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

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

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

1. 研究目的

外国語能力の共通参照枠として Common European Framework of Reference for Languages（ヨーロッパ言語共通参照枠 : CEFR）の利用が世界的に進んでいる。これを日本の英語教育環境に適用したのが CEFR-J [8]であり, 現在の中高での CAN-DO 形式での学習到達目標の作成，今後の学習指導要領改訂，教科書作りなどで利用が広まってきている。CEFR レベル自体は言語中立なので，各国語で参照レベル記述（reference level description）という CEFR レベルに対応する言語材料の配置が行われる。英語では CEFR レベル指標付きコーパスからレベルを判別する基準特性（criterial feature）を抽出する研究が進んでいる[3]が，コーパス準拠で文法項目に該当する特性を抽出する方法を検討し，実際にリストを作ることが本研究の目的である。

レベル別基準特性の特定は，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[8]。

今回の研究では以下の研究設問を設けた：

(1) 基準特性の抽出のために, 中学・高校の文法項目の一括抽出を試みるが, その適合率（precision）, 再現率（recall）は文法項目ごとにどの程度か？

(2) (1)の抽出精度は通常の誤りを含まない英文と学習者データとではどの程度異なるか？

(3) CEFR レベル別教材から基準特性と考えられる文法項目を実際に抽出することが可能か？

2. 文法項目

CEFR の枠組みで，基準特性としての文法項目を挙げたものとしては[5]や[1]がある。しかし，例えば[5]では A1 の基準特性として adverbs of frequency, *going to*, *I'd like* が含まれ，[1]では A2 の基準特性として simple sentences using infinitives, some modals が含まれるなど，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と，日本の学校文法の観点からすると扱いにくい。

そこで，本研究では東京外国語大学佐野研究室が作成した学校文法項目のリスト[7]を利用した。これは中学・高校で学習する学校英文法で取り上げられる主要な部分をカバーするべく，中高の教科書・文法書から 144 の文法項目を抽出したものである[4, 6]。

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。

- CEFR レベル別の海外の ELT 教材：語彙・表現パターンの列挙部分を除いて約 155 万語（A1: 9.2 万，A2: 23.9 万，B1: 42.9 万，B2: 50.7 万，C1: 24.9 万，C2: 2.9 万）
- 現行の学習指導要領に基づく中高検定教科書：約 27 万語（中 1: 1.8 万，中 2: 2.8 万，中 3: 3.5 万，高 1CE1: 18.7 万）
- CEFR レベルに再分類した JEFLL コーパス：約 67 万語（A1: 13.6 万，A2: 31.3 万，B1: 21.5 万，B2: 0.9 万）

データは XML 形式で，全ての語に品詞タグが

付与されている。文法項目を記述するコーパス検索式（CQL）が BNC のデータを対象に作られたものであるため，品詞情報付与には，BNC の品詞タグ付与に用いられた CLAWS[2]を用いた。

4. 各文法項目に該当する用例の抽出

[7]の開発時に作成された，各文法項目に該当する用例を BNC から抽出するための CQL パターンを利用して，用例を抽出した。144 の文法項目は，14 の文種別（肯定文・否定文・肯定疑問文・否定疑問文・疑問詞疑問文 10 種）ごとに，語形・レマ・品詞のパターンとして定義され，組み合わせ上あり得ないものを除いて約 1,300 のパターンが記述されている。以下に[7]で挙げられている文法項目の例を示す。（文法項目の呼称等は当該資料による。）

- 【関係代名詞(主格)】名詞句＋(who|which|that)＋(助動詞|動詞)
- 【現在完了形(be 動詞)】(has|have)＋been
- 【It 動詞 形容詞 (for 目的語) to 動詞原形】It＋(一般動詞|be 動詞)＋形容詞 [＋for 名詞句] ＋to＋一般動詞(原形不定詞)
- 【仮定法過去】if＋...＋(were|was|動詞(過去形))＋...＋(would|should|could|might)＋動詞(原形不定詞)

CQL パターンは，今回利用したコーパスの形式に合うように，機械的に正規表現に変換した上で利用した。

今回の実験では文タイプ 1（肯定文）と 2（否定文）のみ対象とし，検索式が非常に複雑になるものを除いて，計 228 の文法項目に該当する用例を各コーパスから抽出した。教材サブコーパスごとの度数集計結果の一部を表 1 に示す。

表 1. 各文法項目の教材サブコーパスごとの度数集計結果の一部（文種別は肯定文）

文法項目	A1_001	A2_002	B1_003
be going to	0	26	9
can	24	27	55
would	1	3	10
形容詞＋er	14	14	42
more＋形容詞	0	0	1
現在完了 (be 動詞)	0	0	15
SVOC (C は現分)	0	0	0

5. 抽出の精度評価

ELT 教材の一部を対象として，適合率と再現率を目視で検証した結果の一部を表 2 に示す。適合率については概して高い。再現率については網羅的な評価は難しいが，特にパターンが単純な文法項目については十分に高く，文法項目の CQL パターンとしての定義は概ね適切であると考えられる（研究設問 1)。例えば，＜have（＋副詞）＋過去分詞＞として定義される「完了相」の項目を検定教科書で検索すると，中 1・中 2 では 1 例もなく，中 3 で初めてマッチし，さらに適合率も非常に高い。

表 2. 主要文法項目と抽出精度

文法項目	Precision	Recall	F 値
関係代名詞主格	0.98	0.74	0.85
have been	1.00	1.00	1.00
get＋過去分詞	0.99	0.55	0.70
It … 形 to do	1.00	0.98	0.99
仮定法過去	1.00	0.14	0.24

学習者コーパスにも本研究の手法が利用できることが確認できた。しかしながら，検索パターンが長くなれば対応する用例も当然長くなり，そこに英語の誤りが含まれる可能性も高くなる。ゆ

えに，検索パターンが長くなるとマッチしない用例が出てきて再現率は低下する（研究設問 2)。例えば，文法項目の一つとして扱われている will は＜助動詞 will ['ll]（＋副詞）＋動詞原形＞として定義されているが，“And we will practis harder than this year!”では practice の綴りが誤っているために名詞のタグが付与され，上記のパターンにマッチしない。

ELT 教材での CEFR レベル別の文法項目の使用状況を見ると，あるレベルで使用が急増したり，レベルの上昇に従って使用が漸増したりするものが多く見られ，これらはレベル基準特性と考えられる可能性がある（研究設問 3)。図 1 と図 2 に結果の一部を示す。

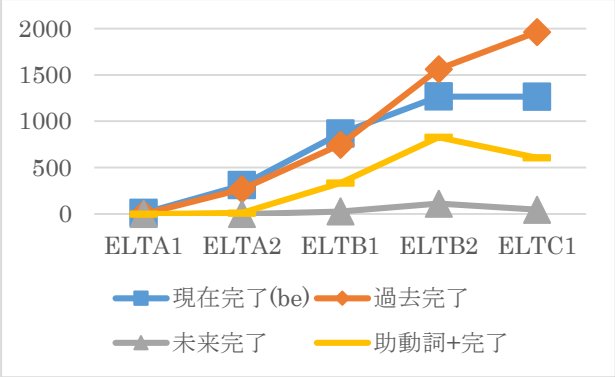

図 1. レベルごとの完了相に関する文法項目の度数（100 万語あたり）

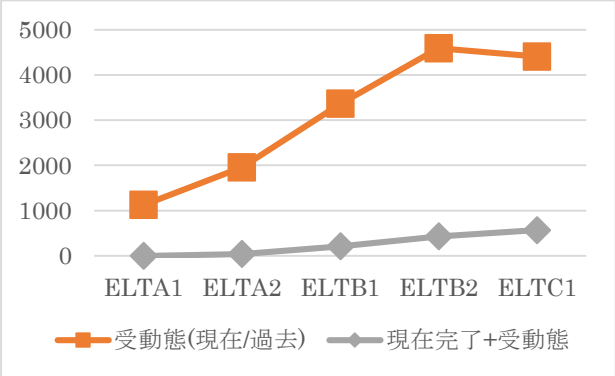

図 2. レベルごとの受動態に関する文法項目の度数（100 万語あたり）

6. 機械学習による言語特徴のレベル別分類

抽出によって得られた，228 の文法項目のレベル（A1～C1 の 5 つ）ごとの度数データを予測変数とし CEFR レベルを目的変数として Weka 3.6.11 [9]を利用し，機械学習による分類タスクを行った。手法としては J48（決定木 C4.5＋データ更新＋枝刈り），Support Vector Machine，Random Forest を用いた。表 3 に結果を示す。

表 3. 機械学習の精度比較（F 値）

	A1	A2	B1	B2	C1
J48	0.963	0.605	0.393	0.318	0.105
SVM	0.538	0.522	0.431	0.426	0.316
RF	0.857	0.634	0.586	0.588	0

決定木（J48）は A1 レベルの分類は極めて優秀であったが，B レベルは低かった。SVM は B1 から C1 にかけての判定が決定木よりも優秀であった。最も優秀だったのは Random Forest で，決定木とほぼ同等の精度を A レベルで出しただけでなく，SVM を超える精度を B レベルでも示した。これらすべてにおいて C レベルの精度が低かったのは，文法事項での判別自体が上級レベルではあまり有効でなくなってきていることが主な原因ではないかと思われる。

7. 機械学習情報による基準特性候補抽出

機械学習の結果，判別に寄与した言語特徴群を知ることにより，基準特性のリスト作成に有益な情報を得ることができる。例えば，図 3 は J48 の結果の判別木を可視化したものである。

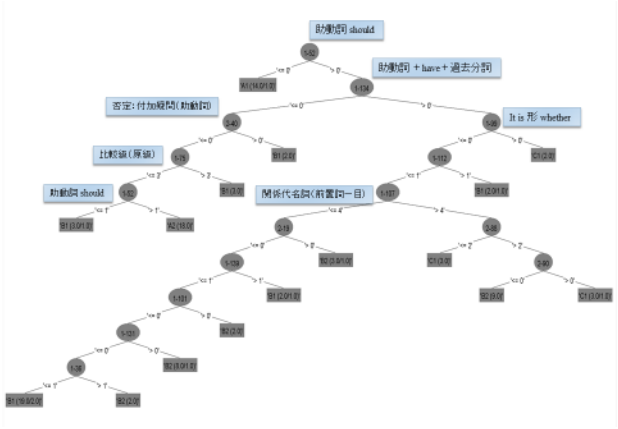

図 3. J48 の分類木構造

これを見ると, A1 と他の上位レベルを分ける最上位ノードの判別に助動詞 should が用いられている。これは少々意外な感じがするが，その後，大きく B1～C1 レベルを分ける基準特性として「助動詞＋have＋過去分詞」,「It is＋形容詞＋whether」,「関係代名詞（前置詞の目的語）」といった項目が判別に用いられており，直観とも合致する結果だといえる。

8. 課題と展望

文法項目は[7]のデータを機械的に正規表現に変換して利用したが，元の CQL パターンリストに誤りや項目選定のバランスの悪さが見られる。また，機械的な変換により非常に複雑な正規表現になり処理コストが高いという問題もある。そのため，項目を整理しながら新たに抽出式を書く作業を現在行っている。

言語特徴のレベル別分類については，異なる機械学習アルゴリズムの相互比較・評価，属性の重みづけを勘案した変数の整理・統合が必要である。その上で，CEFR レベルを判別する言語特徴の具体的な特定と CEFR 準拠コーパスそのものの妥当性を検証したい。

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 Profile: Introducing the CEFR for English*, Version 1.1. 2011. UCLES / CUP.

[2] 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.

[3] 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.

[4] 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.

[5] North, B., A. Ortega and S. Sheehan. 2010. *A Core Inventory for General English*. British Council / EAQUALS.

[6] 佐野洋. 2007. 「多重の制約を利用した英語用例文の提示方式」『語学研究所論集』第 12 号, 東京外国語大学語学研究所, pp. 87-100.

[7] 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)』.

[8] 投野由紀夫(編). 2013. 『CAN-DO 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.

[9] WEKA 3.6.11. 2014. University of Waikato.