社会イノベーション研究
第 11 巻第 2 号（107-122）
2016 年 3 月

2016 年 2 月 29 日掲載承認

英文中の文法項目頻度調査のための項目選定と英文からの抽出法

―CEFR-J の枠組みでの研究―

石　井　康　毅

1.　本研究の背景と目的

テキストの難易度測定や学習者の習熟度測定において重要な指標となり得るものとして，語彙のレベル，タイプ・トークン比，平均文長，文法，（話し言葉の場合）発話速度，流暢さなどが考えられるが，本研究ではこのうちの文法に焦点を当てる。文法項目の使用状況を調査する方法について検討し，実際に文法項目リストを作成し，各項目の使用状況を調べる方法を論じる。

本研究は外国語教育の枠組みとして主流になりつつある CEFR の枠内で文法項目リストを作成することを目標とする。CEFR とは，外国語の学習・教授・評価のための共通枠組みとして世界的に利用が進んでいる Common European Framework of Reference for Languages: Learning, Teaching, Assessment（「外国語の学習・教授・評価のためのヨーロッパ言語共通参照枠」）の略で，ヨーロッパでの 1970 年代からの研究・実践に基づいて，言語中立の外国語能力を示す汎用の枠組みとして 2001 年に公開されたものである（Council of Europe 2001）。CEFR は言語を使って何ができるのかという観点で，言語能力を A1（初級）～C2（上級）の 6 レベルで規定する。例えば B1 レベルは「仕事・学校・娯楽などの場面で普段出会うような身近な話題について，はっきり標準的な話し方・書き方をされれば要点を理解できる。」（レベル概要の一部）

と規定される。CEFR 自体は言語中立でやや抽象的であるため，対象の各言語で具体化する際に，「参照レベル記述」（reference level description; RLD）と呼ばれる CEFR レベルに対応する言語材料（語彙・文法・機能等）の割り当てが行われる。

CEFR は言語教育と評価の枠組みとして世界的に広く利用されるようになってきている。日本の英語教育政策においては，2011 年に文部科学省が『国際共通語としての英語力向上のための 5 つの提言と具体的施策』（文部科学省 外国語能力の向上に関する検討会 2011）によって CAN-DO（「～ができる」）形式での学習到達目標の作成を各中学校・高等学校に求めたが，この CAN-DO は CEFR の枠組みの柱である。CAN-DO の考え方は中等教育の段階ではかなり浸透してきており，平成 28 年度から使用開始となる中学校用の新しい英語の検定教科書では，全 6 社の教科書で，単元ごとに CAN-DO 形式の到達目標が明示され，日本国内で広く利用されている各種英語能力検定試験の中でも CEFR・CAN-DO の考え方を導入するものが増えてきている。

現在，CEFR を日本の英語教育環境に適用した CEFR-J（投野 2013）の開発が行われていて，CEFR の考え方や手法が今後日本の英語教育においてさらに広範囲で利用されると期待されている。このような状況において，本研究は筆者が研究分担者の一人である CEFR-J の RLD（謝辞参照）の一環として行った。したがって，本研究の目的は，CEFR の枠組みと日本の英語教育における学校文法の両方を考慮しながら，英語の学習・教授・評価のために必要な文法項目の選定と頻度調査の方法を確立することにある。

以下，先行研究と文法項目の選定（2 節），文法項目の使用例を抽出・頻度集計するために必要なコーパスの整備（3 節），作成した文法項目定義の精度（4 節），今後の課題と展望（5 節）について述べる。

2. 文法項目の選定

2.1. CEFR の枠組みに基づく文法項目提示の先行研究

CEFR の枠組みで，各レベルの違いを判別する基準特性（criterial feature）としての文法項目を分類し，提示しているものとしては，主に次の 3 点を挙げることができる。

1. 通称 “T-series”（van Ek and Trim 1991a/1998a, 1991b/1998b, 2001; Trim,

2009）

CEFR 公開以前に作られ，CEFR の基盤となった重要な研究である。学習者のニーズ分析を基に，各レベルの学習者が習得することを目標とするべき機能・概念を挙げていて，その中に文法項目も含まれている。

2. *A Core Inventory for General English*（North, Ortega and Sheehan 2010）

CEFR に基づく教材・指導を基に，CEFR レベルごとの特性を提示している。その中に文法項目も含まれ，各文法項目が典型的に指導されるレベルを提示している。

3. *English Grammar Profile*（2015）

約 5,500 万語の Cambridge Learner Corpus を基に，各文法項目がどの CEFR レベルで習得されるかを特定したデータである。

1 の T-series は客観的なデータに基づくものではなく，また 1970 年代～2000 年前後に作られたものであるため，CEFR レベル指標付きコーパスから基準特性を抽出する研究が進んでいる（Hawkins and Filipović 2012）現在では補助的に参照するのにとどめるのが適切であろう。

2 の *Core Inventory* は，例えば A1 レベルの基準特性として adjectives (common and demonstrative), adverbs of frequency, comparatives and superlatives, *going to, how much / how many* and very common uncountable nouns, *I'd like,* imperatives (affirmative/negative), past simple, verb + ing: *like/hate/love* といった項目を挙げている[1)] が，厳密に定義して頻度を明らかにするのが難しい項目が見られる（*how much / how many* and very common uncountable nouns など）。また，語彙・定型表現と切り分けにくいもの（*I'd like* など）は，日本の学校文法の観点からすると文法項目としては扱いにくい場合もある。

3 の *English Grammar Profile* は世界中の英語学習者がどのレベルで各文法項目を習得しているかを明らかにするデータであるが，各文法項目を日本人学習者にどの段階で指導すべきかという判断にそのまま利用できるとは限らないものも多い。

2.2. 日本の学校英文法の枠組みに基づく文法項目リスト作成の先行研究

日本の学校英文法の枠組みで文法項目を網羅的にリストアップした研究の一つに，『文法項目別 BNC 用例集及び文法項目集（1.0 版）』（東京外国語大学佐

1) *Core Inventory* より，筆者が表記を一部変更した上で引用している。

野研究室 2005）がある。（以下，『文法項目集』と略記する。）これはシェアの高い検定教科書 31 点（中学 6 点・高校「英語Ⅰ」8 点・高校「英語Ⅱ」8 点・高校「ライティング」9 点[2)]）と文法書 4 点の調査に基づき，中学・高校で学習する学校英文法で取り上げられる主要な文法項目 144 項目をカバーしたリストである（Minn et al. 2005: 105-108）。各文法項目は肯定平叙文・否定平叙文・肯定疑問文・否定疑問文・疑問詞疑問文 10 種の計 14 種の文タイプで個別のパターンとして定義されている。ただし 696 のパターンは構造上存在しないため，全部で 1,320 のパターンが定義されている。

『文法項目集』の項目の一部を例として以下に示す。（文法項目の呼称等は当該資料による。）

1. 【人称代名詞（I）am 名詞】I＋am＋名詞
2. 【人称代名詞（We, You, They）are 名詞】(We | You | They)＋are＋名詞
18. 【人称代名詞所有格 名詞 be 動詞 前置詞】(my | your | his | her | its | our | their)＋名詞＋(is | are)＋前置詞
36. 【How－感嘆文】How (形容詞 | 副詞)＋.*＋！
41. 【副詞節（when）】when
43. 【時制表現（will）】名詞句＋will＋一般動詞（原形不定詞）
44. 【意志未来（be going to）】名詞句＋(is | am | are | was | were)＋going＋to＋一般動詞（原形不定詞）
69. 【第五文型 主語－述語－目的語－補語（S＋V＋O＋C）】名詞句＋(make | bake | boil | burn | drive | render | dye | paint | set | turn | wipe)＋名詞句＋名詞句
70. 【不定詞（名詞用法）】to 動詞（原形不定詞）
77. 【比較級（形容詞＋er）】be 動詞＋形容詞（比較級）［＋than］
78. 【比較級（more＋形容詞）】be 動詞＋more＋形容詞［＋than］
86. 【現在完了形（be 動詞）】(has | have)＋been
87. 【現在完了形（一般動詞）】(has | have)＋一般動詞（過去分詞）
88. 【現在形・過去形の受動態】(am | is | are | was | were)＋一般動詞（過去分詞）
91. 【助動詞を使った受動態】助動詞＋be＋一般動詞（過去分詞）
97. 【It be 動詞（相当語）形容詞（for 目的語）that 節】It＋(be 動詞 | 一般

2） 科目名は調査当時のものである。

動詞(be 動詞相当語))＋形容詞[＋for 名詞句]＋that

104. 【関係代名詞（whose＋主格)】名詞句＋whose＋名詞句＋(助動詞 | 動詞)

139. 【仮定法過去完了】if＋...＋had＋動詞(過去分詞)＋...＋(would | should | could | might)＋have＋動詞(過去分詞)

144. 【it is ... that（目的語強調)】it＋is＋名詞句＋that＋名詞句

『文法項目集』は小学館の BNC 検索サービス[3] の一部として，各文法項目に該当する BNC の用例を抽出するために作られたものであるため，各文法項目は語形・レマ・BNC の品詞タグに基づいて，当該サービスで利用される CQL（Corpus Query Language）の形式で定義されている。CQL で定義されたパターンは公開されていないが，筆者は『文法項目集』を作成した東京外国語大学佐野洋教授に入手・利用の許諾を得た。文法項目と CQL の例を**表 1** に示す。

表 1：文法項目と CQL の例

ID[4]	CQL	意味
1–1	<or> <cql>^{W="i" P="PNP"} {P="VBB"} {P="N.*"}</cql> <cql>^{W="i" P="PNP"} {P="VBB"} {W="a|an|the" P="AT0"} {P="N.*"}</cql> </or>	文頭位置＋人称代名詞の I＋am＋名詞 または 文頭位置＋人称代名詞の I＋am＋冠詞の a または an または the＋名詞
3–86	<cql>^{P="VHB|VHZ"} {P="PNP|AT0|DT0"} [0,3] {P="VBN"} [0,10] {W="?"}</cql>	文頭位置＋have または has＋人称代名詞または冠詞または決定詞＋0～3 語の任意の語＋been＋0～10 語の任意の語＋疑問符

ID 1–1：I＋am＋名詞の肯定平叙文
ID 3–86：現在完了形（be 動詞）の肯定疑問文

2.3. 文法項目リストの作成

本研究ではあくまで日本の英語教育に適用した CEFR，すなわち CEFR-J の枠内で文法項目リストを作成することを目的としているため，『文法項目集』を基本データとして利用することとした。

ただし，2.2 で述べたように，『文法項目集』は特定環境での利用に特化したものであるため，任意のコーパスを対象として利用することはできない。そ

3) 小学館コーパスネットワーク BNC Online（http://bnc.jkn21.com）

4) 最初の数字が文タイプを，後の数字が項目の ID を表す。

こで筆者は次の手順により任意のコーパスにおける文法項目頻度調査を可能にすることとした。

1. コーパスデータに BNC と同じ品詞タグを付与する。
2. 『文法項目集』の CQL で書かれたパターンを汎用的な正規表現に変換し，手順 1 の品詞タグを付与したコーパスデータを検索する。

手順 1 は，BNC の品詞タグ付与で利用された CLAWS（Garside and Smith 1997）を利用することで可能になる（詳細は 3 節を参照）。

手順 2 の「正規表現」とは，通常の文字と特殊な意味を持つ「メタ文字」を使って，文字列を文字のパターンとして柔軟に表現することを可能にするものである。通常の文字列検索では探し出せない文字列のパターンを検索・置換するために使われる。『文法項目集』のパターン総数は 1,320 と極めて多いことから，CQL パターンを手順 1 で作成される形式のコーパスデータを検索可能な正規表現に自動で変換することとし，プログラムを作成した。例えば上記の ID 1–1 の CQL パターンは次の正規表現に変換された。

<[sv] n="¥d+"[^>]*?> ?<w[^>]+?c5="(PNP)"[^>]+?>(i)</w> ?<w[^>]+?c5="(VBB)"[^>]+?>[^<]+?</w> ?<w[^>]+?c5="(N[A–Z¥d¥–]+)"[^>]+?>[^<]+?</w>|<[sv] n="¥d+"[^>]*?> ?<w[^>]+?c5="(PNP)"[^>]+?>(i)</w> ?<w[^>]+?c5="(VBB)"[^>]+?>[^<]+?</w> ?<w[^>]+?c5="(AT0)"[^>]+?>(alan|the)</w> ?<w[^>]+?c5="(N[A–Z¥d¥–]+)"[^>]+?>[^<]+?</w>

ところが，変換した正規表現によりパイロット調査を行ったところ，いくつかの問題に直面した。

- 変換により得られた正規表現は元の CQL パターンと同様の対象を検索することができるものの，機械的な変換により不必要に複雑な正規表現になり処理コストが非常に高い，すなわち処理に非常に時間がかかる正規表現となってしまった。
- 元の CQL パターンの中には，ある集合Ａを検索し，その中から部分集合Ｂの検索結果を除くといった処理（diff）が行われているものもあり，そのようなものは一つの正規表現とするのが難しいという問題があった。本研究においても差分を差し引くような検索・集計システムを採用することも検討したが，後述するように文法項目に該当する部分をマークアップできるようにすることを目指したため，各文法項目を差分を使わずに一つの正規表現として定義することが必要であった。

- 元の CQL パターンに誤りを含むものや精度の低いものが見られた。
- 『文法項目集』の項目選定に，以下に挙げる例のようなバランスの悪さや不足が見られた。
  - ID 92 の「【進行形受動態】(am | is | are | was | were)＋being＋一般動詞（過去分詞）」は，CEFR レベルの基準特性を考えるという目的では，現在進行の受動態と過去進行の受動態に分けた方がよいと考えられる。
  - 再帰代名詞や used to といった重要度の高いと思われる項目が欠落している。

これらの問題を解消するために，まず 2.1 で挙げた T-series, *Core Inventory, English Grammar Profile* を参照して項目の追加・削除・整理を行うことで，日本の学校英文法だけでなく世界基準の CEFR の視点もできる限り取り入れることとした[5)]。その上で，効率よく，そして同時に高精度で検索・集計することを目指して，全ての項目の正規表現を新たに作成した。こうしてできた新しい文法項目リストの最終的な項目数は 255 項目となった。先行研究では同じ文法項目でも肯定文・否定文等の違いによって CEFR レベルが異なるとされる場合も多いことから，現在完了進行や助動詞類など，動詞や構文に関する項目については『文法項目集』にならい，肯定平叙文・否定平叙文・肯定疑問文・否定疑問文などの文種別の異なる変種を作成した。それらの変種まで数え上げると，新しい文法項目集の総数は 493 種である。ただし，全ての項目は理論上存在し得るものであるものの，このうちの約 5 分の 1 は筆者らが作成したコーパス（3 節参照）では 1 例も使用が確認されていない。

各文法項目は，筆者らが作成した XML 形式のコーパス（3 節参照）に対応するよう，語形・レマ・品詞のパターンとして正規表現で定義した。XML タグの一部を利用すればパターン定義の正規表現を簡略化することができるが，4 節で述べる精度評価や用例の確認の際に，各文法項目に該当する部分をマークアップしながら整形式の（well-formed）XML データとして出力できるように，全ての文法項目は単語（列）の単位で取得できるように定義した。文法項目と正規表現の例を**表 2** に示す。（なお，筆者の作成した文法項目リストの ID は『文法項目集』の ID とは無関係である。）

5） この作業は，筆者が研究分担者の一人である CEFR-J RLD 研究チームの研究代表者・他の研究分担者・研究協力者の協力を得て行った。

表 2：文法項目と正規表現の例

ID	文法項目	正規表現
26	none（不定代名詞）	<w c7="PN" c5="PNI" hw="none" pos="PRON">none </w>
47	比較級 and 比較級（同じ比較級）	(<w c7="(JJR|RRR|RGR)" [^>]+>[^<]+</w>) <w c7="CC" c5="CJC" hw="and" pos="CONJ">and</w> ¥1
64	過去進行形（肯定平叙文）	<w c7="VBD." [^>]+>[^<]+</w> <w c7="V.G" [^>]+>[^<]+</w>
124	have to（肯定平叙文）	(?<!hw="not" pos="ADV")>[^<]+</w> ¥K<w c7="VH." [^>]+>[^<]+</w> <w [^>]+>to</w> <w c7="V.I" [^>]+>[^<]+</w>
142	助動詞＋完了（肯定平叙文）	<w c7＝"VM" [^>]+>[^<]+</w> <w [^>]+>have</w> <w c7＝"V.N" [^>]+>[^<]+</w>

3. 文法項目の使用例を抽出・頻度集計するために必要なコーパスの整備

本研究では，2 節で定義を作成した文法項目リストを利用して各項目の頻度を集計するために必要なコーパスのデータ形式も策定し，実際にいくつかのコーパスを作成した。（後述する教材コーパスに含める教材の選定，電子化，メタ情報付与，品詞タグ付与の具体的な手順や問題については内田（2015: 89-92）を参照されたい。）2.3 で述べたように，パイロット調査の段階では BNC のデータを対象とした検索用 CQL パターンを正規表現に変換して利用したため，コーパスデータとしては必然的に，CLAWS で品詞タグを付与し，BNC XML Edition（*The British National Corpus,* version 3 (BNC XML Edition) 2007）に準拠した形式を想定した（図 1 参照）。

ただし，BNC は CLAWS 以外にも処理を行ってタグの修正やレマ情報の付与を行っている。『文法項目集』でもレマ情報を利用しているため，本研究では CLAWS による品詞タグ付与に加えてレマ情報の付与も可能な Wmatrix（Rayson 2009）を利用した（図 2）。また，CLAWS では，BNC で付与されている C5[6] というタグセットよりも詳細な分類がなされた C7[7] というタグセットでのタグ付与が可能なため，本研究におけるコーパスの形式は BNC XML

6） http://ucrel.lancs.ac.uk/claws5tags.html

7） http://ucrel.lancs.ac.uk/claws7tags.html

図1：BNC XML Edition のデータの一部

```
<s n="2">
  ...
  <w c5="PNP" hw="he" pos="PRON">he</w>
  <w c5="VBD" hw="be" pos="VERB">was</w>
  <w c5="VVG" hw="drive" pos="VERB">driving</w>
  <w c5="PRP" hw="through" pos="PREP">through</w>
  <w c5="NP0" hw="lydsett" pos="SUBST">Lydsett</w>
  <w c5="NN1" hw="village" pos="SUBST">village</w>
  <c c5="PUN">.</c>
</s>
```

図2：Wmatrix の出力データの一部[8)]

0000546010	NNU	#####R#####7	#####r#####7
0000547010	NN2	Families	family
0000547020	VBR	are	be
0000547030	JJ@	like	like
0000547040	NN2	bars	bar
0000547050	IO	of	of
0000547060	NN1	chocolate	chocolate
0000547061	–	–	PUNC
0000547070	RR	mostly	mostly
0000547080	JJ	sweet	sweet
0000547081	,	,	PUNC
0000547090	IW	with	with

Edition を拡張した情報（C7 タグ・C5 タグ・見出し語・大まかな品詞分類）を持つ構造とした。

Wmatrix の処理結果は，プログラムによって BNC を独自に拡張した XML 形式（上述）のデータに変換した（図3）。

文法項目の正規表現は BNC のデータを前提としていたパイロット調査の後に全て新たに書き直したので，その時点で BNC XML Edition に近いデータ構造を保持する必然性はなくなっていたが，将来的に TEI（Text Encoding Initiative；標準的なテキスト電子化方法の確立を目指すコンソーシアム）のガイドラインに準拠したデータとする可能性を考慮し，コーパスのデータ構造は変更しなかった。（BNC XML Edition は TEI のガイドラインに準拠している。）また，同時に品詞タガーとして CLAWS を使い続ける必然性もなくなっていたが，BNC との比較がしやすいため，タガーも変更しなかった。

8） これは後述する ELT Coursebook Corpus の一部であるが，1行目の「#####R#####7」はメタ情報である当該部分のスキル種別（Reading）とページ番号を表している。

図 3：最終的な XML データの一部

```
<part type="Reading" page="7">
    <s n="18">
        <w c7="NN2" c5="NN2" hw="family" pos="SUBST">Families</w>
        <w c7="VBR" c5="VBB" hw="be" pos="VERB">are</w>
        <w c7="JJ" c5="AJ0" hw="like" pos="ADJ">like</w>
        <w c7="NN2" c5="NN2" hw="bar" pos="SUBST">bars</w>
        <w c7="IO" c5="PRF" hw="of" pos="PREP">of</w>
        <w c7="NN1" c5="NN1" hw="chocolate" pos="SUBST">chocolate</w>
        <c c5="PUN">–</c>
        <w c7="RR" c5="AV0" hw="mostly" pos="ADV">mostly</w>
        <w c7="JJ" c5="AJ0" hw="sweet" pos="ADJ">sweet</w>
        <c c5="PUN">,</c>
        <w c7="IW" c5="PRP" hw="with" pos="PREP">with</w>
        ...
```

実際に様々なコーパスを作成するに当たっては，単語単位の情報に加えて，コーパスに関する情報・ページ番号・レベル・教材のスキル種別・テストの内容などのメタ情報も含める必要がある。筆者を含む研究チームで作成したコーパスではこれらの情報を全て含む整形式の XML としてデータを作成している。

筆者を含む研究チームが作成したデータの一例が，CEFR レベル別の海外の ELT 教材 96 点のコーパス（ELT Coursebook Corpus と呼称）である。このコーパスの規模は，語彙・表現パターンの列挙部分を除いて約 164 万語（A1：13.8 万，A2：25.1 万，B1：44.4 万，B2：52.2 万，C1：25.5 万，C2：2.9 万）で，教材別・スキル別（リーディング・ライティング・文法など）などのサブコーパスがあり，サブコーパスを指定しての文法項目検索や頻度調査も可能である。

4. 作成した文法項目定義の精度評価

正規表現として定義した文法項目（2.3）を使って，作成した XML コーパス（3 節）を対象に検索を行うことで，各文法項目の頻度を集計することが可能になる。筆者らが作成した ELT Coursebook Corpus（3 節）を対象とした集計結果の一部を**表 3** と**表 4** に示す。（文法項目の ID は**表 2** 内のものと対応する。）

CEFR レベル別サブコーパスの頻度集計により，各文法項目が異なるレベルでどのような使用頻度の違いを示すかを明らかにすることができる。**表 4** のう

表 3：各文法項目の頻度集計結果の一部（ELT 教材サブコーパスごとの実度数）

ID	001 (A1)	002 (A2)	003 (B1)	004 (B2)	005 (C1)
26	0	0	2	3	3
47	0	0	0	0	0
64	0	3	10	9	20
124	2	12	19	37	41
142	0	0	0	6	13

表 4：各文法項目の頻度集計結果の一部（CEFR レベルごとの 100 万語あたりの相対頻度）

ID	A1	A2	B1	B2	C1
26	0	32	45	134	81
47	0	4	27	34	15
64	6	743	1,239	911	1,152
124	358	696	1,144	1,066	935
142	0	11	210	546	480

図 4：ID 64 過去進行形（肯定平叙文）のレベルごとの使用頻度の違い

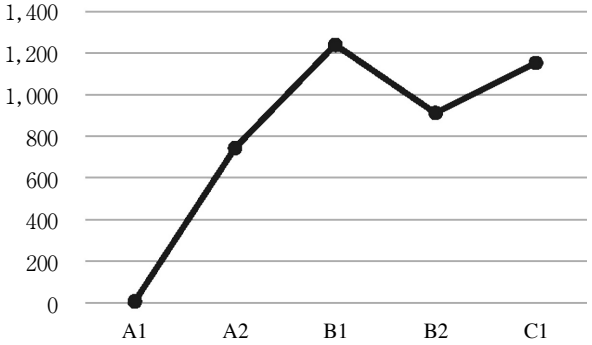
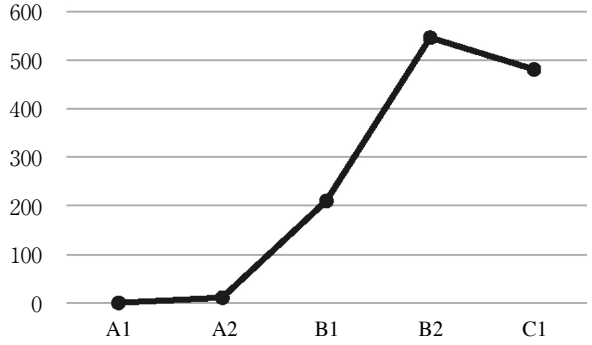


図 5：ID 142 助動詞＋完了（肯定平叙文）のレベルごとの使用頻度の違い



ち ID 64 の過去進行形（肯定平叙文）のレベルごとの使用頻度の違いをグラフで示したものが**図 4**，ID 124 の助動詞＋完了（肯定平叙文）のレベルごとの使用頻度の違いをグラフで示したものが**図 5** である。

これらの結果から，少なくとも筆者らが構築した ELT Coursebook Corpus においては，過去進行形は A2 レベルで使用が始まり，助動詞＋完了は B1～B2 レベルで使用が始まるということが分かる。

本研究で作成・定義した文法項目の精度については石井・投野（2016: 779-780）で調査・評価しているが，ここではその結果の一部を抜粋して示す。

精度の評価基準として，情報検索の分野でよく用いられる適合率（precision；検索したもののうちどれだけが正しいかを示す割合），再現率（recall；正しいもののうちどれだけを検索できているかを示す割合），F 値（F-measure；適合率と再現率の調和平均）を用いる。適合率と再現率は一般にトレードオフの関係にあるため，両者が高い場合に高くなる F 値で評価が行われることが多い。本研究で作成した文法項目 493 種のうち 207 の項目について，東京外国語大学投野由紀夫研究室において，以下の方法で F 値を算出した[9)]。

1. 多くの文法項目が出現すると予想される B1 レベルの代表的な ELT 教材 1 点（語数約 4.7 万語）を対象とし，正規表現による全抽出例を目視で確認し，次式により適合率を得る。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

TP (true positive)：検索されたもので正しいもの
FP (false positive)：検索されたが正しくないもの

2. 複数の語形による検索や目視等の検索方法を適宜組み合わせ，人手により抽出漏れがないかを確認し，次式により再現率を得る。

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

FN (false negative)：検索されなかった正しいもの

3. 次式により F 値を得る

$$\text{F} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

9) 調査した項目数が全体のうちのおよそ 4 割であるのは，この調査後に項目・定義の大改訂を行ったためである。残りの項目については今後調査予定である。

表 5：適合率・再現率・F 値の平均

適合率	再現率	F 値
0.947	0.891	0.892

検証の結果，207 項目のうち，当該の教材で実際に使用が確認できた 124 項目の平均の適合率，再現率，F 値は**表 5** のようになった。

124 項目中 83 項目は F 値 0.9 以上であった。また，0.9 未満の項目でも，文頭位置に限定して定義したがゆえに再現率が低かったものなどは，今後の文法項目リストの改訂時に容易に修正できる見込みである。語形・レマ・品詞のみで特定するのが難しい項目（構文に関するものなど）の F 値が低い，副詞要素が修飾句として挿入される場合の想定が難しい，名詞句を高精度で定義するのが難しいなど，パターン定義の困難点が明らかになったが，F 値が 0.7 未満であるものは 124 項目中 15 項目しかなく，作成した正規表現の精度は概して良好であることが判明した。

5. 今後の課題と展望

本研究で作成した文法項目リストは，筆者が研究分担者の一人である CEFR-J の RLD 研究チームが成果物として公開予定の CEFR-J Grammar Profile というインベントリーの根幹を成すデータとなる。CEFR-J Grammar Profile は，日本人英語学習者向けの網羅的な文法項目リスト，各項目に先行研究が割り当てているレベル情報，各項目の ELT Coursebook Corpus（インプット）と学習者データ（アウトプット）における頻度，何点中何点の教材で使用されているかという分布情報（range），各項目の抽出精度，機械学習に基づく各レベルの基準特性などから成る総合データベースである。CEFR-J Grammar Profile は，CAN-DO ベースの英語学習目標設定が今後本格化する中で，CAN-DO と言語材料を結びつけ，シラバス・教材開発の重要な基礎資料になることが期待される。

本研究で作成した文法項目リストは，より具体的・直接的な形で利用することもできる。例えば，主に語彙や文長などで測られる文章の読みやすさ（readability）を，使用されている文法項目を考慮に入れることにより，より精密に測定できるようになる。また，既存の教材や辞書の評価や，教材・教授法の改

善にも利用できる。例えば Ishii and Minn（2015）は，辞書の用例を，使用されている文法項目の観点で評価した。石井（2016）は平成 28 年度から使用開始となる中学校用の新しい英語の検定教科書と平成 27 年度までの教科書を，使用されている文法項目の違いに着目して比較分析した。

本研究で作成した文法項目リストは，項目選定と定義の両方の点で今後も改善を続ける予定である。公開される文法項目リストが様々な調査・分析に利用され，英語教育の改善に資することを期待している。

引用文献

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

English Grammar Profile. 2015. (http://www.englishprofile.org/english-grammar-profile)

Garside, R. and N. Smith. 1997. “A hybrid grammatical tagger: CLAWS4.” In Garside, R., G. Leech and A. McEnery (eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman, pp. 102-121.

Hawkins, J. and L. Filipović. 2012. *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework* (English Profile Studies). Cambridge: Cambridge University Press.

Ishii, Y. and D. Minn. 2015. “Grammatical items used in EFL dictionary examples: From the Japanese EFL learner perspective.” In Li, L., J. Mckeown and L. Liu (eds.), *Words, Dictionaries and Corpora: Innovations in Reference Science (Proceedings of ASIALEX 2015 Hong Kong)*. The Asian Association for Lexicography, pp. 154-161.

Minn, D., H. Sano, M. Ino and T. Nakamura. 2005. “Using the BNC to create and develop educational materials and a website for learners of English.” *ICAME Journal,* 29, pp. 99-114.

North, B., A. Ortega and S. Sheehan. 2010. *A Core Inventory for General English*. British Council / EAQUALS. (http://clients.squareeye.net/uploads/eaquals2011/documents/EAQUALS_British_Council_Core_Curriculum_April2011.pdf)

Rayson, P. 2009. *Wmatrix: A Web-based Corpus Processing Environment*. Computing Department, Lancaster University. (http://ucrel.lancs.ac.uk/wmatrix/)

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. (http://www.natcorp.ox.ac.uk/)

Trim, J. L. M. 2009. *Breakthrough*. Cambridge: Cambridge University Press. (http://www.Englishprofile.org/images/stories/ep/breakthrough.doc)

van Ek, J. A. and J. L. M. Trim. 1991a/1998a. *Threshold 1990*. Cambridge: Cambridge University Press.

van Ek, J. A. and J. L. M. Trim. 1991b/1998b. *Waystage 1990*. Cambridge: Cambridge University Press.

van Ek, J. A. and J. L. M. Trim. 2001. *Vantage*. Cambridge: Cambridge University Press.

石井康毅．2016．「教科書改訂に伴う文法項目使用状況の変化」『英語教育』2016 年 2 月号．東京：大修館書店，pp. 20-22．

石井康毅・投野由紀夫．2016．「CEFR-J Grammar Profile のための文法項目頻度調査」『言語処理学会第 22 回年次大会発表論文集』，pp. 777-780．

内田諭．2015．「CEFR レベルに基づいた教材コーパス－レベル別基準特性の抽出に向けて」『英語コーパス研究』第 22 号，英語コーパス学会，pp. 87-99．

東京外国語大学佐野研究室．2005．『文法項目別 BNC 用例集及び文法項目集（1.0 版）』．

投野由紀夫（編）．2013．『CAN-DO リスト作成・活用　新しい英語到達度指標 CEFR-J ガイドブック』東京：大修館書店．

文部科学省　外国語能力の向上に関する検討会．2011．『国際共通語としての英語力向上のための 5 つの提言と具体的施策～英語を学ぶ意欲と使う機会の充実を通じた確かなコミュニケーション能力の育成に向けて～』．(http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2011/07/13/1308401_1.pdf)

謝辞

本研究は，筆者が研究分担者の一人である JSPS 科研費基盤研究（A）「学習者コーパスによる英語 CEFR レベル基準特性の特定と活用に関する総合的研究」（課題番号：24242017・代表：投野由紀夫）の助成を受けている。本研究に関連する研究・調査・作業に携わった研究代表者・分担者・協力者に感謝の意を表する。

本研究はまた，JSPS 科研費若手研究（B）「日本人英語学習者の学習・使用実態を反映した重要句動詞リストの作成」（課題番号：26770201・代表：石井康毅）と平成 26～27 年度成城大学特別研究助成「新課程用英語検定教科書における語彙使用状況の調査と分析」の成果も利用している。ここに記して謝意を表する。