K-055

言語運用データを使った英語教育教材の作成
Usage-based Materials for English Teaching

岩倉隆幸† 新井雅之‡ 佐野洋‡
Takayuki Iwakura Masayuki Arai SANO, Hiroshi

1. はじめに

我々の研究の目的は、(1) 言語運用データに基づく英語教育用教材の作成と、(2) これら英語教材をWeb上で共有し、ネットワークを通じて英語教材を提供する教材データバンクの開発にある。本稿では、コーパス(BNC)からの英語教材素材の自動抽出の方法と、抽出された英文の評価(自動抽出の精度と、教育素材化のための判断データの作成)について述べる。

1.1 背景

企業内教育・研修の効率化やコスト軽減のためにWebを利用したe-Learningシステムの活用が盛んである。e-Learningシステムを始めとした情報技術は、学習場所と学習時間の制限をなくしている。学習行為の分散化(必要な事柄を必要な時に学習すること)が進んでいる現在、学習内容の制約の解消、すなわち規格化された従来教科から、多様な教科の実現が求められている。専門分野の英語(ESP : English for Specific Purposes)教科もその一つである。

ビジネス分野では業務の国際展開に伴い、それに対応するため、専門的な分野で活躍できる英語力を有する人材が強く求められている。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認識され、企業内教育ではESP適合の英語教育教材の要望が高まっている。

1.2 取り組み

一方で、教材開発は依然労働集約的な作業であり、作成コスト削減の取り組みが必要である。我々は、電子化テキストやコーパスと自然言語処理技術を応用することで、専門分野に合わせた英語教育教材(ESP用教材コンテンツ)を、その作成コストを低減しながら効率良く作り出すことを目的に語学教育支援システムを研究開発してきた[1,2]。

筆者等は、これまでに個人に適合する外国語教授の方法論を構築した[2]。さらに、分野別の英文データ(英語コーパス)から、文法項目や語彙レベルを主なパラメータとして、学習者のニーズに適合する英語教育素材を自動抽出するN-Cubeシステム(パイロット版)を構築した[1]。2003年度、小学館・マルチメディア局と共同研究を開始し、言語運用データに基づく英語教育用教材の作成を行ってきた。

2. 用例抽出の手順

英語教育用教材の作成にあたって、収集対象となる用例に含まれるべき英語文型の一覧を作成した。文型選択のため、中学・高校で使用されている英語教科書を分析し，文法項目の一覧を整理した。この文法項目を基に、英語コーパスから例文を抽出するツールである小学館 LTB(LTB: Language Toolbox)を利用して、BNC(BNC: British National Corpus)から、各文法項目に対応する英文(教育素材)を抽出した。

2.1 抽出対象項目

一般流通する教科書の、市場において各教科書が占めるシェアを基準として、分析対象の英語教科書を収集した。

日本の中学校・高校で使用されている英語の教科書の市場において合計のシェアが50%を超える上位31種の英語の教科書を収集した。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種である。収集した各教科書名，及び占有率については表1に挙げた。

文法項目は、以下の二点の方針に基づいた教科書分析を通じて行なった。(1) 収集教科書に共通して現れる文法項目であること、(2) 共通して現われない場合，収集教科書を幅広くカバーする項目であること、である。この分析で153の文法項目を整理した。文法項目は肯定形を基本とし、各文法項目に14の下位項目を設定した。整理した153の文法項目を表2に挙げる。下位項目は表3に挙げている。

153の文法項目に対し14の下位項目を設定した場合、計算上、2142文型になるが、命令文や感嘆文のように14の下位項目が整わない文法項目があるため767の文型を除外した。その結果、文型の総計は1395である。

2.2 検索式の作成

1395文型に対して、小学館LTB用の検索式であるCQL式(CQL: Corpus Query Language)を作成した。CQL式を使うと、一つの単語に対し表層形・基底形・品詞タグの三つ組みの指定を行うことができる。そして、その連鎖で文型を特定する。例えば、二重目的語構文などの文型対応式では、検索式の段階で予め動詞を指定することもできるが、動詞を特定せず連続する品詞を限定して検索することでより幅広く例文を抽出することが可能になる。また、文頭や文末の位置指定・キーワードとなる単語間の語数の指定ができる。

しかし、文法項目の表現の抽象度が高い場合などは、CQL式による表現能力には限界があり、意図した文法項目に該当しない例文を抽出する可能性がある。例えば品詞タグを使って文型を特定するCQL式の場合、BNC品詞タグには自動詞/他動詞の区別がないため、通常のSVO構文と補語に名詞句を取るSVC構文が両者とも同一の文型として認識され、正しく抽出できない(このように誤って抽出した文を本稿では「ノイズ」と呼ぶ)。

検索の結果、1395の文型に対して計18228374文の用例が抽出された。この検索用例にはノイズも含まれている。

† 東京外国語大学 大学院地域文化研究科

‡ 東京外国語大学 外国語学部

筆者等は、次にノイズの分析を行い、教育素材化のための判断データの作成を行なった。

3. 抽出結果の評価

本章では、抽出された英文の評価、すなわち自動抽出の精度と、教育素材化のための判断データの作成について述べる。具体的には、抽出された用例が検索式で意図した文法項目にどれほど適合しているかを確認し、適合していない場合、どのような理由で抽出されたのかを分析した。この評価を通じ、検索式の精度を検証すると同時に、抽出された例文の品質(定性分析)を明らかにする。

直感的に予想される通り、文法項目ごとに抽出される用例数は異なる。例えば「主語＋一般動詞(＋前置詞)」という文法項目ではBNCから300万件を超える例文が抽出される。その一方で、「過去完了進行受動態」(1件)、「未来進行受動態」(3件)のように極めて少ない用例数の文法項目もある。抽出した用例数が多い文法項目については、人手で調査可能な分量を超えているため、各文法項目の抽出用例数に対し100例文をランダム抽出して分析を行った。同時に、抽出誤りで得られた用例の定性分析も行った。

3.1 評価対象の文法項目

文法項目番号70～84、90~97の計23項目を評価した。

これら評価対象の文法項目は、検索式の表現(表層形、基底形、品詞タグ)のみでは十分な抽出精度が出ず、抽出用例中に多数のノイズを含んでいると推測されるものである。なお、残りの文法項目については引き続き同様の評価を実施する予定である。

評価では、各文法項目のサンプル用例数100について、
(1) 各用例が該当する文法項目に適合しているかどうか、
(2) ノイズ用例の場合、どのような理由で抽出されたのか、
(3) 100用例に対するノイズ文の比率
を調査した。この評価作業は、本学の英語科専攻の学生を中心に実施した。

3.2 定量分析

抽出された用例中のノイズには、大きく分けて2種類のノイズがあることが分かった。すなわち、各文法項目に共通して現れる遍在的なノイズと、文法項目の構文に依存して現れるノイズである。遍在的なノイズは、どの文法項目の用例においても10%前後の生起率を示す。それに対して構文に依存したノイズの生起率は文法項目によって違っている。構文に依存したノイズの詳細なデータは、各文法項目で提示しなければならず、紙幅の都合で本稿には載せない。機会を別にして示したい。

遍在的なノイズは、以下の3グループに分類することができる。すなわち、(1) 抽出対象のデータ記述(BNC)に原因があるもの、(2) 検索式に起因するもの、(3)その他のものである。

3.3 定性分析

3.3.1. (1)のタイプ

a) 品詞タグのミス

LTBツールでは、BNCコーパスのデータは、タガーによって単語に品詞ラベルが付与されている。ラベル名が間違っていることがある(5%くらいの割合で間違いがあると指摘されている)。そのため、検索式に品詞を指定する部分がある場合、間違った品詞タグの用例にヒットする場合がある。

b) 単語の切れ目がおかしい

ごくまれにある間違いである。文中の単語の切れ目がおかしいため、該当しない例文が抽出されてしてしまうことがある。例えば、以下のような文である。

(HTM 1808) I feel so old ? y I was his sister in the story, why do If eel so old, so cold ?

3.3.2. (2)のタイプ

a) 固有名詞が抽出対象となっている

人や会社の名前、本のタイトルなど、固有名詞が間違って抽出されることがある。

b) 場所を表す名詞句が副詞として機能している

時や場所を表す名詞は単独で副詞として用いられる場合がある。例えば、this week、 last week などである。機能上は副詞だが表面上は名詞であるため、主語や目的語と間違って抽出してしまうことがある。これは遍在的なノイズにするか構文に依存するノイズにするか判断が難しい。

3.3.3. (3)のタイプ

a) 例文が短すぎて内容がない

S+be動詞という項目に対して、"I am." という例文のように、形式の基準は満たしているが、内容がなく学習用教材としては不適当である。

b) 口語のデータであるため、構文を確定できない

小学館LTBツールには、コーパス抽出のための様々な機能が用意されている。その中にKWIC整形機能がある。KWIC整形したデータの文頭にある記号は、3桁のアルファベットと数字の連続である。3桁のアルファベットはデータの出典を表し、特に最初の文字がJもしくはKのものは、口語のデータであることを表す。後続の数字は行番号を表す。これらに該当する口語のデータは、場合によって、おかしな位置で文が途切れている。そのために構文を確定できない、繰り返しや言い直しの一部であるなどが原因で誤って抽出されることがある。

c) 主語がない

口語のデータに多い。主語が省略されている場合がほとんどであろう。

d) 破格の構文(構文の確定不可)

前後の文脈が不十分で構文を確定できないものと重なる場合もある。例文が非文法的である場合、「破格の構文であるため、該当する例かどうか判断できない」として取り扱った。

構文に依存するノイズは、主として検索の対象とする構文の形式上の特徴、もしくは作成した検索式に原因があって生じることが明らかになった。構文に依存するノイズを分析することで、サンプリング調査した文法項目の持つ特徴が明らかになると同時に、検索式の性能および問題点、そして改良の方向性を具体的な形で捉えることが可能になった。

4. 今後の課題

筆者等は、31種の英語の教科書を調査し、153の文法項目を選定し、下位項目14を設けた上で、1395の文型に対応する検索式(CQL式)を作成した。小学館LTBを利用し、

BNCから、用例抽出を行った。検索式の能力制限等のことから、我々は検索結果をサンプル評価した。

評価結果から、(1) すべての文法項目に対して偏在的な抽出誤りが 10%あることが判明した。(2) 構文に依存する抽出誤りには、その度合いに違いがあり、本稿では、その定性分析を説明した。定量的な誤り率は、別の機会に示す予定である。

この結果は、教材データバンクの各教材に付随する説明データになる。利用者は、この評価結果を参照することで、英文教材素材の品質を確認することができる。また、同時に評価結果は、CQL式を改善するための基礎データとする予定である(減算式という考え方で抽出結果を改善することができる)。改善された検索式は、おそらく構文に依存するノイズを減らすことができるだろう。それによって、言語運用データ(BNC)を使った、更に信頼性の高い英語教材データを作成することが可能になる。また、個々の文法項目の誤り分析に伴う定量分析結果は、新たな英語教育に対しての知見を提供するだろう。

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

参考文献

[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

表 1. 調査教科書一覧と市場占有率

中学校教科書			
占有率順位	発行者	教科書名	2002年度出版社別占有率(%)
1	東書	New Horizon1,2,3	41.1
2	開隆堂	Sunshine1,2,3	22.6
		計6冊	計63.7
高校教科書(英語I)			
占有率順位	発行者	教科書名	2002年度出版社別占有率(%)
1	東書	New Horizon English course I	12.5
2	三省堂	The CROWN English Series I	9.8
3	三省堂	VISTA English Series I	6.9
4	文英堂	UNCORN ENGLISH COURSE I	6.2
5	啓林館	MILESTONE English Course I	4.1
6	第一	Evergreen English Course I	3.9
7	桐原	SPECTRUM ENGLISH I	3.9
8	開隆堂	Sunshine ENGLISH COURSE I	3.6
		計8冊	計50.9
高校教科書(英語II)			
占有率順位	発行者	教科書名	2002年度出版社別占有率(%)
1	文英堂	UNCORN ENGLISH COURSE II	14.5
2	三省堂	VISTA English Series II Revised Edition Step1,2	9.9
3	啓林館	MILESTONE English Course II	8.9
4	数研	POLESTAR English Course II	4.4
5	東書	New Horizon English course II	3.5
6	三省堂	The CROWN English Series II	3.4
7	大修館	Genius English Course II	3.3
8	大修館	CLIPPER ENGLISH COURSE 2	3.3
		計8冊	計51.2
高校教科書(ライティング)			
占有率順位	発行者	教科書名	2002年度出版社別占有率(%)
1	文英堂	SECOND EDITION POWWOW WRITING	9
2	啓林館	REVISED MILESTONE English Writing	7.5
3	数研	Revised POLESTAR Writing Course	6.5
4	開拓	NEW ACESS to English Writing New Edition	5.9
5	文英堂	NEW EDITION UNICORN ENGLISH WRITING	5.7
6	三省堂	Orbit English writing New Edition	5.5
7	桐原	SPECTRUM ENGLISH WRITING Second Edition	5
8	三省堂	The CROWN English Writing New Edition	4.7
9	第一	Evergreen WRITING	4.2
		計9冊	計54

表 2. 採用した文法項目リスト

番号	文法項目
1	I+am+名詞
2	[We|You|They]+are+名詞
3	[He|she]+is+名詞
4	I+am+形容詞
5	[We|You|They]+are+形容詞
6	[He|she]+is+形容詞
⋮	
148	仮定法未来
149	助動詞完了形
150	It+[is|was]+前置詞句+that
151	It+[is|was]+副詞+that
152	It+[is|was]+名詞句（主語）+that+[V|助動詞]
153	It+[is|was]+名詞句（目的語）+that+S+[V|助動詞]

表 3. 下位項目一覧

番号	下位項目(展開項目)
1	肯定文
2	否定文
3	疑問文
4	否定疑問文
5	疑問詞が単独で主語の疑問文
6	疑問詞＋名詞が主語の疑問文
7	疑問詞が単独で目的語の疑問文
8	疑問詞＋名詞が目的語の疑問文
9	when疑問文
10	where疑問文
11	why疑問文
12	how疑問文
13	how+形容詞･副詞疑問文
14	疑問詞の前に前置詞を伴う疑問詞疑問文