様式 C-19

科学研究費補助金研究成果報告書

平成２３年６月２４日現在

機関番号：３４３１５
研究種目：若手研究(B)
研究期間：2009～2010
課題番号：２１７２０２１１

研究課題名（和文）教材の自動評価のための学校文法に基づいた安全な英文解析システムの開発

研究課題名（英文）Development of an English parser with its validation based on school grammar for material evaluation

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

研究成果の概要（和文）：
本研究は，教材の自動評価を念頭に，学校文法に基づいた精度保証付き英文解析システムの開発を行った．まず，学校文法の項目を精査，類別し，科学文法との対照性を調査した．次に，科学文法上の情報を活用した学校文法における文法項目の検出ルールを与え，それらの検出精度を見積もった．そして，それらを統合し，英文解析システムを構築した．現在，プロトタイプを Web 経由で利用できるよう公開準備を進めている．

研究成果の概要（英文）：
This project developed an English parser with its validation based on school grammar for automatic evaluation of educational materials. First, I investigated items in school grammar from the view point of scientific grammar and showed that they were regularly correlated with various features in scientific grammar. Second, on the basis of the results of this investigation, I specified detection rules for items in school grammar and estimated the accuracies of these rules. Finally, the project developed the English parser by their rules. A prototype of this system will be released on the Internet as a Web application.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	2,300,000	690,000	2,990,000
2010 年度	1,100,000	330,000	1,430,000
総　計	3,400,000	1,020,000	4,420,000

研究分野：外国語教育、知能情報学
科研費の分科・細目：言語学・外国語教育
キーワード：学校文法、科学文法、言語処理、機械学習

１．研究開始当初の背景

英文をさまざまな観点から客観的かつ手軽に自動評価する技術の開発は，外国語教育を支える言語教育工学の重要な課題の一つである．近年，主に言語処理分野で開発が進んでいた語彙レベルの言語解析（形態素解析，タギングなどと呼ばれる）の完成度が高まったのに加え，コーパス（電子化された大規模用例集）を用いた頻度情報に基づいた語彙リストや，日本人英語学習者にとっての親密度に基づいた語彙リストなどの客観的な語彙知識の整備が進み，教材の語彙レベルでの自動評価は充実しつつある．しかし，この現状

には次のような，少なくとも 2 つの大きな問題がある．

(1) 学校文法レベルの情報をとり扱えていないこと
学校文法は，英語を母語としない多くの日本人にとって，英語理解のもっとも重要な観点の一つである．中学生ー高校生のまだ網羅的に文法を習得していない学習者はもちろんのこと，ひととおり文法を修めた学習者であっても，特定の文法項目に過度な苦手意識をもっている場合もある．したがって，教材の妥当性や難易を診断するのに，学校文法に関わる情報を欠かすことはできない．しかし，現在の教材の自動評価では，学校文法レベルの情報が欠けており，英文に含まれる文法項目を考慮した評価や提示は，十分には実現できていない．

(2) 学校文法に基づいた英文解析システムがないこと
計算機で文を解析・生成する言語処理分野で，「文法」として学校文法が注目されることは稀有である．なぜならば，学校文法は人が教授する，なおかつ人が学習することを前提としたもので，規則性や厳密性の問題から計算機上で形式的に取り扱うことが難しいからである．言語処理における「文法」には，もっぱら句構造文法や単一化文法といった，いわゆる科学文法が採用されている．このような状況から，学校文法に基づいた英文解析システムは代表者の知る限りでは皆無である．これも前項で述べた教材の自動評価において，学校文法に関する情報が欠落する原因の一つである．
なお，東京外国語大学佐野研究室が開発しているコーパス検索システム N-Cube は，学校文法を織り込んだ数少ないシステムである．N-Cube は，予め文法項目ごとに人手で形態素レベルの検索式を記述し，実装している．学校文法の観点を導入したという意味で画期的ではあるものの，あくまでもコーパス検索が前提のシステムであり，形態素レベルによる記述力の限界もある．

２．研究の目的

本研究の主目的は，前節で述べた 2 つの問題の解決を目指し，学校文法に基づいた英文解析システムを開発することである．そして，次のような 3 つの小テーマを設定した．

(1) 英文解析を念頭に置いた学校文法における文法項目の類別と科学文法の対照調査

(2) 学校文法に基づいた安全な英文解析システム構築のための方法論の提案
①文法項目の検出ルールの記述法の開発
②文法項目の検出ルールの精度評価

(3) 学校文法に基づいた英文解析システムの構築と公開

なお，英文解析システムを実際に活用し，学校文法レベルの教材プロファイル等を試行することも，本研究の副次的目的である．

３．研究の方法

[2009 年度]
(1) 英文解析を念頭に置いた学校文法における文法項目の類別と，科学文法との対照調査
佐野ら[1]などの学校文法に関する従来研究・文献を参照し，学校文法の文法項目を，以下のように 30 程度を一つの目安に選分，類別した．なお，括弧内は当該項目がとりうる値で，下線が付されているものがデフォルト値である．

①主に文・構文に関わる項目
文型（1-5 文型），文の種類（平叙・疑問・命令・感嘆），文の単複等（単文・重文・複文・混文），疑問文の種類（一般・特殊・選択・間接・付加），極性（肯定・否定），否定の種類（全否定・部分否定），倒置構文，比較級+比較級構文，存在 there 構文，分詞構文（主語一致・主語不一致・慣用表現），強調構文，動名詞の慣用構文

②主に動詞に関わる項目
時制（現在・過去・未来），態（能動・受動），法，法（直接・仮定・命令），相（進行・完了），to 不定詞（名詞的・形容詞的・副詞的），原形不定詞，分詞（限定・叙述），動名詞，助動詞

③形容詞・副詞に関わる項目
比較（原級・比較級・最上級），同等比較

④関係詞に関わる項目
関係代名詞の格（主格・目的格・所有格），関係代名詞の用法（制限・継続），関係副詞，複合関係詞（代名詞・副詞・形容詞）

⑤その他の品詞に関わる項目
接続詞（等位・従属），疑問詞・疑問詞+to+V

⑥その他の項目
話法（直接・間接），時制の一致（一致・不一致），数量表現

これらの学校文法の文法項目と，科学文法の一つである句構造文法（正確には，文脈自

由文法: CFG）の文法的特徴（単語/品詞の混合列や部分的な構文構造など）と対応関係について調査し，本研究で検出の対象とする文法項目を絞り込んだ．

その結果，2/3 程度の学校文法の項目は品詞/単語の混合列である程度同定できることが見込まれた．一方，残りの 1/3 程度については，多くが部分的な構文構造で同定できることが予想された．しかし，精査を進めていくなかで，その一部には構文レベルでも峻別が難しいものも存在することが明らかとなり，それらは改めて検出対象としての文法項目からは外すこととした．

(2) 科学文法の情報を活用した文法項目の同定方法の検討

(1)の調査で，品詞/単語の混合列で一定程度同定されることが予想された学校文法上の 21 項目について，その検出ルールの整備を試みた．形態素レベル（品詞/単語の混合列）の検出ルールは基本的に人手で記述する．品詞/単語の混合列が当該項目の表層的特徴を表しており，それが形態素解析を通した入力英文に含まれていれば即，当該項目と判断する非常に単純な決定的ルールである．たとえば，受動態は補文構造なども含めて入力英文の動詞部分ごとに，

/BE 動詞/ + (*/副詞/*)
+ */動詞/過去分詞

という単語/品詞列が照合すれば検出される．なお，α/β/γは一形態素情報で，αが表記，βが品詞，γが活用形を指定しており，これらは論理積として働く．(δ)はδが随意的な形態素であること，*は任意であることを表す．α，β，γは正規表現で，形態素間の任意のギャップを表現することもでき，本質的には佐野ら[1]がベースとした CQL(Corpus Query Language)と同等の記述力を有す．

そのようにして記述した検出ルールに対して，別のプロジェクトで整備を進めている学校文法情報付き英文データベース（英文に対して使用されている学校文法の文法項目に関する情報が付与された文例集で，その英文は主に中高生向け英語文法参考書から収集されている）を使用して，その検出精度（適合率と擬似再現率）を与えた．擬似再現率とは，データベースの英文には，全ての文法項目の使用の是非が与えられていなかったため，データベースからランダムに抽出した英文に対して，事前に当該項目の使用の是非を別途チェックした上で与えた近似的な再現率である．

(3) 機械学習を活用した文法検出ルールの自動記述

(2)のように形態素レベル，そして人手や内省で記述が難しいものには，文内の比較的広い範囲をみなければならなかったり，構文レベルでの記述が求められたりすることが推測される．そこで，検出ルールを効率的に記述するために，構文解析を導入し，機械学習理論のなかでも構造の取り扱いに向いた決定株を弱学習器としたブースティング（BACT）[2]を援用することを検討し，予備的な実験を行った．

[2010 年度]

(1) 再現率を重視した文法項目の検出ルールの整備

BACT を活用すると比較的容易に構文構造まで考慮した検出ルールが記述される（ただし，BACT が構築する分類器は単語/品詞の混合列のように単純な照合ではなく，部分構造の重みつき投票による判別で，それが検出ルールに相当する）．それらの検出精度を評価すると，学習データの量やヴァリエーションなども影響してか，適合率は高いものの，やや再現率が低くなる．応用研究によっては，滅多には使われないものの，唯一の使用が重要な指標となるような文法項目を求めることもある．そこで，このような BACT による検出ルールの自動記述の結果・傾向を踏まえ，再現率に重きを置いた検出ルールの補完を次のように進めた．

①浅い構文解析（チャンキング）の導入

2009 年度に用いていた CFG ベースの構文解析よりも言語情報的には浅い，チャンキングとよばれる構文解析を導入した．チャンキングは，文中の名詞チャンク（NC）・動詞チャンク(VC)・前置詞チャンク（PC）といった構文的まとまり（チャンク）を同定する．一部のチャンク間の構造的関係も明らかになるものの，言語情報の観点では，ちょうど形態素解析と CFG ベースの構文解析の間に位置するような解析である．たとえば，

If I was superman, I could help you.

に対するチャンキングの結果は以下のとおりである．

[$_{S}$[$_{IN}$ if] [$_{NC}$ [$_{PP}$ I]] [$_{VC}$ [$_{VBD}$ be]] [$_{NC}$ [$_{DT}$ a] [$_{NN}$ superman]] [$_{,}$,] [$_{NC}$ [$_{PP}$ I]] [$_{VC}$ [$_{MD}$ could]] [$_{VV}$ help]] [$_{NC}$ [$_{PP}$ you]] [$_{SENT}$.]]

なお，このようなチャンキングの結果を，ここではチャンク構造とよび，チャンク構造の構造的包含関係については[2]に準じる．

②チャンク構造に対する前処理

文法項目には，仮定法における if，as，

wish や関係代名詞における who, which など，いくつか特徴的な単語が存在することがある．このような単語の多くは，専門家でなくとも文法書等から比較的容易に収集することができる．そこで，各英文のチャンク構造から，対象としている文法項目の特徴的な単語以外については予め削除する前処理を施す．たとえば，①で挙げた仮定法を含む例文は，前処理の結果，次のようなチャンク構造に簡素化される．

[S[IN if] [NC [PP]] [VC [VBD]] [NC [DT] [NN]] [,] [NC [PP]] [VC [MD could]] [VV]] [NC [PP]] [SENT]]

このように文法項目には直接関わらない単語を予め削除しておくことで，当該項目には本質的ではない単語を含むような検出ルール生成の抑制を図る．

③BACT が選択した素性の活用

特定の文法項目に関して簡素化した英文のチャンク構造から BACT によって構築される分類器（検出ルール）を精査すれば，どのような素性（部分チャンク構造）が当該項目に深く関わっているかが分かる．動詞の態や相といった局所的な情報だけで同定されるものではない，たとえば仮定法や分詞構文のような英文中の比較的広い範囲をみて同定される文法項目では，BACT が選択したこの素性は強力な手掛かりとなる．そこで，再現率を高めたい場合には，BACT が選択した素性を参照しつつ，人手で構造レベルの決定的な検出ルールを加筆する．たとえば，仮定法の際の上位 5 位の素性は次のとおりである（このような部分チャンク構造を含む場合に「仮定法である可能性が高い」という判定を下しやすくなる）．

・[S [NC [PP]] [NC [PP]] [VC [MD should]] [VV]] [ADVC [RB]] [SENT]]
・[S [NC [NP]] [SENT]]
・[S [ADVC [RB]] [NC [DT]] [SENT]]
・wish
・if

そして，以下が機械学習を直接的そして間接的に活用した特定の文法項目の検出ルールの具体的な記述法である．

Step 1. 各英文に対して TreeTagger チャンキングを適用し，チャンク構造を得る．

Step 2. 各英文のチャンク構造から，当該項目に特徴的ではない単語を削除する．

Step 3. 英文のチャンク構造に BACT を適用し，当該項目に関する分類器（検出ルール）を構成する．

Step 4. 分類器を精度評価し，再現率が低い場合には，素性（部分チャンク構造）を手掛かりに決定的な検出ルールを記述，補完する．

(2) 英文解析システムの開発と公開

各文法項目の検出ルールを適用し，その結果を統合した，学校文法に基づいた英文解析システムを実装する．その出力には，検出したルールの種別（形態素レベルの単語/品詞の混合列・BACT・BACT の素性等をもとにした再現率を重視したルール）と，その検出の根拠となった英文中の箇所および推定精度を付記する．これらの付帯情報は，利用者がある程度，自身で各結果を取捨選択するための判断材料である． たとえば，

Being tired, he sat down to rest.

に対する結果の一部は，以下のようになる．

態=受動 <- Being_1 tired_2 (0.95/1.00)
時制=過去 <- sat_5 (0.96/1.00)
文種類=複文 <- Being_1 ... ,_3 ... sat_5 @ [S [VC [V*G]] [, ,] [VC [VV*¥=V*G]] (0.41/0.04)
TO 不定詞=副詞的用法 <- (0.74/0.55)
分詞構文 <- Being_1 ... ,_3 ... sat_5 @ [S [VC [VBG] [, ,] [VC [VV*¥=V*G]] (0.89/0.08)

結果の中の“α <- β (P / R)”は，α が検出された文法項目で‘=’を含む場合は値，β はその検出の根拠となった箇所で，P が適合率，R が擬似再現率を表している．β が ε（空）のものは BACT による検出である．β が ε ではなく，なおかつ@区切りがないものは単語/品詞の混合列，一方@区切りのあるものは@以降に提示される部分チャンク構造等によって検出されたものである．

[1] 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報告，第 2000 巻，第 117 号，pp. 5-12，2000 年

[2] 工藤 拓，松本 裕治: 半構造化テキストの分類のためのブースティングアルゴリズム，情報処理学会論文誌，第 45 巻，第 9 号，pp. 2146-2156，2004 年

４．研究成果

本研究は，主に次のような成果をあげた．

(1) 学校文法における文法項目の類別と，科学文法との対照

従来研究を参照しつつ，学校文法における文法項目を選分し，英文解析システムを念頭に類別化した後，科学文法でも特に句構造文法との対応関係を調査した．

(2) 文法項目の検出ルールの記述法と整備

①人手による検出ルールの記述

単語/品詞の混合列のような形態素レベルで文法項目の検出ルールを記述，蓄積した．このような形態素レベルの検出ルールは比較的精度が高いことを確認した．

②機械学習による検出ルールの自動記述

一般の構文解析よりも言語情報的には浅いチャンクキングを導入し，なおかつ各文法項目に関する特徴的ではない単語を削除する前処理を施した．その結果，素直に構文解析した結果を用いるよりも，過学習を抑制しつつ，高い適合率と再現率を有す検出ルールの自動記述が可能となった．

②機械学習を援用した検出ルールの補完

比較的広い範囲を見なければならないような，内省に基づいて記述するには不安が残る文法項目に対する記述の支援法を検討した．特定の文法項目を対象に BACT を適用した際に得られる素性を補足的に活用し，検出ルールを記述することを試行した．その結果，再現率をより高める検出ルールが補完された．

(3) 学校文法に基づいた英文解析システムの構築と公開

上述のような方法で整備した学校文法の検出ルールを集約し，結果に検出根拠等の情報を合わせて提示する英文解析システムを構築し，Web 上での公開を進めた．

以上が主な研究成果であるが，最後に今後の課題および展望を記しておく．

英文解析システムのインタフェースがやや不十分である．特に精度表示については，適用された検出ルールのみの提示で，同項目を検出する他のルールとの関係性を制御できていない．したがって，現在の方式では信頼性の判断が難しい場合がある．また，応用研究への実際的適用が少ない．今後，ここまでの成果と合わせて，その応用や論文等への発表に努めたい．

なお，本研究テーマからは外れるが，機械学習を使用した検出ルールの記述支援の過程で得られる素性については，文法項目間の構造的類似性などを測る手掛かりとなることが偶発的に示唆された．それは，これらの素性の当初意図した利用法ではないものの，今後そのような視点から学校文法を捉え直すという新しいテーマも検討していきたい．

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 8 件）

① 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 号，pp.61-74，2011 年（査読有）

② 田中 省作：Web コーパスの言語情報処理基盤，英語コーパス研究，第 18 号，pp.97-112，2011 年（査読有）

③ 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.61-72，2011 年（査読有）

④ Miyazaki, Yoshinori, Ikemoto, Takanori and Tanaka, Shosaku: Development of Web Application to Help Write Technical Documents in English -Using Corpus for Language Teaching-, Proceedings of *the ICTATLL 2010 Kyoto Conference*, pp.149-158, 2010 年（査読有）

⑤ 多田 一馬，田中 省作：類義語を手掛かりとした未知語推測ストラテジのための基礎調査，統計数理研究所共同研究リポート，第 239 号，pp.43-54，2010 年（査読無）

⑥ 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，第 239 号，pp.13-30，2010 年（査読無）

⑦ 木村 恵，田中 省作，八島 等，依田 みずき：言語資源とその処理技術を活用した L2 語彙の習得レベル判定，英語コーパス研究，第 16 号，pp.1-14，2009 年（査読有）

⑧ 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 -データベース・ソフトウェアを用いた英語学習教材の自動作成-，自然言語処理，第 16 巻，第 2 号，pp.45-58，2009 年（査読有）

〔学会発表〕(計 7 件)

① 田中 省作, 宮崎 佳典, 池本 孝徳, 小山 由紀江: 英作文支援のためのクラス n-gram モデルに基づいた文例の汎化, 応用数理学会環瀬戸内応用数理研究部会第 14 回シンポジウム, 2011 年 1 月 23 日, 岡山理科大学 (岡山県)

② 田中 省作: 質問紙分析法, 母語・継承語・バイリンガル教育 (MHB) 研究会リサーチメソッド学習会, 2010 年 11 月 7 日, 立命館大学 (京都府)

③ 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 -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 年 10 月 3 日, 青山学院大学 (東京都)

④ 田中 省作: 習得度測定のための回答時間の基礎的調査 -正答受験者の回答時間分布の特性-, 日本教育工学会第 25 回全国大会ワークショップ, 2009 年 9 月 19 日, 東京大学 (東京都)

⑤ Kitao, Kenji and Tanaka, Shosaku: Authorized Junior High School English Textbooks: a Corpus-based Study of Vocabulary Level and Readability, EuroCALL2009, 2009 年 9 月 11 日, Universidad Politécnica de Valencia (Spain)

⑥ 田中 省作, 小山 由紀江: 日本の英語教科書を基準とした ESP 特徴表現の抽出, 第 49 回外国語教育メディア学会全国研究大会, 2009 年 8 月 6 日, 流通科学大学 (兵庫県)

⑦ 田中 省作: 分割表に対する近似検定としての χ^2 検定, 第 33 回英語コーパス学会ワークショップ, 2009 年 4 月 25 日, 神戸大学 (兵庫県)

〔図書〕(計 1 件)

① Miyazaki, Yoshinori, Ikemoto, Takanori and Tanaka, Shosaku: Development of Web Application to Help Write Technical Documents in English -Using Corpus for Language Teaching-, Corpus, ICT and Language Education (Scotland, UK: University of Strathclyde Publishing) (by Weir, G. and Ishikawa, S. (eds.)), pp.225-234, 2010 年 (査読有)

※雑誌論文(5)の再録

〔その他〕
http://www.cl.ritsumei.ac.jp/SGA/

６．研究組織

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