Vol.2012-CH-93 No.5
2012/1/27

学校英文法の学参例文データベースとその応用: 日本人英語科学論文における文法項目の使用傾向

田中省作[†1]　小林雄一郎[†2]　徳見道夫[†3]
後藤一章[†4]　冨浦洋一[†5]　柴田雅博[†5]

英語の学校文法（学校英文法）は，非英語母語話者にとって重要な英語理解の観点である．それにもかかわらず，著者らの知り得る限りでは，このような学校英文法に関する情報を詳細に付与したような英文データはない．そこで，まず日本の学校英文法に対する学習参考書の例文（学参例文）を電子化し，それと同時に学校文法の項目検出ルールを整備した．この応用研究の一つとして，検出ルールを活用し，日本人科学論文における学校文法上の使用傾向を試験的に分析した．その結果，従来から指摘されていた文法項目の使用傾向が確認された．

Database of Sentence Examples in a Reference Book for the English School Grammar: Characteristics of English Scientific Papers Written by Japanese

Shosaku Tanaka ,[†1] Yuichiro Kobayashi ,[†2]
Michio Tokumi ,[†3] Kazuaki Goto ,[†4]
Yoichi Tomiura [†5] and Masahiro Shibata [†5]

Although the English school grammar is one of the most important perspectives of understanding the language for nonnative English speakers, there are as yet no English data that annotate the structure of school grammar. As a first step, the project collects and digitizes sentences from grammar reference books and annotates the sentences with grammatical information. Furthermore, rules for identifying elements of school grammar in sentences are described in the data. This paper includes a pilot study that applied these rules to analyze papers written by Japanese. The study yields almost the same results as previous studies with respect to some grammatical elements.

1. はじめに

近年，さまざまな言語でコーパスなどの電子化大規模用例集が整備されつつある．特に英語は言語資源がもっとも充実した言語であり，単語・構文から意味にわたって多様な情報を付与したコーパスが構築，公開されている．しかし，そのような英語にあっても，我々が知り得る限りでは，英語の学校文法（学校英文法）に関する情報が付与されたデータはない[⋆1]．一般の英語学習者や英語教員にとっては，句構造文法といった形式文法よりも学校文法の方が身近であり，このような英文データのニーズは極めて高い．そこで我々は，現在，学校文法に関する情報が付与されたコーパス（学校英文法コーパス）の構築を念頭に，既存の各種情報処理技術も活用したデータ構築や，これらの応用を模索している[5),6),24)]．

本研究では，まず高校生を主たる対象とした学校文法に対する学習参考書の例文（学参例文）を 24) にならい電子化（学校英文法の学参例文データベース）し，それをもとに学校文法上の文法項目を検出するルールを記述した[⋆2]．これらの検出ルールは，データの拡充と再度のルール精密化という具合に，上記作業へ循環的に寄与するものである．

また，これらの文法項目の検出ルールを活用した応用研究も検討している．現在，コーパスなどの分析に利用される言語の自動解析のほとんどは，データに付与される情報同様，形式文法がベースである．本研究の検出ルールを活用すれば，英文章内の学校文法の項目を直接観測することができ，一般の英語学習者や英語教員にとっても可読性の高い知識となることが期待される．本稿では，このような学参例文データベースの応用研究の一つとして，

†1 立命館大学文学部
College of Letters, Ritsumeikan University
†2 大阪大学大学院言語文化研究科 / 日本学術振興会
Graduate School of Language and Culture, University of Osaka / Japan Society for the Promotion Science
†3 九州大学大学院言語文化研究院
Faculty of Language and Cultures, Kyushu University
†4 摂南大学外国語学部
Faculty of Foreign Studies, Setsunan University
†5 九州大学大学院システム情報科学研究院
Faculty of Information Science and Electrical Engineering, Kyushu University

⋆1 辞書出版社などで，学校文法に類する情報が付与されたコーパスが構築されているものもあるようだが，残念ながら研究等には利用することはできない．

⋆2 以降，単に「文法項目」と記す場合は，学校文法の文法項目を指すこととし，そのためのルールを「文法項目の検出ルール」もしくは簡単に「検出ルール」と呼ぶ．

日本人英語科学論文の特徴分析へ試験的に活用した事例を示す．

2. 関 連 研 究

学校文法と言語処理・コーパスが関わる研究には，15) や 26) がある．15) は学校文法項目について中高の英語教科書や市販の文法書を極めて詳細に分析し，それらの難易度に関する順序関係，教材の難易度計算の枠組みを提案している．26) では 15) を受け，1,320 の文法項目を設定し，コーパスから用例を抽出するための検索式を，項目ごとに表層・品詞列レベルで記述している．それらを実装したシステムは，British National Corpus から任意の文法項目を含んだ用例を得ることができる画期的なものである．しかし，これはあくまでも用例抽出を主目的としているもので，表層・品詞レベルの記述力の限界や，正確な精度保証がなされていないという点では，本研究が最終的に意図している学校英文法コーパスに替わるものではない．こういった用例抽出の精度を保証する，という意味でも学校文法の情報が付与されたデータの必要性は高い．

3. 学校英文法の学参例文データベース

本節では，学校英文法に関する情報が付与された英文データを蓄積するにあたり，まず文法項目と付与単位，そしてその作業の現状を大まかに述べる．

3.1 文 法 項 目

文法項目については，網羅的に設定するのではなく，15) をベースに日本人英語学習者の英文理解に強く関わるであろう項目を優先し，設定した．今回対象とした文法項目を表 1 に示す．今後もこの文法項目については，継続的に議論と改訂を行う．

現在，我々が進めているデータ構築において，このような文法項目を付与する言語単位は，主に次の 2 通りがある．

(1) **単文・節単位**

文を単文・節に分解し，その上で文の主要素（S, V, O, C）やその他の修飾部分（M）に区分し，文法項目を付与する．たとえば，

If I were a superman, I could help you. (1)

では，

$$\begin{aligned} I_0: &\ [\boxed{I_1}]_M,\ [\mathrm{I}]_S\ [\text{could help}]_V\ [\text{you}]_O. \\ I_1: &\ [\mathrm{If}]_M\ [\mathrm{I}]_S\ [\text{were}]_V\ [\text{a superman}]_C \end{aligned} \quad (2)$$

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

と分解し，(1) に加え I_0, I_1 も作業者に提示され，文全体とその主要素ごとに文法項目が付与される．表 1 の文法項目は，この単位での情報付与を想定したものである．

(2) **文単位**

単文・複文等の区別はせず，その文中に含まれる文法項目を一括して付与する．さきの (1) であれば，節境界等の区別はなしに，「第 2 文型」「第 3 文型」「仮定法」といった情報が振られることになる．情報付与の作業は比較的簡便化されるものの，このように単純化すると，状況を正確に復元できないという問題もある．なお，動詞部分についてはそれぞれの箇所で，態や相・法などの区別が必要となることが考えられるため，その出現箇所ごとに区別して文法項目を振っている．

24) では，Penn Treebank[13] の Brown Corpus 部分からランダムに抽出した約 5,000 文，中高英語教科書中の英文・ロイヤル英文法[28] 中の一部例文約 2,000 文に対して単文・節単

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

位で情報を付与している．

3.2 学参例文データベース

一般に Penn Treebank の Brown Corpus 部分のようにオーセンティックな文は，教科書や学習参考書の文に比べ，意味も構造も複雑で，初期段階での単文・節単位での付与作業はかなり煩雑なものとなる．そこで，実際に使用された英文に対する単文・節単位で情報を付与する前に，比較的単純な学参例文への文単位の情報付与を先導することとした．

電子化の対象を，このように文単位で情報付与することには，次のような利点がある．

- 英文法に関する学参例文は，それが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，文が単純化されている．
- 文が単純なため，参考書で解説されている文法項目については，作業者が比較的容易に情報付与できる．
- 次節で述べる文法項目の検出ルールの記述でも，文が単純で標的とする文法項目とは無関係な情報が少ないため，効率よく実践される．

実作業では，最初から網羅的に全ての文法項目に関する情報を付与するのではなく，例文とそれが配置されている章節で解説されている文法項目を，ダブルチェック体制を組み優先的に付与していった．その学参例文データベースの現状を，表 2 に示す．なお，表内のその他 7 冊は，対象とした学習参考書内の全例文の電子化，もしくはダブルチェックを終えていないものである．

4. 文法項目の検出ルール

4.1 ルールの記述方針

学校文法の文法項目には，単語/品詞列・構文木の一部などの言語特徴から一意に同定できるものも多い．そこで，英文に人手で文法項目に関する情報を付与しつつ，並行して既述情報から文法項目を対応づける検出ルールを記述する．たとえば，(1) に対して仮定法の検出ルールは「+1（検出する/使用/含む）」を，未来時制の検出ルールは「−1（検出しない/未使用/含まない）」を返す．

この検出ルールを整備する目的の一つは，文法項目の情報付与支援である[⋆1]．初期は粗い検出ルールなので，データ構築作業の効率化にはさほど寄与しないことが予想される．しかし，データの充実化に伴って検出ルールが精密化されれば，項目の自動付与の精度も向上し，データ拡充のペースが上がる．その結果，再び検出ルールの精密化がさらに進むことが期待される．この相互作用を繰り返すことで，作業の効率化が図られる．また，4.4 で述べるように，各検出ルールは整備された英文データで逐次精度保証することを前提としており，教材評価等の応用研究への適用可能性も判断しやすくなる[⋆2]．

このような検出ルールを記述する際，英文のどの言語特徴に着目するかは，文法項目それぞれの性質に応じて変わる．本研究では，英文に形態素解析・浅い構文解析（チャンキング）・深い構文解析の言語解析を施し，これらを必要に応じて加工し，活用する．(1) であれば，表 3 のような情報が判断材料の元となる．ただし，形態素情報は TreeTagger[27] による品詞解析の結果，浅い構文情報は TreeTagger のチャンキングの結果を一部加工したもの，深い構文情報は Charniak Parser[3] の解析結果である．検出ルールの記述は，人手による方法，人手で前処理をした上で機械学習を活用する方法[23]，さらに機械学習の結果から改めて人手でまとめ上げる方法など，多角的に試みている[⋆3]．

4.2 人手による記述

態や相のように文法項目を同定する言語特徴を作業者が容易に想起可能であるような場合には，人手で文法項目の検出ルールを記述する．たとえば,「受動態」は次のような形態素列を含んでいるかどうかで，比較的高い精度で検出ができる．

$$\text{*\_VB*\_*} \quad (\text{*\_RB*\_*}) \quad \text{*\_V*N\_*} \tag{3}$$

ただし，‘*’ はワイルドカード，‘(α)’ は α は随意的な要素であることを表す．実際には随意

⋆1 英文に対して一から各文法項目の情報を付与するよりは，多少精度が低くとも，事前に文法情報が付与され，それをチェックする，という作業の方が負荷軽減される．

⋆2 検出精度は，英文章の性質によって大きく変化することが予想される．検出ルールを利用する際には，適用する英文章で再度精度を見積もり直す必要がある．ここで付与された精度は，クローズドテストのような甘い見積もりであるものの，この段で不十分であれば他種の英文章でも十分な精度は期待できず，ルール使用/不使用の判断材料にはなる．

⋆3 この詳細については，紙面の都合上別稿に改める．

形態素情報	IF_IN_if I_PP_I were_VBD_be a_DT_a superman_NN_superman ,_,_, I_PP_I could_MD_could help_VV_help you_PP_you ._SENT_.
浅い構文情報	(S if_IN (NC I_PP) (VC be_VBD) (NC a_DT superman_NN) ,_, (NC I_PP) (VC could_MD help_VV) (NC you_PP) ._.)
深い構文情報	(S1 (S (SBAR (IN If) (S (NP (PRP I)) (VP (AUX were) (NP (DT a) (NNP superman))))) (, ,) (NP (PRP I)) (VP (MD could) (VP (VB help) (NP (PRP you)))) (. .)))

表 3 “If I were a superman, I could help you.” の各種情報

的な要素の数や形態素間の距離（語数）なども規定することができ，15) と同等の記述が可能である．

また，上述したような構文情報を活用し，部分構文木（部分木）で文法項目を検出することもできる．たとえば，「受動態」を深い構文情報で見直すと，次のような部分木を含むかどうかで判断される．

$$\text{VP AUX} \uparrow \text{VP VBN} \tag{4}$$

なお，この部分木は 9) の記法にならっており，部分木を先順走査した際のノードのラベルを表し，“↑” は親ノードに上がることを意味するメタ記号である．

4.3 機械学習を活用した記述

文法項目の検出は，構文木から当該の文法項目を含む構文木集合と，そうでない構文木集合への分類問題と考えることもできる．そこで，本研究では作業者の内省では着目すべき言語特徴が分からない，もしくは網羅性に不安があるような場合には，機械学習を積極的に活用している．

4.3.1 部分木を素性とした分類

構文木のようなラベル付き順序木の分類問題に対して，部分木を素性とする決定株と，その決定株を弱学習器としたブースティングによって分類器を構成する手法が提案されている[9]．決定株は入力データのクラスを，1 つの素性の有無によって決定する単純な分類器である．ここで素性としてラベル付き順序木を考え，素性の木 x, t とクラス $y \in \{+1, -1\}$ の決定株 h を次のように定義する．

$$h_{\langle t,y\rangle}(x) \triangleq \begin{cases} y & t \subseteq x \\ -y & \text{otherwise} \end{cases} \tag{5}$$

ここで $t \subseteq x$ は，t が x の部分木であることを表している．$\langle t, y\rangle$ は決定株のパラメタで，学習データに対する誤分類率を最小にするように推定される．

決定株は単純な分類器で，通常その分類精度は高くない．そこで，この決定株を弱学習器としたブースティングを適用する．それまでに構成された分類器では分類が難しいデータを中心に学習した弱学習器が逐次生成され，データに対するクラスはこれらの重み付き多数決によって決定される．その結果，x のクラスは，

$$\text{sgn}\left(\sum_{t,y} \alpha_{\langle t,y\rangle} h_{\langle t,y\rangle}(x)\right) \tag{6}$$

と決定される．ここで，$\alpha_{\langle t,y\rangle}$ は，$h_{\langle t,y\rangle}(x)$ に対する重みである．この分類器の利点の一つは，どのような素性が有効に働いているかを容易に確認できることである．したがって，当該の文法項目を上手くとらえているかどうかを，素性（部分木）という点からも検証しやすい．また，データ量が十分にあれば，従来想定されていなかった当該文法項目の構文的特徴などの発見も期待される．

4.3.2 英文の前処理

この方法で人がかかわるべき作業は，主に次の 2 点である．

- 文法項目に応じた適切な情報レベルの選定
- 文法項目に無関係な単語等の除去

例として仮定法の検出を考えてみる．仮定法は文内で広範囲に関連する表現が現れるため，英文の浅い構文情報を素性の元とする．学習参考書では “if”, “would” などの助動詞，“wish” などが，これに関わる重要な表現となっている．よって，このような表現以外については事前に除去し，BACT にとっての英文の素性として考える．その結果，(1) の浅い構文情報は次のようになる．

$$\begin{aligned}&\text{(S if\_IN (NC *\_PP) (VC *\_VBD) (NC *\_DT *\_NN) ,\_,}\\ &\text{(NC *\_PP) (VC could\_MD *\_VV) (NC *\_PP) .\_.)}\end{aligned} \tag{7}$$

個別の文意に関わる情報が，ほとんど落ちていることが分かる．このようにすることで，少量のデータでも仮定法の検出にかかわる素性を早い段階で得，効率よく分類器を構成することができる．

4.3.3 検出ルールの例

BACT を活用した仮定法の検出ルールの一例を示す．学参例文データベースにおいて，仮定法が使用されている例文とそうでない例文を分け，その自動分類問題として BACT による分類器を構成した．4.3.2 節で示したように英文の浅い構文情報で，参考書から仮定法に

順位	素性（部分木）
1	S NC PP ↑ ↑ VC MD should ↑ ↑ VV ↑ ↑ ADVC RB ↑ ↑ SENT
2	S NC NP ↑ ↑ SENT
3	S ADVC RB ↑ ↑ NC DT ↑ ↑ SENT
4	wish
5	if
6	S ADVC RB ↑ ↑ NC DT
7	S IN if
8	S VC VB ↑ ↑ SENT
9	VB
10	S NC ↑ VC VV ↑ ↑ NC ↑ PC NC ↑ ↑ SENT

表 4 仮定法の分類器に関する重み 10 位の素性（部分木）

かかわり得る表現である “if”,“wish” や助動詞以外の単語は除去したものを素性として考える．

このようにして構築した分類器において，仮定法を主張する重み上位 10 位までの素性を表 4 に示す．1 位は典型的な仮定法の帰結節の形となっており，4,5 位には “wish”,“if” といった表現がある．一方で，1 位の素性は帰結節の主語が代名詞 (PP) で助動詞は should に，2 位の素性は条件節もしくは帰結節の主語が固有名詞 (NP) に限定されるなど，やや一般性に欠ける．実際，この分類器で一般の英文章を入力とすると，上位の素性が思いの外，適用されないことが分かる．そこで，再度，前処理の段階に戻り，浅い構文情報における名詞チャンク (NC) より下の品詞情報を削除する前処理を施し，学習し直すことで，より一般的な検出ルールに精錬される．

4.4 検出ルールの精度

検出ルールについては，それぞれの精度として，再現率 (Recall: R) と適合率 (Precision: P) を見積もっている．X を検出すべき対象の集合，Y を検出ルールが検出した対象の集合とすると，R, P は各々以下のように算出される．

$$R = \frac{|X \cap Y|}{|X|}$$
$$P = \frac{|X \cap Y|}{|Y|} \tag{8}$$

4.3.3 節で示した BACT による仮定法の検出ルールの精度は，$R = 72.8\%, P = 70.1\%$ であった．

なお，同じ文法項目を対象とした検出ルールは，一般に複数記述されており，論理和や論理積といった複合化の仕方によって精度は変化する．目的に応じて検出ルールの組み合わせを選択することになる．

5. 学校文法に基づいた言語データ分析

5.1 言語データ分析における学校文法

コーパスなどの言語データ分析の際，最終的に得られた言語特徴を学校文法で説明することは少なくない．たとえば，23) は論文の「表現」という点で良質な英語科学論文と日本人が書いたそうではない英語科学論文を，品詞 trigram 分布に基づいた文書分類モデルという観点から比較している．表 5 は，分類モデルにおいて差が大きかった品詞 trigram 分布の条件部である．品詞 trigram 分布の差は，主に文法的な特徴に起因することが予測される．しかし，表 5 からその示唆を直接読み取ることは極めて難しい．23) では，これらの分布を複数人の専門家が該当する実例を参照しつつ，表 6 のような学校文法上の項目に翻訳している．一般の英語学習者や英語教員には，この段になってはじめてその成果が有機的な知識として，受け止められることとなる．

現在，コーパスなどを対象とした言語データ分析では，サイズが大規模であったり，あるいは着目すべき言語特徴がより深いレベルであったりするため，形態素解析や構文解析といった言語解析を適用することが一般的である．このような言語解析は，その多くが形式文法をベースとしているために，観測できる言語特徴は形式文法上の情報で，それらがそのまま教育的示唆につながることは稀である．その結果，このような形式文法ベースの言語特徴から学校文法への翻訳作業が求められることになる．その過程で実例と作業者の経験や専門的知識を活用することによる利点もあるものの，その作業負荷は少なくなく，客観性と網羅性という観点で問題がないわけではない．現に 23) は，差異として得られた品詞 trigram 分布を全て精査できてはいない．このようなことからも，学校文法上の言語特徴を直接観測するための解析技術が重要となる[⋆1]．

5.2 検出ルールを活用した頻度分析

本研究で記述し，蓄積している文法項目の検出ルールは，前節で述べた問題に対する試みの一つとしても位置づけられる．ただし，現状では少なくとも次の 2 つの問題がある．

一つは，本研究の検出ルールが実現するのは，単純で単独の文法項目の検出で，表 6 のよ

⋆1 学校文法でとらえられるのは主に構文から一部の意味レベルのもので，現在の自動解析のものを全てカバーするわけではない．

1	`‘‘ JJ`	11	`, WRB`	21	`DT NN`	31	`#1 RB`	41	`NN NNS`
2	`NPS IN`	12	`#1 PP`	22	`IN JJ`	32	`VBZ VVN`	42	`CC VVN`
3	`SYM NP`	13	`IN VVZ`	23	`, NP`	33	`NP TO`	43	`NP IN`
4	`JJ SYM`	14	`DT VBZ`	24	`RB VVN`	34	`NNS NN`	44	`NN VVD`
5	`VBZ NN`	15	`JJ VVN`	25	`NN NN`	35	`#1 CC`	45	`IN NN`
6	`NN ’’`	16	`RB CC`	26	`CC NN`	36	`JJ TO`	46	`, EX`
7	`PP VBD`	17	`NNS VVZ`	27	`, DT`	37	`IN DT`	47	`NNS (`
8	`VV TO`	18	`, PP`	28	`NN IN`	38	`NN VBP`	48	`, NN`
9	`NN VHP`	19	`NN VVG`	29	`NN (`	39	`CC RB`	49	`NN VBZ`
10	`VBZ DT`	20	`#0 #1`	30	`IN NP`	40	`JJ NN`	50	`VVN ,`

表 5 日本人英語科学論文の特徴的な品詞 trigram 分布の条件部

過剰使用（+）	過少使用（−）
名詞による名詞の修飾・重出，現在分詞による名詞の修飾，関係節（先行詞主格）による後置修飾，“to” を除く前置詞句による名詞の後置修飾	形容詞の限定用法，過去分詞による名詞の後置修飾，TO 句による後置修飾，形容詞の重出
文頭の前置詞，文頭の接続詞・副詞（連結語），受動態	文頭の名詞句，主語・述語間の副詞，副詞節前文（受動態）の分詞化，副詞節後文の分詞化，前置詞句における名詞句の省略

表 6 表 5 の特徴（破線より上は名詞句の修飾にかかわるもの）

うな複数項目の組み合わせをとらえる枠組みとはなっていない．

もう一つは，このような文法項目でも，その検出ルールは相当数あり，しかもそれぞれ着目すべき言語特徴が異なるために，即時に全て高精度で実現することが難しいということである．そこで，これを英文章中の文法項目の頻度分析に利用する際には，各検出ルールで見積もられている精度で次のような補正を行う．検出ルールの再現率と適合率がそれぞれ R と P，項目ルールが検出した集合を Y とすると，検索ルールから $|Y|$ が得られるので，これに P/R を乗じたものを補正頻度とする．ただし，R や P は当然，分析対象の文章の性質によって大きく変化する．したがって，各文法項目の検出ルールにおける R, P を，事前に別途見積っておく必要がある．

5.3 日本人英語科学論文の頻度分析

本節では，これらの成果を活用した試験研究の一部を示す．試験研究は，論文の表現という点で良質な英語科学論文（G クラス）と日本人によるそうではない論文（JP クラス）を学校文法の諸項目で直接対比し，日本人英語科学論文の特徴を探るものである．

レベル	誤りの種類と回数（観点 A）	学術雑誌への掲載（観点 B）
L5	十分に良質で修正の必要はない	そのまま掲載可
L4	軽微な誤りが 250 語あたり 2 箇所以下，なおかつ NNS 特有の誤りは皆無である	
L3	軽微な誤りと NNS 特有の誤りがいずれも 250 語あたり 2 箇所以下，または NNS 特有の誤りが 250 語あたり 3,4 箇所ある	そのままで掲載可，または軽微な修正の上掲載可
L2	NNS 特有の誤りが 250 語あたり 8 箇所以下である	掲載不可
L1	NNS 特有の誤りが 250 語あたり 8 箇所より多い	

表 7 科学論文における表現の質区分

5.3.1 デ ー タ

Web から収集した IMRAD 型に類する文章構成を含む英語科学論文で，それに英文校正の専門家が表現上の質判定を行った[16),25)]．その基準は表 7 のとおりで，「英文章中の表現の誤りの種類（軽微な誤り/非母語話者 (NNS) 特有の誤り）と回数」（観点 A）と「各分野で高い評価を得ている学術雑誌にそのまま掲載できるものかどうか」（観点 B）によって規定される．なお，「軽微な誤り」とは科学論文に通じた母語話者 (NS) でも犯すようなミススペリングや編集ミスといったもの，「非母語話者特有の誤り」とは NS は決して犯さない文法的誤りや不自然なコロケーション，科学論文としては不自然な表現（まわりくどい表現，古風/カジュアルな表現）などである．

この試験研究では，このうち L5,L4 の論文を G クラス，L2,L1 の論文で日本人が第一著者であるものを JP クラスとし，学校文法の文法項目の観点から頻度分析を行う．論文数はそれぞれ G クラス 384 編，JP クラス 397 編である．

5.3.2 文法項目の使用頻度

論文ごとに文法項目の検出ルールを適用し，各項目の使用頻度を算出し，それをクラス間で次のように比較した．なお，事前に実験データから 200 文ランダムに抽出し，各文法項目の R, P を見積もり，頻度を補正している[⋆1]．

クラス C に属する論文の集合を $\{c_1, c_2, \cdots, c_m\}$，論文 c に含まれる文集合を $\{s_1, s_2, \cdots, s_n\}$ とする．文 s に対して得られた文法項目 g の頻度補正値を $f_g(s)$ とする

⋆1 この 200 文で未観測だった文法項目については，学参例文データベースで見積もられた精度を代わりに使用した．

と，論文 c における文法項目 g の 1 文あたりの平均使用頻度 $a_g(c)$ は，次式で与えられる．

$$a_g(c) = \frac{1}{|c|}\sum_{s\in c} f_g(s) \tag{9}$$

さらに，(9) をクラス内の論文間で平均化した，クラス C における g の「1 文あたりの平均使用頻度」の平均 $A_g(C)$ は次のようになる．

$$A_g(C) = \frac{1}{|C|}\sum_{c\in C} a_g(c) \tag{10}$$

また，クラス C における (9) の分散 $V_g(C)$ は，

$$V_g(C) = \frac{1}{|C|-1}\sum_{c\in C}\{a_g(c) - A_g(C)\}^2 \tag{11}$$

となる．(10),(11) より項目 g におけるクラス C_1, C_2 間の差の尺度として t-score を考え，次のように算出する．

$$t = \frac{|A_g(C_1) - A_g(C_2)|}{\sqrt{V_g(C_1)/|C_1| + V_g(C_2)/|C_2|}} \tag{12}$$

ただし，この t-score はあくまでも項目間の使用頻度の差に優先順位を付すためで，統計的仮説検定にまで帰着するものではない．

t-score が 3 以上・−3 以下のものを表 8 に示す[⋆1]．3 以上の項目つまり G クラスの方が平均使用頻度が高い項目が圧倒的に多い．これは G クラスの節数が JP クラスのそれよりも高く[⋆2]，絶対数を比較する t-score では正の方に多くの項目が挙げられることになる．このようななかでも，JP クラスで過剰使用となっているのが「現在形」と「受動態」である．この 2 項目については従来から指摘されているものである．また，動詞部分で時制・相・法・態などの組み合わせで t-score を上記のように計算すると，「現在時制・受動態」($t = -9.69$) がやはり最上位になる．それに加え，「現在時制・完了進行相」($t = -4.02$) が JP クラスでは特に高い，という意外な結果が得られる．これについては，今後，質的に分析を進める予定である．

(9) では文数で平均化しているが節数に以上のような差があるので，節数で平均化して見直してみる．論文 c における総節数を T_c とすると，論文 c における文法項目 g の 1 節あたりの平均使用頻度 $a'_g(c)$ は，次式のようになる．

⋆1 この数値，表 9 における ±1 という数値は，あくまでも本稿での暫定的なものである．
⋆2 平均文長（語数）は G クラスは 23.9，JP クラスは 20.9 で，不偏標準偏差はそれぞれ 9.9，14.7 である．G クラスの方がより複雑な文を産出することによると予想される．

文法項目 (g)	t-value	A_g(G)	A_g(JP)
現在形	−5.34	0.631	0.676
受動態	−4.75	0.163	0.186
存在 THERE	3.37	0.030	0.023
同等比較	3.46	0.063	0.053
過去形	3.85	0.141	0.113
原形	4.29	0.215	0.200
仮定法	4.63	0.007	0.004
助動詞	4.63	0.201	0.173
関係代名詞	4.72	0.902	0.781
従属接続詞	5.15	0.226	0.199
疑問詞	5.19	0.218	0.189
形容詞・最上級	6.45	0.038	0.027
等位接続詞	6.97	0.632	0.592
副詞・比較級	7.53	0.016	0.103
数量表現	7.64	0.840	0.795
形容詞・比較級	7.92	0.085	0.061
副詞・最上級	8.62	0.019	0.011
比較級＋比較級	8.97	0.013	0.090
分詞構文	9.10	0.107	0.074
節数	10.85	2.652	2.358

表 8 G/JP 間で平均使用頻度の差が大きい文法項目

$$a'_g(c) = \frac{1}{T_c}\sum_{s\in c} f_g(s) \tag{13}$$

(10),(11) 内の $a_g(c)$ を $a'_g(c)$ に置き換え，t-score が 1 以上・−1 以下となったものを表 9 に示す．このようにしてみると，JP クラスではもともと形容詞があまりうまく活用出来ていないことが知られているが，さらに比較級・最上級といったより巧みな用法は，やはり不十分であることが分かる．副詞についても同様である．また，分詞構文・仮定法も従来から指摘されていたような，JP クラスでは十分に使用できていない傾向がみられる．

6. おわりに

本稿では，英語の学校文法に関する情報が付与された英文データ構築の方法論を俯瞰し，現況を述べた．そして，文法項目の検出ルールの具体的な試験的応用として，日本人英語科学論文の特徴分析の事例を示した．より実際的な活用には，検出ルールの精度の問題，そして諸項目の組み合わせをどのように織り込んでいくかという問題を検討していく必要がある．

文法項目 (g)	t-value	A_g(G)	A_g(JP)
現在形	−1.94	0.062	0.079
受動態	−1.66	0.238	0.287
形容詞・最上級	1.27	0.014	0.011
仮定法	1.37	0.003	0.002
形容詞・比較級	1.37	0.032	0.026
比較級＋比較級	1.48	0.048	0.038
分詞構文	1.65	0.040	0.032
副詞・比較級	1.79	0.006	0.004
副詞・最上級	2.23	0.071	0.045

表 9 節単位で G/JP 間で平均使用頻度の差が大きい文法項目

謝辞 本研究の成果の一部は，立命館大学学内提案公募型研究推進プログラム，文部科学省科学研究費補助金によるものである．

参 考 文 献

1) Aarts, J. and Granger, S.: Tag Sequences in Learner Corpora: a Key to Interlanguage Grammar and Discourse, In Granger, S.(ed.) pp.132-141 (1998).
2) 工藤 拓：BACT: a Boosting Algorithm for Classification of Trees, `http://chasen.org/~taku/software/bact/` .
3) Eugene Charniak's Home Page, `http://www.cs.brown.edu/~ec/` .
4) Granger, S.: Learner English on Computer, Addison Wesley Longman (1998).
5) 小林雄一郎, 田中省作, 後藤一章, 徳見道夫, 朝尾幸次郎: 学校英文法コーパスの提案 - デザインと応用可能性-, *NLP* 若手の会第 *3* 回シンポジウム, 4 page (2008).
6) 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).
7) 小池 浩: 必修英文法問題精講, 旺文社 (2006).
8) 小寺茂明: デュアルスコープ総合英語 三訂版, 数研出版 (2006).
9) Kudo,T. and Matsumoto, Y.: A Boosting Algorithm for Classification of Semi-Structured Text, *EMNLP 2004* (2004).
10) 中原道喜: 基礎英文法問題精講 三訂版, 旺文社 (2003).
11) 日本物理学会: 科学英語論文のすべて 第 2 版, 丸善 (1999).
12) 塊 タカユキ, 石黒昭博: 総合英語 Forest 5th edition, 桐原書店 (2006).
13) Penn Treebank Project, `http://www.cis.upenn.edu/~treebank` .
14) 佐伯里子: プリズム総合英語, 美誠社 (2002).
15) 佐野 洋，猪野真理枝：英語文法の難易度計測と自動分析，情報処理学会コンピュータと教育研究会（CE）報告，Vol.2000，No.117，pp.5-12 (2000).
16) Shibata, M., Tomiura, Y., Mizuta, T.: Identification among Similar Languages Using Statistical Hypothesis Testing, *Proceeding of Pacific Association for Computational Linguistics*, pp.47-51 (2009).
17) 清水周裕: チャート式現代英文法, 数研出版 (1996).
18) 霜崎 實: クラウン総合英語, 三省堂 (2008).
19) 杉山忠一: 英文法詳解, 学習研究社 (1998).
20) 鈴木希明: 高校総合英語 Harvest 第 3 版, ピアソン桐原 (2008).
21) 高沢節子, 豊島克己, 町田 健: depth 英語総合, 河合出版 (2002).
22) 田中 実: ラーナーズ高校英語 五訂版, 数研出版 (2009).
23) 田中省作, 藤井 宏, 冨浦洋一, 徳見道夫: NS/NNS 論文分類モデルに基づく日本人英語科学論文の特徴抽出, 英語コーパス研究, 第 13 号, pp.75-87 (2006).
24) 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, 4 page (2008).
25) 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).
26) 東京外国語大学佐野研究室：文法項目別 BNC 用例集 — N-Cube, `http://scn02.corpora.jp/~n-cube/` .
27) TreeTagger - a language independent part-of-speech tagger, `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/` .
28) 綿貫 陽，宮川幸久，須貝猛敏，高松尚久：ロイヤル英文法 改訂新版，旺文社 (2000).
29) 山口俊治: コンプリート高校総合英語, ピアソン桐原 (1989).