CEFR-J Grammar Profile のための文法項目頻度調査

石井 康毅
成城大学社会イノベーション学部
ishii@seijo.ac.jp

投野 由紀夫
東京外国語大学大学院総合国際学研究院
y.tono@tufs.ac.jp

1. 研究目的

ヨーロッパ言語共通参照枠（Common European Framework of Reference for Languages : CEFR）は，言語中立の外国語能力を示す汎用枠であるが，世界的にその利用が進んでおり，日本の英語教育政策でも文部科学省が CAN-DO 形式での学習到達目標の作成などに CEFR を意識した枠組みを推奨している。この CEFR を日本の英語教育環境に適用したものが CEFR-J [7]であり，今後広範囲の利用が期待されている。CEFR 自体は言語中立なので，各国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の配当が行われる。本研究の目的は，学校文法を視野に入れながら，CEFR-J レベルに即した文法項目の導入順序と，各レベルの違いを判別する基準特性（criterial feature）となり得る文法項目をコーパス準拠で特定することである。昨年度，NLP2015 のテーマセッションでこのプロジェクトが紹介され[8]，その後この文法項目頻度調査は総合的な CEFR-J の RLD の一環として継続されてきた。本調査の結果は，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，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。

以下，文法項目リスト作成の過程と，各項目に該当する用例をコーパスから抽出した結果の精度評価，今後の課題と展望について述べる。

2. 文法項目

CEFR の枠組みで基準特性としての文法項目を分類し，提示しているものとしては，次のようなものがある。

- CEFR 公開以前に作られ，CEFR の基盤となった T-series [9-12]
- CEFR に基づく教材・指導から各文法項目が典型的に指導されるレベルを提示した *Core Inventory* [3]
- 約 5,500 万語の Cambridge Learner Corpus を基に各文法項目の CEFR レベルを特定した *English Grammar Profile* [1]

しかし，例えば[3]では A1 レベルの基準特性として *how much / how many* and very common uncountable nouns, *I'd like*, verb + ing: *like/hate/love* といった項目が挙げられていて，語彙と文法が入り交じっていたり，項目の括りが荒かったりして，日本の学校文法の観点からすると扱いにくい点が目立つ。また，[1]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。

そこで，本研究では中学・高校で学習する学校英文法で取り上げられる主要な項目をカバーした東京外国語大学佐野研究室作成の文法項目リスト[5]をベースに，前述の[9-12]，[3]，[1]で取り上げられている項目を追加した。

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

各文法項目は，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。各文法項目に該当する部分を含む文を，当該部分をマークアップしながら 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）単位で定義した。例を表 1 に示す。

表 1. 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	<w c7="VM" [^>]+>[^<]+</w> <w [^>]+>have</w> <w c7="V.N" [^>]+>[^<]+</w>

最終的な文法項目の数は 255 項目，文種別（肯定平叙文・否定平叙文・肯定疑問文・否定疑問文など）の異なる下位項目まで数え上げると 493 種である。ただし，そのうちの 93 種は理論上は可能な文であるものの，3 節で述べるコーパスでは 1 例も確認されなかった。

3. 使用コーパス

以下の 3 種のコーパスデータを構築し，利用した。（括弧内のバージョン番号は本プロジェクトにおける内部バージョン番号である。）

- CEFR レベル別の海外の ELT 教材 96 点（Ver. 3.0）：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）。教材別・スキル別などのサブコーパスがある。
- JEFLL Corpus [6]の CEFR レベル付きオリジナル作文データと誤り訂正済みデータ（Ver. 1.3）：オリジナルデータは約 66 万語（A1: 13.2 万，A2: 31.0 万，B1: 21.2 万，B2: 0.9 万）。
- NICT JLE Corpus [2]（Ver. 1.0）：約 97 万語。テストの内容別のサブコーパスがある。

これらのコーパスは，Wmatrix [4]を利用して全ての語に品詞タグとレマの情報を付与した上で，XML 形式のコーパスとして整備した。

4. 各文法項目に該当する用例の集計

作成した文法項目の正規表現により，全ての（サブ）コーパスにおける各文法項目の度数を集計し，相対頻度を計算した。また，ELT 教材コーパスでは何点の教材に出現しているかという分布情報（range）も集計した。

集計結果の一部を表 2～表 4 に示す。（ID は表 1 内のものと対応する。）

表 2. 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

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

表 3. 各文法項目の頻度集計結果の一部（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

表 4. 各文法項目の頻度集計結果の一部（レベルごとの分布）（括弧内は当該レベルの教材点数）

ID	A1 (17)	A2 (21)	B1 (26)	B2 (23)	C1 (8)
26	0	4	13	19	7
47	0	1	12	12	3
64	1	16	26	23	8
124	6	18	26	23	8
142	0	1	17	23	8

5. 文法項目抽出の精度評価

作成した文法項目の精度を評価するために，493 種のうち 207 の項目について，以下の方法で適合率と再現率を検証した。

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，適合率を得る。
2) その他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，再現率を得る。

検証の結果，207 項目中，当該の教材で出現が確認できた 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。

表 5. 適合率，再現率，F 値の平均

適合率（precision）	再現率（recall）	F 値
0.947	0.891	0.892

作成した正規表現の精度は概して比較的高く，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たがゆえに再現率が低かったものなどはすぐに修正が可能である。

表 6 に F 値が低かった項目の例を示す。

表 6. F 値が低かった文法項目とその理由

文法項目	F 値	理由
関係代名詞（目的格）の省略	0.191	複合名詞・前置詞句内の名詞＋節などにマッチ
関係副詞（先行詞あり）	0.400	接続詞の when にマッチ
as が導く副詞節	0.509	as ... as 名詞句の 2 つ目の as にマッチ
名詞を後置修飾する過去分詞	0.567	現在完了の疑問文の主語＋過去分詞部分がマッチ
S+V(give/pass/send/show/teach/tell)+IO+DO（肯定平叙文）	0.600	想定外の数詞を含む目的語にマッチせず
名詞を後置修飾する現在分詞	0.682	分詞構文などにマッチ
助動詞 could（肯定疑問文）	0.688	想定外の..., please?・副詞を含むものなどにマッチせず

who・which・when などの関係詞は品詞タグ上で疑問詞や接続詞などの他の用法との区別ができないため，正確な抽出がかなり困難である。また，目的格の関係代名詞の省略として「先行詞となる名詞の後に名詞句＋（助）動詞が続くもの」というパターンを想定したが，名詞＋名詞の複合名詞にマッチしてしまうなど，構文パターンが特定しにくい項目の F 値は低い。その他，副詞要素が修飾句として挿入される場合や，高精度での名詞句の定義の難しさなど，パターン定義の困難点が明らかになった。

しかし F 値が 0.7 未満であるものは 124 項目中

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

15 項しかなく, 作成した正規表現の精度はおおむね良好であることが判明した。

我々は, CEFR レベル別の各種コーパス（ELT 教材・書き言葉学習者コーパス・話し言葉学習者コーパス）におけるこれらの文法項目の頻度データを CEFR-J Grammar Profile の基礎データとして, 上記の F 値を添えた形で公開する方向で検討している。

6. 課題と展望

CEFR-J Grammar Profile の構築のために，網羅的な文法項目リストを作成し，それらの頻度をインプットのソースとなる ELT 教材コーパス, アウトプットの学習者データから自動抽出して集計する試みを紹介したが，この頻度データを基礎に同じ研究グループの東京工業大学奥村学研究室で機械学習を行い，レベル別提示順序の判定とレベル判別に寄与する文法項目の特定という作業を行う。その際に，擬似的 CEFR-J レベルを出すために，ページ情報を基準にして A1 用テキストを A1.1・A1.2・A1.3 に三分割（A2～B2 はそれぞれ二分割）したデータを利用することで，より細分化されたレベル判定が可能かどうかを検証する。さらに ELT 教材コーパスには言語材料を listening・speaking・reading・writing の 4 技能に分けるセクション・タグも付与しているので，今後スキルごとの語彙・文法の分析も可能になる。これらの一連の試みは, CEFR に準拠した RLD を行う際の方法論的な手続きを客観的に示すという意味で，その意義は大きい。

謝辞

本研究は科学研究費基盤研究（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）の助成を受けたものである。

参考文献

[1] *English Grammar Profile*. 2015. http://www.englishprofile.org/english-grammar-profile

[2] 国立研究開発法人 情報通信研究機構. 2012. *NICT JLE Corpus Version 4.1*.

[3] 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

[4] Rayson, P. 2009. *Wmatrix: a web-based corpus processing environment*. Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/

[5] 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)』.

[6] 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』東京: 小学館.

[7] 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.

[8] 投野由紀夫・石井康毅. 2015. 「英語 CEFR レベルを規定する基準特性としての文法項目の抽出とその評価」『言語処理学会第 21 回年次大会発表論文集』, pp. 884-887.

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

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

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

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

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.