様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成 24 年　5 月　9 日現在

機関番号：17102
研究種目：基盤研究(C)
研究期間：2008～2011
課題番号：20520504
研究課題名（和文）　英文法コーパスの構築とその応用

研究課題名（英文）　Constructing School Grammar Corpus of English and its Application

研究代表者
徳見　道夫　（TOKUMI MICHIO）
九州大学・大学院言語文化研究院・教授
研究者番号：90099755

研究成果の概要（和文）：
本研究は，日本人英文学習者にとって英文理解で最も重要な観点の一つである学校文法に関する情報を付与した実用的な英文データ（広義の学校英文法コーパス）を，学習参考書に掲載されている例文を中心に約1万文分，整備した．その英文データの集積過程と並行して半自動的に当該情報を検出するようなルールを蓄積・活用し，作業の効率化を図った．また，これらを他課題と連携し，応用研究にも一部活用し，その有用性を確認した．

研究成果の概要（英文）：
Our research has provided about 10,000 English sentences together with specific grammatical rules for reference books used by learners of English. This has proved to be essential for Japanese learners trying to understand English texts. We have also designed and used a mechanism that retrieves almost automatically referent information beside the list of selected English sentences. Moreover, we have amongst other tasks successfully put the data's usability to the test through further applied research.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	600,000	180,000	780,000
2009 年度	900,000	270,000	1170,000
2010 年度	1,100,000	330,000	1430,000
2011 年度	700,000	210,000	910,000
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：人文学
科研費の分科・細目：言語学・外国語教育
キーワード：コーパス，学校英文法，機械学習，自然言語処理

１．研究開始当初の背景

近年，大量の電子化用例集（広義のコーパス）が言語研究で重要な役割を果たしている．このコーパスには平文だけが大量に集積された生テキストレベルのものから，単語認定がなされ，品詞や原形といった形態素情報が付与された形態素レベル，文の構文構造（句

構造文法などの形式文法が仮定され，それによって記述される構造）が与えられた構文レベル，さらには単語の語義や文の意味などが記述された意味レベルの情報が付与されたものまで様々である．このような状況のなかで主に品詞～構文レベル，ときに意味レベルまで含むような英語の学校文法（学校英文法）に関する情報が付与された研究等に利用できる英語コーパスはない．学校英文法は一般の英語教員や，英語を母語としない日本の英語学習者にとっては，句構造文法といった形式文法よりも身近であり，英文等の理解の重要な切り口の一つである．本研究の各推進者の研究やその周縁的な研究領域においても，このような情報が付与されたデータに対する潜在的なニーズは高かった．

２．研究の目的

前節であげた背景を受け，本研究では主に次のような３つの目的を置いた．

(1) 学校英文法に関する情報を付与した研究利用可能なコーパス（学校英文法コーパス）を構築すること

(2) 英文に含まれる文法項目を検出するようなルール（文法項目の自動検出ルール）を並行して蓄積し，前述の作業を効率化すること

(3) 学校英文法コーパスや検出ルールを用いた応用研究を展開すること

(1)は具体的なデータ構築で，(2)はそれを援用する手段の蓄積である．(1)の初期段階では(2)のルールは粗く，(1)への貢献は薄いことが予想される．しかし，データの充実化にしたがってそれらが精密化され，ひとたび(2)が一定の精度に達すれば(1)のデータ拡充のペースが上がり，(1),(2)の好循環が整う．また，ある程度の量の(1)や精度の(2)が整えば，(3)にあげた学校英文法の項目に関する統計的分析や教材評価といった応用研究へ活用することが可能となる．それらの応用研究から，学校英文法コーパスの不備，再設計すべき方向性が示唆されることも期待される．このように，これらの３つの目的は，互いに相乗的に進展する仕掛けとなっている．本研究はコーパス構築の新しい方法論の提案でもある．

なお，本研究では上記の英文データを，従前，「（学校）英文法コーパス」と呼んでいた．研究開始後，狭義のコーパスを考えた場合にはこの名称が不適当であることが懸念された．そこで，「学校文法情報付き英文データ」あるいは集積した英文が学習参考書上の例文（学参例文）であった際には「学校英文法の学参例文データベース」と名称を変更している．ただし，この報告書に限っては，申請書との対応と広義のコーパスという還観点から，これらのデータを申請時に使用した「学校英文法コーパス」と記すこととする．

３．研究の方法

前項目であげた研究の目的に合わせて，本研究の方法について述べる．

(1) 学校英文法コーパスの整備

初年度に学習指導要領・検定教科書・参考書および従来研究を精査し，学校英文法コーパスの基本的な設計（コーパス・デザイン）を行った．文法項目の種類等を表１にあげる．

文法項目の情報付与の粒度については，単文・節単位と文単位の二方式を検討した．前者の方式の方が，情報量が高く正確ではある．しかし，特にオーセンティックな文を対象とした場合には，試行結果から情報付与の作業がかなり煩雑となることが見込まれたため，2 年目以降は文単位の方式に加え，動詞部分を細分化した形で作業を進めている．

情報を付与する対象としては，主に次の二つの資料とした．一つは，代表的な構文解析済み英文コーパスの一つである Penn Treebank の一部（Brown Corpus 部分）で，比較的広い構文情報によって同定される文法項目（たとえば，仮定法や分詞構文など）を対象とした．なお，この Penn Treebank の一部については，単文・節単位で情報を付与している．もう一つは，主に高校～大学生レベルの英語参考書の例文（学参例文）である．

文法項目	下位項目
文型	1-5 文型
文の種類	平叙・疑問・命令・感嘆
疑問文の種類	一般・特殊・選択・間接・付加
否定	全否定・部分否定
時制	未来・現在・過去
態	能動・受動
法	直接・仮定（・命令）
相	進行・完了
話法	直接・間接
to 不定詞	名詞的・形容詞的・副詞的
原形不定詞	名詞的・形容詞的・副詞的
形容詞	原級・比較級・最上級
副詞	原級・比較級・最上級
同等比較	
分詞	現在・過去
動名詞	
助動詞	
疑問詞	
接続詞	等位・従属
関係詞	代名詞（主格/目的格/所有格）・副詞
数量表現	
倒置	
比較級+比較級構文	
存在 there 構文	
分詞構文	

表 1 文法項目

学参例文は，配置されている章節で解説された文法項目が顕在化するよう単純化されており，一定の英語知識を備えた者であれば，問題なく作業できるという利点がある．

また，蓄積したこれらのデータの公開方法についても合わせて検討を行った．

(2) 文法項目の検出ルールの蓄積

研究推進者らが強くかかわる他課題とも連動し，(1)の作業の効率化を念頭とした文法項目の検出ルールの整備を進めた．単に人手で記述するだけでなく，機械学習を活用した自動的な記述なども試みた．それぞれの文法項目で，検出の精度を見積もり，本課題の情報付与作業へ還元した．なお，(1)で選択した学参例文は，この文法項目の検出において，対象とする文法項目とは無関係な情報が含まれにくいといった点で，非常に都合が良かった．

(3) 応用研究

応用研究については枚挙すれば限りがないため，特に研究推進者らが従事している他課題での応用を優先した．主に英語科学論文の表現上の質に関わるデータに対して，(2)のルールを活用し，質間の使用頻度傾向の基本的な調査，他の言語特徴に比した質判定への寄与を検討した．

４．研究成果

(1) 学校英文法コーパス

文単位で文法情報を付与した学参例文は，表 2 の通りである．このなかで「その他（7 冊)」はダブルチェックを経ていない，あるいは学習参考書内の例文をすべて電子化できていないものである．その他を除く 9 千文のうち完全な文法情報（不正確という意味ではなく，文中の全ての文法情報の使用の是非が付与されている，という意）が付与されているものが 4 千文弱である．

また，Penn Treebank を対象としたものについては，5 千文弱に対して仮定法や分詞構文などの比較的広い構文情報に関わる情報が付与された．

データについては，著作権上の問題から英文そのものは省き，付与した文法情報のみの公開とした．ただし，英文部分を適当な形式で準備した際に，公開する文法情報とマージするプログラムを別途作成した．

(2) 文法項目の検出ルール

検出ルールのなかでも精度が高いものについては，(1)の作業に活用した．作業効率の向上等について客観的な検討を付していない．しかしながら，文法項目の付与作業において，簡易とはいえ一から情報を付与していくのと，多少誤りが含まれる可能性はあるものの推定された値が提示され，作業者はそれを確認するのみで良い，というのは全体の作業量を勘案すれば負荷軽減の効果は小さくないと考えられる．

(3) 応用研究

表現上の質情報が付与された英語科学論文に対して，(1)や(2)の成果を学校英文法の観点から基礎的調査を幾通りか試みた．表現上の質が良い論文と，日本語を母語とする著者が中心として書いたあまり表現上の質が高いとはいえないものを比較した．このような場合には，日本人著者が中心となった表現上あまり良質とは言いがたい論文では，従来から指摘されているような「現在時制」「受動態」が過剰使用である，「比較級」「最上級」「仮定法」などの使用が少ない，といったことが半自動的に明らかになった．また，質判定という問題で考えた際には，このような学校文法上の項目の多寡よりは，談話標識や n-gram といったものの方が相当に優位で，おおむね予想通りの結果であった．

なお，上記の研究を推進するにあたっては，後述する研究組織にあげた研究分担者・連携研究者の他に，研究協力者として小林雄一郎氏（大阪大学大学院・言語文化研究科・院生／日本学術振興会・特別研究員）も重要な役割を果たしたことを申し添えておく．

学習参考書	文数
ロイヤル英文法	2,165
depth 英語総合	1,925
必修英文法問題精講	1,914
プリズム総合英語	1,458
チャート式現代英文法	1,450
基礎英文法問題精講	1,042
チャート式ラーナーズ高校英語	815
その他（7 冊）	2,886
計	1,1730

表 2　電子化した学参例文の内訳

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 13 件）

① Kitao, K. and Tanaka, S.: Characteristics of Japanese Junior High School English Textbooks, 文化情報学, 第 4 巻, pp. 1-14 (2009)

② 神谷健一, 田中省作, 北尾謙治: 言語

処理技術と教材作成の連携，自然言語処理，第 16 巻，第 2 号，pp.45-58 (2009)

③ 田中省作，小山由紀江：構文情報を考慮したESPコーパスからの特徴表現の抽出，統計数理研究所共同研究リポート，No.239，pp.13-30（2009）

④ 徳見道夫：『ヘンリー六世』三部作における女性の登場人物について -家父長制への揺さぶりと再建-，英語英文学論叢，第61巻，1-15（2011）

⑤ 徳見道夫，田中省作：英文法コーパス構築の有用性，言語科学，第 46 巻，pp.61-74（2011）

⑥ 徳見道夫：『エドワード三世』の作者について -歴史劇における三つの親子関係からの推測-，言語科学，第 46 巻，pp.47-59（2011）

⑦ 田中省作，柴田雅博，冨浦洋一：Web を源とした質情報付き英語科学論文コーパスの構築法，英語コーパス研究，第18号，pp.61-71（2011）

⑧ 徳見道夫，田中省作：標準化テストと九州大学における英語教育，大学教育，第16巻，pp.93-108（2011）

⑨ 徳見道夫：『エドワード三世』と『ヘンリー五世』－『エドワード三世』の作者の推測－，言語文化論究，第 27 巻，pp.17-30（2011）

⑩ 徳見道夫：翻訳『エドワード四世』第一部，言語文化論究，第 28 巻，pp.233-309（2012）

⑪ 徳見道夫，田中省作：大学入試センターのリスニングテストと九州大学の標準化テスト成績の連関性，英語英文学論叢，第62巻，pp.19-37（2012）

⑫ 小林雄一郎，田中省作，冨浦洋一：N-gram を素性とするパターン認識を用いた英語科学論文の質判定，情報処理学会研究報告自然言語処理，No.2012-NL-205，pp.1-6（2012）

⑬ 田中省作，小林雄一郎，徳見道夫，後藤一章，冨浦洋一，柴田雅博：学校英文法の学参例文データベースとその応用，情報処理学会研究報告人文科学とコンピュータ，No.2012-CH-93，Vol.5，pp.1-8（2012）

〔学会発表〕（計 10 件）

① 田中省作，小林雄一郎，徳見道夫，朝尾幸次郎：学校英文法コーパス構築の試み，人工知能学会第 22 回全国大会（2008.6.12，ときわ市民ホール・北海道）

② 小林雄一郎，田中省作，後藤一章，徳見道夫，朝尾幸次郎：文法情報の自動検出技術を用いたリーディング教材の作成と評価，語彙研究フォーラム 2008 第1回 JACET リーディング研究会・英語語彙研究会合同研究大会（2008.12.6，関西学院大学・兵庫県）

③ 田中省作，小山由紀江：日本の英語教科書コーパスを基準としたESP特徴表現の抽出，外国語教育メディア学会第 49 回全国研究大会（2009.8.6，流通科学大学・兵庫県）

④ Kitao, K. and Tanaka, S.: Authorized Junior High School English Textbooks in Japan: a Corpus-based Study of Vocabulary Level and Readability, EuroCALL 2009 (2009.9.11, Universidad Politécenica de Valencia・Spain)

⑤ 田中省作，冨浦洋一，安東奈穂子，柴田雅博：Web を源とした英語科学論文コーパスの構築法 -技術的方法論と法的観点からの検討-，英語コーパス学会第 34 回大会（2009.10.3，青山学院大学・東京都）

⑥ 後藤一章：多変量アプローチに基づく BNC における名詞の統合構造の分析，言語研究と統計 2010（2010.3.28，統計数理研究所・東京都）

⑦ 池本孝徳，宮崎佳典，田中省作：n-gram と科学英語の特徴表現を活用した例文提示に基づいた英作文支援ツール，NLP 若手の会第 5 回シンポジウム（2010.9.16，国立情報学研究所・東京都）

⑧ 田中省作，冨浦洋一，徳見道夫：学校文法に基づいた英文解析による言語データの頻度分析，英語コーパス学会第 37 回大会（2011.10.1，京都外国語大学・京都府）

〔産業財産権〕

○出願状況（計０件）

○取得状況（計０件）

〔その他〕
http://www5a.biglobe.ne.jp/~tokumi/

６．研究組織
(1)研究代表者
徳見　道夫（TOKUMI MICHIO）
九州大学・大学院言語文化研究院・教授
研究者番号：90099755

(2)研究分担者
朝尾　幸次郎（ASAO KOJIRO）
立命館大学・文学部・教授
研究者番号：40102462

冨浦　洋一（TOMIURA YOICHI）
九州大学・大学院システム情報科学研究院・教授
研究者番号：10217523

田中　省作（TANAKA SHOSAKU）
立命館大学・文学部・准教授
研究者番号：00325549

(3)連携研究者
後藤　一章（GOTO KAZUAKI）
摂南大学・外国語学部・講師
研究者番号：90397662