2G3-2

学校英文法コーパスの構築の試み

Development of a School Grammar Corpus of English

田中省作[*1] 小林雄一郎[*2] 徳見道夫[*3] 朝尾幸次郎[*1]
Shosaku TANAKA Yuichiro KOBAYASHI Michio TOKUMI Kojiro ASAO

[*1]立命館大学文学部 College of Letters, Ritsumeikan University
[*2]法政大学 Hosei University
[*3]九州大学大学院言語文化研究院 Faculty of Languages and Cultures, Kyushu University

Although the school grammar of English is one of the most important perspectives of the language that non-native speakers of English require to understand in order to learn the language, there is no English corpus annotated with the structure of school grammar. The purpose of this research is to develop a prototype of such a corpus and describe rules to identify items of school grammar. This prototype is annotated with syntactic structure of science grammar by hand. Since most of the items of school grammar are partially related to the syntactic information, we can obtain the rules to identify these items in a sentence from its surface and syntactic information. Furthermore, a machine learning method generates these rules efficiently using the prototype data in order to become efficient at developing the above-mentioned corpus.

1. はじめに

近年，さまざまな言語で電子化大規模用例集（コーパス）が整備され，言語処理や言語教育をはじめとして，新しい方法論の確立や成果が得られている．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る限りでは，研究等に自由に利用できる学校文法に関する情報を付与したコーパス（**学校英文法コーパス**）はない[*1]．一般の英語学習者や英語教員にとっては，科学文法よりも学校文法の方が身近で，学校英文法コーパスのニーズは極めて高い．そこで現在，我々は学校英文法コーパスの効率的な構築を検討している．

本研究では，まず小規模な学校英文法コーパスを構築し，そのデータを使って文法項目を検出するルール（**文法項目の検出ルール**）を記述する．この初期の粗い文法項目の検出ルールを使ってコーパスを拡充し，再び検出ルールの精密化を行う．本研究の特徴は，このような「コーパスの拡充」と「検出ルールの精密化」という良循環を意識しつつ，コーパスの整備を進める点である．本稿では，学校英文法コーパスのプロトタイプの現状と，機械学習（決定株を弱学習器としたブースティングによる分類器）を活用した初期の文法項目の検出ルールについて述べる．

2. 学校英文法コーパス

2.1 背景

学校文法コーパスに関わる研究として [Sano 00] や [N-Cube] がある．[Sano 00] 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．[N-Cube] では [Sano 00] を受け，1320 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，正確な精度保証がなされていないという点では，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コーパスの必要性は高い．

2.2 構築方針

学校文法項目には，表層表現や品詞列のレベル，科学文法の構文レベルで一意に同定できるものも多い．そこで，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学校文法項目を対応づける文法項目の検出ルールを記述する．初期は粗い検出ルールで，コーパスの構築の作業の効率化はさほど高くないことが予想されるが，データの充実化に伴って検出ルールが精密化されることが期待される．この相互作用を繰り返すことで，コーパスの拡充と検出ルールの精密化が進み，全体としての作業の効率化につながる．また，文法項目の検出ルールは整備されたコーパスで精度保証されるため，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．

現在，構築している学校英文法のプロトタイプについて述べる．付与する対象は構文解析済みコーパスである Penn Treebank[PTB] の Brown Corpus 部分からランダムに抽出した 5167 文である[*2]．

文法項目については，網羅的に設定するのではなく，[Sano 00] 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ている．今回対象とした文法項目を表 1 に示す．取り扱っている文法項目は未だ初歩的なものであり，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．

付与の単位は，文を単文・節（C_i）に分解し，その上で文の主要素（S, V, O, C）・修飾部分（M）に区別したものである[*3]．したがって，重文・複文，to 不定詞・関係節などは，互いの関係は明示しつつも，文法項目は全て別単位で付与される[*4]例えば，“I love music although I can’t play a musical instrument.”

連絡先: 田中省作，立命館大学文学部，京都市北区等持院北町 56-1，(075)466-3301，sho@lt.ritsumei.ac.jp

*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，残念ながら利用することはできない．

*2 生テキストでは，中学校の平成 14 年度検定済み英語教科書（NEW CROWN, NEW HORIZON, SUNSHINE）に付与している．

*3 単文単位であれば，本動詞に関わる文型・時制・態・法・相は，それぞれ独立に与えれば良く，文法項目数も削減される．

*4 付与の単位としては別でも，分解された文・節における互いの構文的・意味的な関係は保存して，文法項目の付与はなされる．

文法項目	値
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・比較・最上
副詞	原・比較・最上
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

では，

C_0: [I]$_S$ [love]$_V$ [music]$_O$ [although $\boxed{C_1}$]$_M$.

C_1: [I]$_S$ [can't play]$_V$ [a musical instrument]$_O$

と作業者に提示され，C_0, C_1 と文の主要素ごとに文法項目が付与される．

3. 機械学習を活用した文法項目の検出

3.1 部分木を素性とした分類

本節では，前節で整備した学校英文法コーパスに基づいた文法項目の検出ルールの記述について述べる．文法項目の検出ルールも一から人手で記述することも考えられるが，小規模なデータからでも半自動的に獲得できたり，記述のための情報が与えられれば作業の効率化につながる．そこで，本稿では機械学習の手法を応用した，文法項目の検出ルールの記述を検討する．

特定の文法項目の検出は，科学文法の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ができる[*5]．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 [Kudo 04].

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．

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

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分類エラー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，

$$\mathrm{sgn}\left(\sum_{t,y} \alpha_{\langle t,y\rangle} h_{\langle t,y\rangle}(x)\right) \quad (1)$$

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容易に観察できることである．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．

3.2 実験

2.2 節で述べた Penn Treebank の部分データを元とした学校英文法コーパスを使って，文法項目のうち仮定法と分詞構文に対して，文法項目の検出実験を行った．仮定法については，2 人の作業担当者の付与した情報が一致した 4,796 文（うち仮定法は 75 文），分詞構文は 5,167 文全て（うち分詞構文は 165 文）を対象とした．

分類器の生成には，Google の工藤 拓氏が公開している BACT を使用した [BACT]．各英文を Penn Treebank に付与されている構文木で表現し，10-交差検定で評価した．

その結果，仮定法は適合率 93.2%・再現率 73.3%，分詞構文は適合率 61.4%・再現率 53.9%となった．

仮定法については比較的高い適合率が得られたものの，両項目とも現状ではコーパスの整備に直接利用したり，応用研究へ活用したりすることは難しい．そこで，まず次節で分類に働いた素性の例を簡単に考察し，3.4 節で本ルールのコーパス構築への利用法について検討する．

3.3 素性の例

仮定法と分詞構文のそれぞれの分類器で，重み上位 20 位までの素性をそれぞれ表 2,3 に示す[*6]．なお，表中の素性は部分木を先順走査した際のノードのラベルを表しており，")" は親ノードに上がることを意味するメタ記号である．例えば，"α) β" は α と β が兄弟関係で，α が β よりも左に存在することを表すことになる．

仮定法の素性を見ると，学校文法で示される "if S V 〜, S would[should,could,*etc.*] V 〜"，"I wish 〜" といった典型的なパタンを連想させる部分木や，"if"，"could" や "wish"，"as if/though" といった表現も得られている．"as if/though" については，後続する表現が句か節かによって，本動詞の法（仮定法/直接法）が変化する場合が多い．しかし，両事例に関

*5 文法項目の検出ルールは，[Sano 00] のように表層・品詞レベルで同定できるものも多い．したがって，全ての文法項目をこのように取り扱うわけではない．

*6 それぞれ分類器を構成する際，当該文法項目を含んだ構文木のラベルを $y = +1$ とした場合である．

重み	素性
0.170	could
0.152	S VP MD would
0.108	S NP-SBJ) VP MD 'd)) VP VB
0.064	if
0.050	IN though
0.043	IN as
0.042	SBAR-ADV
0.040	SBAR-ADV IN If
0.040	VP VBP wish)) SBAR
0.040	S NP-SBJ
0.034	S NP-SBJ PRP)) VP SBAR S VP VBD
0.025	S VP VP
0.024	VBD
0.022	wish
0.020	SBAR-ADV IN) .
0.019	S VP VBD were
0.018	VP VP VBN
0.018	S NP-SBJ PRP)) VP
0.016	S NP-SBJ NP)) VP VBD had)) VP
0.015	S VP VP VB))) .

表 2: 重み上位 20 位の素性（仮定法）

重み	素性
0.047	SINV S-ADV
0.045	,
0.042	S VP ,) S-ADV VP VBG)))) .
0.038	NP NP NN)) ,) VP VBG
0.036	S VP VP ,) S-ADV NP-SBJ -NONE-)) VP
0.035	S S-ADV) VP
0.025	S VP VP VP) ,) VP
0.023	VP VBG
0.022	S S-ADV) ,) NP-SBJ-1) .
0.020	S VP ,) S-ADV NP-SBJ -NONE-)) VP
0.020	S NP-SBJ-1 DT) JJ)) .
0.019	VP S
0.019	S VP S-ADV NP-SBJ -NONE-)) VP PP IN))))) .
0.018	VP S) S-ADV NP-SBJ -NONE-
0.018	S NP-SBJ-1 PRP)) VP VBD) S-ADV VP ADVP
0.016	S-ADV VP SBAR S NP-SBJ
0.016	NP NP) PP IN) NP CC and
0.015	VP VBD was
0.015	PP IN from)) NP
0.015	S ,) VP VP VP))) . .

表 3: 重み上位 20 位の素性（分詞構文）

わるデータが過疎なために，実際にはうまく分類できていない（仮定法に分類される）．“wish” についても同様の傾向が見られる．

また，分詞構文については，確かに分詞が導く副詞節の存在を示唆する部分木や，“V-ing ～, S V ～” といったパタンを連想させる部分木も得られているものの，名詞句の構造に関するような無関係なものも比較的上位に存在している．

以上の点から，文法項目を適切に細分化し検出ルールを再設定することや，低頻度な文法項目についてはある程度恣意的に文法書などのデータを投入することも考えられる．また，今回は素直に Penn Treebank の構文木をそのまま使って分類器を構成したが，文法項目の特徴に応じた適切な構文木の前処理も重要な今後の課題である．

3.4 文法項目の付与への援用

本節では，3.2 節で得られた分類器のコーパス整備への援用について述べる．3.1 節の手法では x のクラスを，$h_{\langle t,y\rangle}(x)$ の

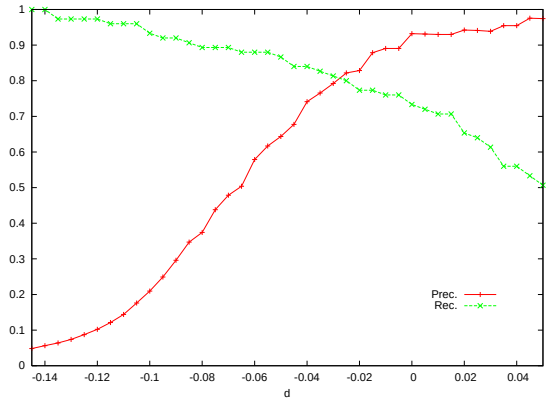

図 1: d と仮定法の適合率（Prec.）・再現率（Rec.）

重み付き多数決の正負で決定する．ここで重み付き多数決の値そのものの大きさは，概ね分類結果の確信度として考えられる．そこで分類器構成後に，分類基準を次のように変更する．

$$\begin{cases} +1 & \displaystyle\sum_{t,y} \alpha_{\langle t,y\rangle} h_{\langle t,y\rangle}(x) > d \\ -1 & \text{otherwise} \end{cases}$$

$d = 0$ のときが，ちょうど (1) 式の分類基準である[*7]．d を大きくすると適合率が上がり・再現率が下がる，d を小さくすると適合率が下がり・再現率が上がる傾向が予測される．実際に 3.2 節の仮定法と分詞構文の実験データに対して，d を変化させた場合も概ねそのような傾向となることが確認される（図 1,2）．

コーパスの構築で作業担当者がチェックすることを前提とすれば，各項目の再現率を最重視すればよい．再現率が 100%となるまで d を下げていくと，仮定法では -0.14 周辺となる．このとき適合率は 5.6%と極めて低いものの，その際チェックすべき文数は 1,253 文で約 26.1%に削減されたことになる．同様に分詞構文についても d を下げていくと，-0.13 周辺で再現率が 100%となる．このとき適合率は 5.2%とやはり低いものの，チェックすべき文数は 3,034 文で約 58.7%に削減されたことになる．

[N-Cube] で記述しているような表層・品詞レベルでも，このような削減効果は期待できるので，この削減率は一概には評価できない．ただ，初期の粗いルールでも少なからず作業の効率化には寄与すること，そして今後のデータの充実化や前節で述べたようなルール（分類器）の作成過程の改善によって，より高い効果も期待されるものと考えられる．

4. おわりに

本稿では，学校英文法コーパスの構築の試みについて述べた．現在取り扱っている文法項目は [Royal 00] 等で規定される学校文法の項目のごく一部であり，今後，随時再検討しつつ，データの整備も進めていく予定である．また，文法項目の検出ルールは構文レベルで記述したが，文法項目によっては，構文レベルまで考える必要はなく，むしろ表層・品詞レベルの方が望ましい場合もある．したがって，文法項目の性格を踏まえた，検出ルールのレベルの設定法や記述法も考えていく必要がある．

*7 正確には等号成立時は，unknown またはランダムに $+1, -1$ を選択する．

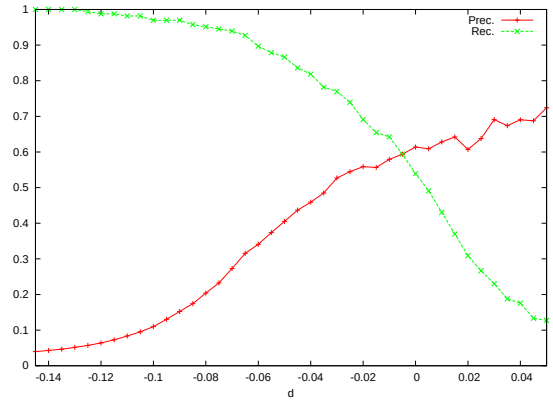

図 2: d と分詞構文の適合率（Prec.）・再現率（Rec.）

立命館大学言語教育情報研究科 池田洋子さん，同大学文学部 加藤孝幸さん・北野宏樹さん・高屋敷大さんには，プロトタイプの文法項目の付与に尽力頂いた．ここに記して感謝の意を表する．

本研究の成果の一部は，2007 年度立命館大学学内提案公募型研究推進プログラム・基盤的研究（課題番号: 30），文部科学省科学研究費補助金・若手研究 (B)（課題番号：19720149）および日本学術振興会科学研究補助金・基盤研究 (C)（課題番号:20520504）によるものである．

参考文献

[BACT] 工 藤　拓：BACT: a Boosting Algorithm for Classification of Trees, `http://chasen.org/~taku/software/bact/` .

[Kudo 04] Kudo, T. and Matsumoto, Y.: A Boosting Algorithm for Classification of Semi-Structured Text, *EMNLP 2004* (2004).

[N-Cube] 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/` .

[PTB] Penn Treebank Project, `http://www.cis.upenn.edu/~treebank` .

[Royal 00] 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).

[Sano 00] 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).