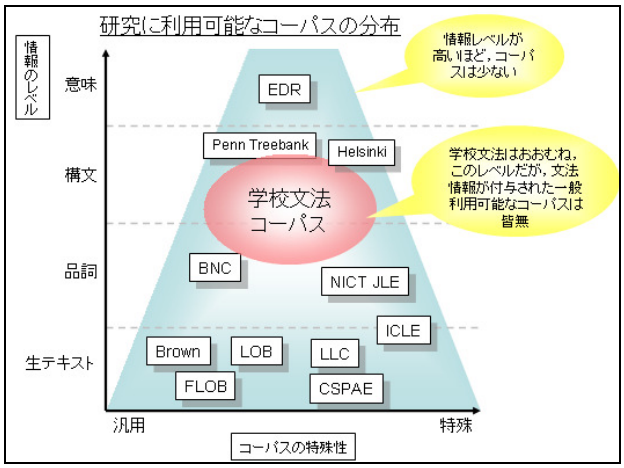
学校英文法コーパスの提案
―デザインと応用可能性―

小林雄一郎 [1*] 田中省作 [2] 後藤一章 [3] 徳見道夫 [4] 朝尾幸次郎 [2]

[1] 法政大学 [2] 立命館大学 [3] 大阪大学 [4] 九州大学
* kobayashi0721@gmail.com

1. はじめに

近年，大規模言語データベース(コーパス) が言語研究にパラダイム・シフトを起こしている．それまでは主観的分析が主流であった言語学研究において，コーパス・情報科学の融合は，データに裏づけされた客観性の高い研究を可能にし，同時に言語教育や情報科学分野の新たな知見を明らかにしている．特に，英語は言語資源が最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[1]．しかしながら，そのような英語にあっても，我々が知る限り，研究など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)は存在しない．



学校文法は，英語の構造を説明するだけでなく，英語を母語としない多くの日本人にとって重要な理解の観点の一つであり，英語教育分野においても，学校文法に関する情報が付与されたコーパスのニーズは極めて高い．そこで，本研究の目的は，英語を対象言語として，英文の学校文法レベルの情報を付与したコーパスの構築と公開にある．

2. 学校英文法コーパスの構築

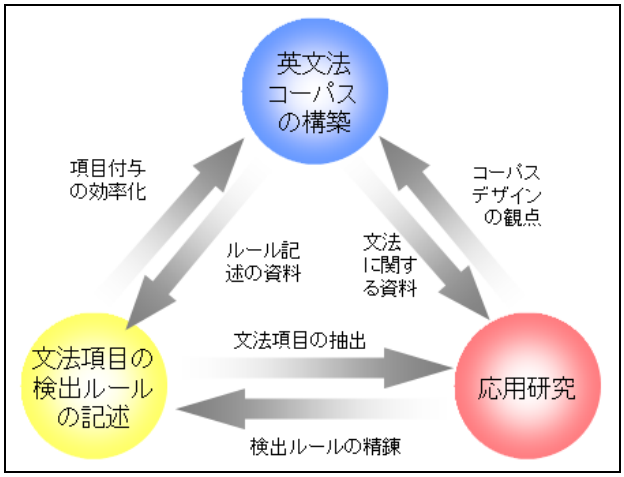
2.1. 背景

学校英文法コーパスに関わる研究として，[2]がある．[2]は，学校文法項目について中学高校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たものである．「文法項目別 BNC 用例集―N-Cube」は，[2]に基づいて，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．[1] それらを実装したシステムは，British National Corpus (BNC) 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．しかしながら，これは，あくまで用例抽出を主目的とするものであり，表層・品詞レベルの記述力に限界があり，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代わるものではない．こういった用例抽出の精度を保証するという意味でも，本研究が構築しているコーパスの必要性は高い．

[1] http://scn02.corpora.jp/~n-cube/

2.2. 構築方針

学校文法の項目には，文字列や品詞列のレベル，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する一方で，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も並行して記述する．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，教材評価などの応用研究への適用可能性も判断しやすくなる．



以下，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約 4000 文である．文法項目については，網羅的に設定するのではなく，[2]をベースに日本人英語学習者の英文理解に強く関わるとされる項目を優先している．日本人の学習過程を意識する理由の一つは，日本人英語の分析などへの応用を強く志向しているためである．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，いま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．

表 1: 文法項目

文型	1-5 文型
文の種類	平叙･疑問･否定･感嘆
文の単複など	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
話法	直接･間接
時制	現在･過去･未来
時制の一致	一致･不一致
態	能動･受動
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	
独立不定詞	
比較(形容詞)	原級･比較級･最上級
比較(副詞)	原級･比較級･最上級
分詞	限定･叙述
動名詞	
助動詞	
接続詞	等位･従属
疑問詞	
疑問詞 ＋to＋V	
関係代名詞	主格･目的格･所有格
関係代名詞の種類	制限･継続
関係副詞	
複合関係詞	代名詞･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	主語の一致･主語の不一致，慣用表現
強調構文	

(個々の項目の定義については，[3]を参照)

3. 学校英文法コーパスの活用

3.1. 言語処理との関連

学校英文法コーパスのプロトタイプが整備されたならば，文法項目ごとに，どのような単語列や品詞列・部分的な構文構造の下で成立するのかを，人手だけでなく情報科学の機械学習に基づいて，文法項目の検出ルールを完備する．これにより，任意の英文とその構文構造から，その英文の文法項目を網羅的に検出できることになる．これらのルールは，学校英文法コーパス自体でその精度を検証することができ，精度保証されることとなる．そして，文法項目ごとの検出ルールが整備されれば，様々なテクストに一定の精度で文法情報を付与することができ，コーパスの大規模化・多様化が可能になる．

文法項目の自動検出に関しては，パイロット・スタディながら，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）を用いて，仮定法や分詞構文が一定の精度で同定できることがすでに報告されている[4]．

3.2. 言語学との関連

語彙研究や音声研究など，言語学には多くの下位分類が存在するが，今も昔もその中心は文法研究である．また，[5]の出版以来，とりわけ英語学の分野では，コーパスに基づく記述的な文法研究が盛んである．しかし，[5]で用いられている Longman Spoken and Written English (LSWE) Corpus は，ゼロ代名詞のような項目にも情報付与がなされている点で非常に画期的なものだが，一般には公開されていない．また，[5]における文法項目の量的分析は，基本的に品詞・構文情報とn-gram抽出に基づくものである．例えば，本研究で構築しているコーパスからは不定詞の用法別頻度を抽出することが可能だが，[5]では，“verb + *to*-clause”や“verb + for NP + *to*-clause”のような連鎖パターンが数量化されているだけである．より詳細な情報が付与されたコーパスを用いて，個々の文法項目に関する頻度や共起関係を調査することによって，これまで見過ごされてきたような事実が得られるに違いない．

また，コーパスによって導かれる新たな知見は，辞書編纂にも大きく貢献するであろう．現在の辞書編纂において自動化されているのは，単語レベルの頻度や共起関係の抽出までで，人手によるコンコーダンス・ラインの吟味が依然として作業の中心を占めている[6]．現在，*COBUILD* や *LDOCE* のような辞書には単語の頻度情報が明示されている．文法情報の自動抽出が可能になれば，用例の記述も今より充実するであろう．

3.3. 言語教育への応用

言語教育への応用例は，大きく分けて二つある．一つは，教材の開発および評価である．英語教育の分野では，検定教科書をはじめとする教材コーパスの整備が進められており，韓国・中国・台湾のような近隣諸国の教科書との量的比較がなされている[7]．現時点では語彙レベルの比較が中心だが，文法レベルの比較も非常に興味深いものである．また，昨今の「多読」ブームにより，レベル別に語彙や文法を制限した Graded Readers (GR) も注目を集めている．このような副読本のコーパスも既に整備されているため[8]，GR における文法項目の導入過程に関する調査も可能になる．さらに，英語教育の現場では，文法項目が自動的に検出できることによって，英語教師による授業教材やテストのための用例検索が容易になるであろう．

そして，二つめの応用例は，学習者コーパス研究である．学習者コーパスとは，外国人学習者によって産出された言語資料を機械可読な形式で集積したものである[9]．日本語母語話者による英作文を集めたコーパスとして，International Corpus of Learner English (ICLE) の日本版[10]、Japanese EFL Learner (JEFLL) Corpus [11]、Nagoya Interlanguage Corpus of English (NICE) [12]、Corpus of English Essays Written Japanese University Students (CEEJUS) [13] などが知られている．これらのコーパスにおける文法項目の数量化が可能になれば，日本人母語話者による学校英文法の習得過程が明らかになり，彼らの苦手な表現を特定することもできるようになる．

4. おわりに

本稿では, 学校英文法コーパスのデザインと応用可能性について述べた. 現在取り扱っている文法項目は, [3]などで規定される学校文法の項目のごく一部であるため, 今後も随時再検討しつつ, データの整備を進めていく予定である. なお, コーパスは 2009 年春に公開予定であり, 文法項目の検出ルールなども随時公開していく予定である.

謝辞

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

[1] McEnery, T., R. Xiao, and Y. Tono (2006) *Corpus-Based Language Studies: An Advanced Resource Book.* London: Routledge.

[2] 佐野洋・猪野真理枝 (2000)「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』No. 117 (pp. 5-12).

[3] 綿貫陽・宮川幸久・須貝猛敏・高松尚弘 (2000)『ロイヤル英文法』(改訂新版) 旺文社.

[4] 田中省作・小林雄一郎・徳見道夫・朝尾幸次郎 (2008)「学校英文法コーパス構築の試み」『2008 年度人工知能学会全国大会(第 22 回)論文集』CD-ROM.

[5] Biber, D., S. Johansson, G. Leech, S. Conrad, and E. Finegan (1999) *Longman Grammar of Spoken and Written English.* Harlow: Pearson Education.

[6] Rundell, M. (2008) "Recent Trends in English Pedagogical Lexicography." T. Fontenelle (ed.), *Practical Lexicography: A Reader.* Oxford: Oxford University Press (pp. 221-243).

[7] 小池生夫 (編) (2008)『第二言語習得研究を基盤とする小、中、高、大の連携をはかる英語教育の先導的基礎研究』平成 16 年度～平成 19 年度科学研究費補助金 (基盤研究(A)) 研究成果報告書.

[8] 小林雄一郎 (2006)「Graded Readers における語彙の統計的分析」『第 45 回(2006 年度)JACET 全国大会要綱』(pp. 47-48).

[9] Granger, S. (ed.) (1998) *Learner English on Computer.* London: Longman.

[10] Kaneko, T. (2004) "What are ICLE and LINDSEI?" *Handbook of an International Symposium on Learner Corpora in Asia* (pp. 66-68).

[11] 投野由紀夫 (編) (2007)『日本人中高生一万人の英語コーパス"JEFLL Corpus"―中高生が書く英文の実態とその分析』東京: 小学館.

[12] 杉浦正利・阪上辰也・成田真澄 (2007)「英語学習者コーパスにおける作文テーマの影響―英語母語話者コーパスとの比較分析」英語コーパス学会第 29 回大会 (2007 年 4 月 28 日、同志社大学) 研究発表資料.

[13] 石川慎一郎 (2008)『英語コーパスと言語教育―データとしてのテクスト』東京: 大修館書店.