BNC を利用した英語教材作成とその提供 Web サイト

須藤武文† 佐野 洋† 中村隆宏‡

†東京外国語大学 外国語学部 〒183-8534 東京都府中市朝日町 3-11-1
‡小学館コミュニケーション編集局電子辞書編集室 〒101-0051 東京都千代田区神田神保町 2-30 昭和ビル 4F
E-mail: †sano@tufs.ac.jp, ‡takahiro@shogakukan.co.jp

1. まえがき

1.1. 研究目的

本稿は，コーパスを利用した英語教育用教材の自動作成の方法と，作成した英語教材を，ネットワークを通じて学習者へ提供するための教材素材ウェブサイトの開発について述べる。

筆者等は，ことばの 4 技能・能力(「読む」，「書く」，「聞く」，「話す」)全てに関係する「文法」能力の効果的な育成を支援する教授法と教育教材の開発を目指している。この目的のため，中学校・高等学校で利用される全種類の英語教科書を調査し，英語教育課程で学習されている英語文型を網羅的に調査した。その結果を基に，(株)小学館マルチメディア局との共同研究により，BNC(British National Corpus)から文型パターン検索を行った。抽出した英文用例に文型パターンに対応する文法項目情報等を付加し XML 化した。そして XML データから，インターネットを通じて英語教育素材を提供するウェブサイトを構築した。本サイトは，小学館コーパスネットワーク(SCN)のサービスとして提供される(平成 17 年 2 月を予定)。

1.2. 背景

情報通信基盤が地球規模で整備され，高度なネットワーク化が進展し，様々な情報がネットワークを通じて伝搬するようになってきた。こうした状況に対応してゆくには，企業のみならず，行動主体である個々人がネットワークを活用する担い手にならなければならない。

グローバル化によるビジネス分野における業務等の国際展開と経済活動のボーダレス化に伴って，専門的な分野で活躍する英語力を持った人材が求められている。近年の日本における英語教育の改善・質的向上は，教育課程における問題に留まらず，上述の環境変化に随伴して社会的な要請強い政策課題になっているのだ。例えば，昨年度(平成 16 年度)，文部科学省が募集した「現代的教育ニーズ取組支援プログラム」のテーマのひとつには、『仕事で英語が使える日本人の育成』が挙がっている。

金融・IT・医療などを中心に専門分野の英語力向上を目指した学習は，英語教育分野でも必要性が指摘され，企業内教育でも ESP 用教材の需要が高まっている。そして ESP を指向する英語教育の改善と質的な向上には，教育効果を向上する仕組みが必要であると考えられる。

1.3. アプローチ

従来の教材は，言語知識(いわゆる語彙・文法)の学習と理解と，言語運用知識(いわゆる会話・対話)の学習と習得が個別に行われていた。言語知識における教育項目は，言語学的な抽象化が過ぎていて，一方，言語運用知識における教育項目は，発話状況に依存し過ぎたスキルの訓練に偏っている。

筆者等は，教育効果を向上する手段として，学習時間を効果的に使うことのできる教材の開発を目指している。

発話状況毎に，発話の中で使われる表現は，現実の状況(言語運用知識の利用)で使用された(言語知識から生成された)表現が対応する。すなわち教材内容が言語知識だけを含むのではなく，同時に言語運用知識も含むようにすればよい。現状の学習項目規模に合わせ，学習の効率化に結びつく教材を開発するほうが，学習者への負担は軽い。実際に使われる表現を教材に用いることが成果の効率向上につながるものと考える。

上記の教材は，特に ESP 教育には適しているだろう。これらの教材の特徴は，専門分野の語彙と文型が意識されていること，リーディングとライティング能力の育成・向上に焦点あること，学習者の言語運用目的と到達

目標に応じた内容を持つこと, 効率的な学習が可能であることなどである。

学習の効率化に加えて，この取り組みの背景には，(1) 教材作成が労働集約的な作業であって, 短時間に多様で且つ大量の教材が作成できないこと，(2) 作成された教材内容の品質を, 教授者の経験や学習量といった属人的力量に依存せずに決めたいこと，(3) EFL を前提とすると，作成者の多くは英語母語話者ではないので, 英文表現に多様性が乏しいこと, など例文の品質が定性的に保証できない問題がある。こうした問題の解決も視野に入れている。

1.4. Web サイトの位置づけ

筆者等はこれまで，多様な学習要求に適合する語学教育方法論(N-Cube)を研究してきた[1,2]。N-Cube は，ESP(English for Specific Purposes)適合の教材作成の効率化とその教授法確立を目指した語学教育支援枠組みである。この枠組みは，(1) 教育メソッドとして, 認知力と母語の運用能力を活用することを特徴にした Cognition-Based Rule-Driven/Data-Driven (RD/DD) Interactive Learning 方法論の研究と，(2) 学習支援として, 言語運用データと自然言語処理技術を使った効率的な教育教材作成を目指すものである。

図 1 は，研究目的に対応する全体の工程を示している。本稿でカバーする事柄は，図中の (3)，(5)，(6)，(7) である。(4) の LTB(Language Tool Box)は, (株)小学館が開発したコーパス利用のためのワークベンチである。詳細については，[5,6,7]を参照されたい。

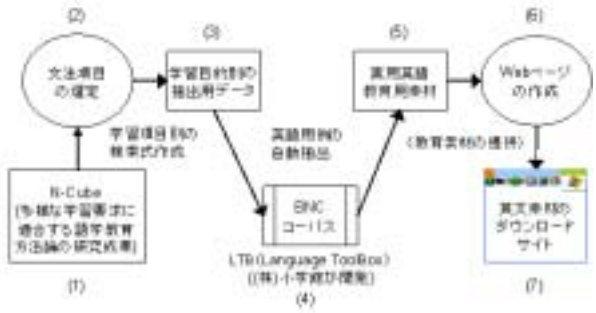

図 1. 全体構成図

我々は，教科書の学習項目を文型パターンとして(3)，LTB を利用し，BNC から文型パターンを含む用例を抽出した(5)。抽出した英文用例に文法項目情報等を付加し, それらデータを XML 化した上で，インターネットを通じて英語教育素材を提供するウェブサイトを構築した(7)。この Web サイトは，小学館コーパスネットワーク(SCN)のサービスとして提供される。

以下，2 章では，教科書の学習項目に対応する検索式の作成と，LTB を使った BNC からの用例抽出について説明する。3 章は，教育素材提供のための Web サイトの構築について説明する。4 章で本稿をまとめる。

2. 例文自動抽出の方法

2.1. 文法項目の選定

日本国内で販売される英語教科書に関する市場調査(売り上げ高)を基準に選択した 31 種の英語の教科書(中学校英語教科書：6 種, 高校英語教科書：英語Ⅰ：8 種, 英語Ⅱ：8 種, ライティング 9 種), 及び日本の英語教育で広く使われている参考書 4 種を対象に文法項目を調査した[1,2]。

2 点の選択指針, (1) 主要教科書に共通して現れる項目であること，(2) そうでない場合，(1)の条件を緩和し，各教科書を幅広くカバーすると同時に, 言語運用規則として生産性の高い項目であることに基づいて 144 の文法項目を整理した。この 144 項目は，"I am+名詞"といった単純な文法項目から，"SVC"，"強調構文"，そして"仮定法未来"などの難しい文法項目まで含む。

144 の文法項目は肯定文を基本とする。そして各文法項目に対し, 否定文や疑問文など 14 の下位項目を設けた。144×14(2016)項目から，英語構文として存在しない 634 のパターンを除き, 総計で 1,382 文型を整理した。

2.2. 検索式の作成と用例抽出

1,382 の文型に対応する検索のための文型パターン(CQL[i]式)を作成した。例えば「how を使った感嘆文」の場合，CQL 式は次の検索式になる。

```
^{W="how"} {P="AJ0|AV0"} [0,10] {L="!"}$
```

この式は,『"How"という単語で文が始まり，形容詞もしくは副詞が続き，0 個以上 10 個以下の単語を間に挟んで"！"で終了する文を検索』するパターンを意味する。このような検索式を使って，1382 文型にすべてについて，文型パターンを作成し，LTB を用いて BNC から英文用例を抽出した。

2.3. 用例評価と検索式の精緻化

教育教材の品質を確保するために検索式の精度向上を行った。

LTB で利用する BNC は，形態素解析結果までの情報を参照することができる。CQL 式を利用して，特定できる言語的な情報は，一つの語について，表層形，基底形(辞書形)，そして品詞分類の 3 つである。

例えば，品詞分類では，自動詞と他動詞の区別がない(BNC の仕様)。そのため，SVO(O=that～)『目的語に that 節を取る他動詞構文』文型は，V に対する他動詞品詞の指定ができないにも関わらず，O の表層形の指定が可能なために，CQL 式で検索すると，93%の精度で用例を抽出することができるのだ。

それに対して，SVOC 構文『目的語に名詞をとる構文』文型は，V の品詞指定ができず，O が名詞であることを頼りに CQL 式で検索するために，50%程度の精度でしか用例を抽出できない。

構文の抽出には，本来は構文解析結果を用いて行うべきであろうが，1 億語という大規模コーパスが利用できる利点があること，形態素解析の精度に比べ，構文解析結果の精度は低く，抽出精度が向上しても解析精度が高くなければ結果としての抽出精度は上がらないことなどから，我々は解析精度の高い形態素解析結果を基にして，(1) 検索式を作成し，用例を抽出する，(2) 抽出用例を評価し，検索誤りが生じている語連鎖を特定する，(3) 評価結果をもとに，誤り部分の語連鎖だけを検索する減算式を作成する，(4) 減算式を含めて再抽出を行って誤用例を正用例から除いた用例を得た。抽出用例の精度向上を示す概念図を図 2 に示す。赤い部分が，減算後の用例抽出結果になる。

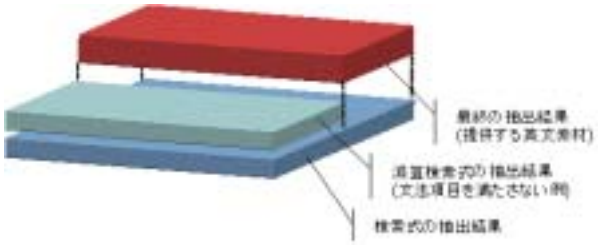

図 2. 用例評価と精度向上の概念図

3. 教材提供 Web サイトの構築

3.1. サイトデザイン

BNC から用例を抽出することなどから，英語教育素材を提供する Web サイトの利用者として，筆者等は，もっか中学生，高校生や大学生に英語を教授する立場にある人を想定している。文型に対応する用例データは，恐らく利用者の要求として，補助資料やテストなどを作成する際の英文参照や，作為例文のためのサンプル利用が考えられる。

教育現場で利用される英語教科書においては「現在完了」や「使役構文」といった学習すべき文法項目の名称が目次として強調されていることが多い。利用者がキーワードとして探しやすいのは，この文法項目名であると考え，まず文法項目の一覧から目的の項目を選び，次に肯定文や否定文といった下位の項目を選択する，という二段階のステップを踏んで目的の例文に到達する構造とした(図 3)。

また、図 3 に示すように同じ文法項目でも教科書によって多少表現や記載内容が異なることもあるので，文法項目名の定義を明確化するため，文法項目を選択した段階で，各項目について各種教科書で使われている用語を用いた文法説明データを表示できるようにデザインした。

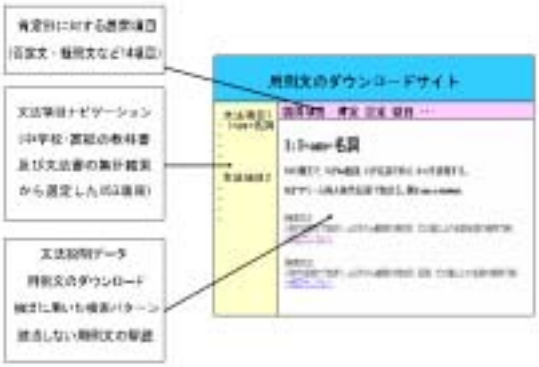

図 3. Web デザイン

3.2. サイトの構築

以下に構築した Web サイトを示す。

図 4. Web サイトの様子

図 4 の画面は，実際に文法項目と下位項目を選択し，例文を表示させた状態である。縦軸に配置される文法項目の数は 144 あるが，利用の便を考えて 1～最大 14 個の単位でまとめられている。また BNC から抽出される例文の数は項目によって寡多があるため，例文数は 100 を最大とし，それを超える項目については抽出された例文から無作為に 100 文を抽出して掲載している。

4. 今後の課題

ダウンロードサイトの公開を行い，利用者評価を実施する。必要文法項目とその数，１文法項目あたりの必要用例数，文長や語彙レベルの制御等の評価が必要である。すでに語彙フィルタ(小学館プログレッシブ基本語水準，JACET8000 語水準)の作成を終え，現在，抽出実験を行っている。語彙フィルタリングした用例集を使った Web サイトも提供する予定にしている。利用者の要求はサイト改善に反映される。

謝辞

本研究は以下の助成を受けた。
(1) 平成 14-16 年度文部科学省科学研究費(基盤研究(B)(2)) 「全電子化検定済み教科書データの解析と大規模日本語コーパスの構築」 (研究代表者：佐野洋)
(2) 平成 15 年度 (株) 小学館・マルチメディア局委託研究

文　献

[1] 佐野洋：「ESP 適合の教材コンテンツを実現する語学教育支援システム」,『最新外国語 CALL の研究と実践』, コンピュータ利用教育協議会(CIEC)・外国語教育研究部会(34～44,10 頁),2003 年 3 月.

[2] 佐野洋，猪野真理枝，宇野陽一郎：「多様性適合の学習環境を実現する語学教育支援システム」,情報処理学会，情報学シンポジウム講演論文集(55～62,8 頁),2002 年 1 月.

[3] 新井 雅之，渡辺 亜美，佐野 洋：「言語運用に基づく英語教育教材とその提供 Web サイト開発」,教育システム情報学会第 29 回全国大会講演論文集，pp.257-258，教育システム情報学会,2004 年 8 月.

[4] 岩倉隆幸，新井雅之，佐野洋：「言語運用データを使った英語教育教材の作成」, FIT2004 第 3 回情報科学技術フォーラム,2004 年 9 月.

[5] Nakamura, T. and Tono, Y. (2003) Lexical profiling using the Shogakukan Language Toolbox. In Murata, Yamada & Tono (eds.) ASIALEX 2003 Proceedings. Dictionaries and Language Learning: How can Dictionaries Help Human & Machine Learning?, pp. 170-176.

[6] Nakamura, T., Tateno, J. and Tono, Y. (2004) Introducing the Shogakukan Corpus Query System and the Shogakukan Language Toolbox. Williams, G. and Vessier, S. (eds) EURALEX 2004 Proceedings . The Eleventh EURALEX International Congress, July 6-10, 2004, Lorient, France, pp. 147-152.

[7] 中村隆宏 相澤弘 渡辺亮嗣 (2004)「自然言語文の検索方法および検索装置」特許出願 特願 2004-047377

[i] LTB で利用するコーパス検索言語(Corpus Query Language)である。表層語，辞書形，品詞分類名の三つ組で一語を表現しその後連鎖で文型パターンを表す。文頭・文末や任意語数など正規表現ライクな記述も可能である。