International Symposium "Exploiting Parsed Corpora: Apllication in Research, Pedagogy and Processing"

４．発表年
2017年

１．発表者名
窪田悠介

２．発表標題
「ツリーバンク検索への「UNIX 的」アプローチ」

３．学会等名
国語研究所言語資源活用ワークショップ

４．発表年
2017年

１．発表者名
Alastair Butler, Ai Kubota, Shota Hiyama and Kei Yoshimoto

２．発表標題
Treebank annotation of FraCaS and JSeM

３．学会等名
Logic and Engineering of Natural Language Semantics (LENLS 13)

４．発表年
2016年

１．発表者名
Alastair Butler

２．発表標題
From meaning representations to syntactic trees

３．学会等名
Logic and Engineering of Natural Language Semantics (LENLS 13)

４．発表年
2016年

１．発表者名
Alastair Butler

２．発表標題
Deterministic natural language generation from meaning representations for machine translation

３．学会等名
2nd Workshop on Semantics-Driven Machine Translation

４．発表年
2016年

１．発表者名
プラシャント・バルデシ

２．発表標題
ワークショップ「イントロダクション」統語・意味解析情報付き日本語コーパスの構築に向けて

３．学会等名
日本言語学会第153回大会

４．発表年
2016年

１．発表者名
プラシャント・バルデシ

２．発表標題
ワークショップ「まとめと将来の展望」統語・意味解析情報付き日本語コーパスの構築に向けて

３．学会等名
日本言語学会第153回大会

４．発表年
2016年

１．発表者名
吉本啓

２．発表標題
ワークショップ「アノテーション方式とコーパスの特色」統語・意味解析情報付き日本語コーパスの構築に向けて

３．学会等名
日本言語学会第153回大会

４．発表年
2016年

１．発表者名
アラステア・バトラー、窪田愛、窪田悠介

２．発表標題
ワークショップ「デモンストレーション」統語・意味解析情報付き日本語コーパスの構築に向けて

３．学会等名
日本言語学会第153回大会

４．発表年
2016年

１．発表者名
Alastair Butler

２．発表標題
Parsed Corpus Semantics

３．学会等名
New Landscapes in Theoretical Computational Linguistics

４．発表年
2016年

１．発表者名
Alastair Butler, Shiro Akasegawa, Prashant Pardeshi and Kei Yoshimoto

２．発表標題
A parsed corpus of Japanese enriched to reach levels of semantic analysis

３．学会等名
Brandeis University, Boston, USA Colloquium

４．発表年
2016年

１．発表者名
窪田悠介

２．発表標題
形式意味論と計算言語学の最近の動向

３．学会等名
第13回東海意味論研究会

４．発表年
2016年

１．発表者名
吉本啓

２．発表標題
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

３．学会等名
台湾日本語言文藝研究学会第15回定例学会

４．発表年
2016年

１．発表者名
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

２．発表標題
統語・意味解析情報付き日本語コーパスのアノテーション

３．学会等名
言語処理学会 第22回年次大会

４．発表年
2016年

１．発表者名
吉本啓・プラシャント・パルデシ

２．発表標題
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

３．学会等名
関西言語学会ワークショップ

４．発表年
2015年

１．発表者名 Kei Yoshimoto and Alastair Butler
２．発表標題 Development of Japanese Corpus Tagged with Syntactic and Semantic In formation
３．学会等名 The 18th Joint Workshop on Linguistics and Language Processing. Korean Society for Language and Information. Kyung Hee University, Seoul
４．発表年 2015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
研究分担者	岸本　秀樹 (Kishimoto Hideki) (10234220)	神戸大学・人文学研究科・教授 (14501)	
研究分担者	野田　尚史 (Noda Hisashi) (20144545)	大学共同利用機関法人人間文化研究機構国立国語研究所・日本語教育研究領域・教授 (62618)	
研究分担者	吉本　啓 (Yoshimoto Kei) (50282017)	東北大学・高度教養教育・学生支援機構・教授 (11301)	
研究分担者	窪田　悠介 (Kubota Yusuke) (60745149)	大学共同利用機関法人人間文化研究機構国立国語研究所・理論・対照研究領域・准教授 (62618)	
研究分担者	長崎　郁 (Nagasaki Iku) (70401445)	大学共同利用機関法人人間文化研究機構国立国語研究所・理論・対照研究領域・プロジェクト非常勤研究員 (62618)	

６．研究組織（つづき）

	氏名 （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	バトラー　アラステア (Butler Alastair) (90588873)	弘前大学・人文社会科学部・准教授 (11101)	
研究分担者	ＨＯＲＮ　Ｓ．Ｗ. (Horn Stephen Wright) (70801538)	大学共同利用機関法人人間文化研究機構国立国語研究所・理論・対照研究領域・プロジェクト非常勤研究員 (62618)	
研究分担者	影山　太郎 (Kageyama Taro) (80068288)	大学共同利用機関法人人間文化研究機構国立国語研究所・理論・対照研究領域・その他 (62618)	