高校英語教科書のCEFRレベル：CEFR-J Wordlistに基づいた語彙の数量的分析

畔元, 里沙子
九州大学文学部

内田, 諭
九州大学大学院言語文化研究院

出版情報：Proceedings of the Annual Meeting of the Association for Natural Language Processing. 24, pp.468-471, 2018-03. 言語処理学会
バージョン：
権利関係：

高校英語教科書の CEFR レベル
—CEFR-J Wordlist に基づいた語彙の数量的分析—

畔元 里沙子* 内田 諭**
*九州大学文学部 **九州大学大学院言語文化研究院

1. はじめに

本研究は、高等学校で取り扱う英語教科書を対象に、学年ごとの CEFR レベルに基づいた語彙レベルの推移とその要因を探ることを目的としている。また、教科書に一定以上の頻度で出現する単語で CEFR-J Wordlist[1]に掲載されていない語群と、CEFR-J Wordlist において A1 レベルと分類されたものの教科書に出現していない語群について考察し、その傾向を探る。

CEFR-J Wordlist は、小・中・高・大の一貫する英語コミュニケーション能力の到達基準の策定とその検証を目的として、中国、台湾、韓国の小中高の主力教科書を CEFR 基準に大まかに分類したものをベースに、各国・地域の CEFR レベル・テキストに取り扱われている共通語彙を抽出したリストであり、日本の英語教育の文脈に合わせて作られたものである（cf. 投野［編］2013）。

英語教科書の語彙分析の研究は、村岡（2014）は 6 社の中学校検定教科書各学年 1 冊ずつの計 18 冊を対象に分析を行い、出版社ごとに使用する語群が大きく異なっており、6 社すべてで使用されている語は 6 社合計語彙数全体の 19.9%にあたる 441 語のみであることを明らかにした。山本（2016）は、文部科学省検定済教科書の語彙リストと話し言葉コーパスの比較を行い、話し言葉では高頻度に用いられているものの、教科書では上位 3000 語に含まれてない語彙についてリスト化し、その語群の特性について考察した。長谷川・中條・西垣（2008）は、1980 年代の教科書と 2000 年代の教科書における語彙の変化を分析し、高等学校教科書の延べ語数は調査対象すべての教科書において減少していることを明らかにし、高校生が触れる英語の分量が減少していることを明らかにした。

以上の研究では、3 学年まとめての語彙分析は行われているが、各学年での語彙レベルの推移については明らかにされていない。また、CEFR レベルとの関係性については議論されていない。よって本研究では、高等学校の英語教科書における各学年の語彙レベルの推移とその要因を探るために、まず近年発刊されている「コミュニケーション英語」の教科書を学年別にコーパス化する。その後、TreeTagger[2]で POS タグを付与し、機能語を除いた内容語に CEFR レベルを付与して各学年の CEFR レベルごとの単語の相対頻度の推移を品詞ごとに見ていく。最後に、CEFR レベルを付与することができなかった語と、CEFR レベルリストにおいて最も基本的な A1 レベルに分類されるものの中で教科書に出現しない語についての考察を加える。

2. 教科書コーパス（CETJ）

本研究で用いるデータは、「平成 28 年度使用都立高等学校後期課程用教科書教科別採択結果」[3]から、コミュニケーション英語の教科書を対象に採択シェアが約 8 割を満たすように第 1 学年から 12 冊、第 2 学年から 14 冊、第 3 学年から 11 冊を選び、それらを基に教科書コーパス CETJ (Corpus of English Textbooks in Japan)を作成し

[1]『CEFR-J Wordlist Version 1.3』 東京外国語大学投野由紀夫研究室.（http://www.cefr-j.org/download.html より 2017 年 12 月ダウンロード）

[2] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[3] http://www.metro.tokyo.jp/INET/OSHIRASE/2015/08/DATA/20p8r300.pdf

た（総語数 202,052）。オプショナルレッスンを除いたリーディングパートのみをコーパス化の対象とした。

CETJ では TreeTagger を使用して品詞タグ付けを行った。これにより、2 つ以上の品詞を持つ単語に対して POS 情報を基に CEFR レベルを付与することができる。例えば、advance の動詞は B1、名詞は B2 レベルに類されるが、これらを区別することが可能となる。表 1 は各学年の内容語の数を CEFR レベルごとに分別した結果を表している。また、表 2 は各学年のサブコーパスの総語数を基準に 1 万語あたりの相対頻度に直したものである。なお、#N/A は CEFR-J Wordlist に掲載がない等の理由で CEFR レベルが付与できなかった語を表す。

表 1　CEFR レベルと各学年の関係

ALL	1	2	3	総計
A1	11958	17713	17290	46961
A2	4066	7840	8378	20284
B1	2041	4787	5750	12578
B2	732	1814	2055	4601
#N/A	2958	5261	5747	13966
総計	**21755**	**37415**	**39220**	**98390**

表 2 CEFR レベルと各学年の関係（相対頻度）

ALL	1年	2年	3年
A1	2258.3	1998.6	1910.0
A2	767.9	884.6	925.5
B1	385.4	540.1	635.2
B2	138.2	204.7	227.0
#N/A	558.6	593.6	634.9

全ての学年において A1 レベルが最も多いが、その数は学年とともに減少し、A2 以上の語については学年と共に増加していることがわかる。

3. 品詞ごとの相対頻度の分析

表 3〜6 は内容語（名詞・形容詞・副詞・動詞）のそれぞれの品詞の相対頻度を表している。

表 3　名詞の相対頻度

名詞	1年	2年	3年
A1	1000.0	896.8	862.0
A2	356.2	400.4	412.3
B1	192.6	256.7	304.8
B2	91.0	118.9	120.1
#N/A	225.9	258.2	265.0

表 4　形容詞の相対頻度

形容詞	1年	2年	3年
A1	276.1	237.1	231.0
A2	95.9	112.8	121.5
B1	64.0	96.4	112.7
B2	12.1	27.1	33.8
#N/A	204.3	207.8	214.3

表 5　副詞の相対頻度

副詞	1年	2年	3年
A1	285.5	260.2	258.4
A2	131.1	139.0	149.6
B1	26.4	41.3	56.9
B2	4.5	9.5	13.5
#N/A	41.4	39.3	50.9

表 6　動詞の相対頻度

動詞	1年	2年	3年
A1	184.7	232.3	242.1
A2	102.4	145.8	160.8
B1	30.6	49.2	59.7
B2	87.1	88.3	104.6
#N/A	1101.4	1120.2	1125.9

名詞、形容詞、副詞においては、どの学年も A1 レベルの語の出現が最も多く、B2 レベルの語の出現が最も少ない。また、A2、B1 レベルはその中間に位置し、CEFR レベルの低い語ほど多く使われる傾向が読み取れる。また、A1 レベルは第一学年の使用頻度が最も高く、学年が上がるにつれて使用頻度は低下するが、A2 以上のレベルの語は学年が上がるにつれて使用頻度が上昇する。しかし、動詞のみ異なった推移を示している。動詞に限っては、B2 レベルの語彙の使用頻度のほうが B1 レベルの使用頻度よりも高い。また他の品詞では学年が上がるにつれて低下していた A1 レベルの使用頻度も、動詞のみは上昇している。

動詞のみ B2 レベルの使用頻度が B1 レベルの

使用頻度よりも高い要因として考えられるのは、CEFR-J Wordlist に含まれている B1 レベルの動詞の種類（464）が B2 レベルの動詞の種類（547）より少ないこと、そして教科書から B1 レベルの動詞、例えば appoint や exit などが抜け落ちていることが挙げられる。また、A1 レベルの動詞の使用頻度が学年とともに上昇する要因としては、学年が上がるにつれて様々な文法を取り入れた構文が増え、文の構造に動詞を含む修飾部が増えるなどして、一文が複雑且つ長くなり、動詞の密度が上がったことが考えられる。I know (that) ~の文構造を例にとると、第一学年では I knew he wouldn't forget us. （『ELEMENT English Communication I』啓林館）であるのに対して、第三学年では I knew without looking that the papers were the ones on which I had listed all the good things each of Mark's classmates had said about him. （『ELEMENT English Communication Ⅲ』啓林館）などのようになっている。

4. CEFR レベルが付与できない語群

CEFR レベルを機械的に付与することができなかった単語が一定数存在する。それらは表 1~6 で#N/A と示された語群であり、表 2 から 5%～6%の割合を占めることがわかる。高校で扱う教科書であることを考慮すると、特異な単語は出現しないと予想されるが、以下ではその原因を探る。

CEFR レベルを付与することができなかった単語は 3654 種類であったが、主に二つのパターンがある。まず一つ目は、CEFR レベルを付与する前に行う TreeTagger が付与する POS タグと CEFR-J Wordlist の POS タグの差異によるものである。例えば、many は CEFR-J Wordlist では determiner（決定詞）としてのみ記載されているが、TreeTagger の POS タグでは JJ（形容詞）が付与される場合がある。CEFR レベルが付与されなかった単語のうち頻度が高かった上位 20 単語中 17 単語はこのパターンに分類される。二つ目は、CEFR-J Wordlist に含まれるべきであると考えられる語が含まれていないことである。例えば English, Japanese などの国名の形容詞形などは CEFR-J Wordlist に掲載されていない。これらの単語は固有表現であるため、掲載有無の境界線の判断が難しい場合であるが、その一方で aborigine など非常に固有的であると思われる単語が CEFR-J Wordlist に含まれていることを考えると、頻度の高いものに関しては含める余地が十分にあると言えるだろう。

さらに diversity, sustainable, globalization など、重要語が欠落している場合がある。これらの単語は英語教育の現場でよくトピックとして扱われるものであるため、CEFR-J Wordlist に含める妥当性は高いと考えられる。また、これらの単語は図 1 の Google N-gram Viewer[4]からわかるように 1990 年以降、急激に使用頻度が高くなってきた語である。

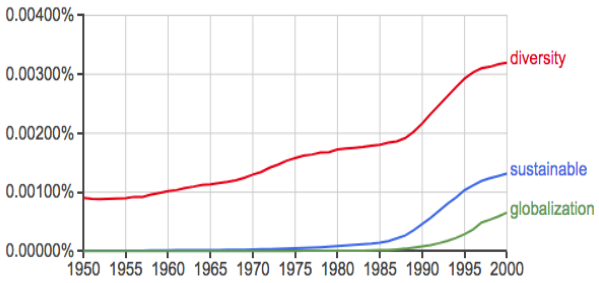

図 1 Google N-gram Viewer の検索結果

例えば globalization という単語は、1991 年にソビエト連邦が消滅、アメリカ合衆国の単独覇権が確立したことにより、社会主義が消滅、自由貿易の拡大やグローバル企業の誕生を迎えたことをきっかけに世界に広く普及していったものだと考えられ、昨今も広く用いられる。従って、この単語をリストに含めることは十分に妥当であると言え、時代背景に合わせた語彙リストの更新の必要性を示唆する。

[4] https://books.google.com/ngrams

5. 教科書に出現しない語群

前節では、教科書には出現するが CEFR-J Wordlist には含まれていない語の考察をおこなったが、本節では逆に CEFR-J Wordlist に掲載されているが、教科書には出現しない単語について分析する。ここでは CEFR-J Wordlist で最も基本的な単語(A1)と分類されているにも関わらず、教科書で扱われていない単語に注目する。

CEFR-J Wordlist に載っていて教科書に載っていない単語は 7890 件中 2630 件（約 33%）である。そのうち A1 レベルは 220 件であり、107 件は名詞である。月名（e.g. August, February）、曜日名（e.g. Monday, Saturday）、スペースを含む単語を除いた 83 件を対象とすると[5]、kick, shake, try などが目立つ。これらは主に動詞としての用法を持つ語であるが、教科書では、名詞として使用されることがないことが示唆される。次に、衣服に関係がある語（e.g. jeans, pants）と、食べ物や飲み物に関係がある語（e.g. burger, butter）が目立つ。よって、身の回りにある具体物の名称が教科書には欠けていると言える。

対して、最も少ない品詞は動詞であり、タグ付けの誤りを除けば、A1 レベルの動詞の中で教科書に出現しないのは phone のみ（名詞としては出現あり）となった。このことから、教科書には基本的な動詞は網羅されていると言える。

6. まとめ

本研究では、教科書コーパス CETJ の作成を行った後に、各学年の CEFR レベルごとの推移を分析した。その結果全体の語彙レベルとしては、A1 レベルは学年とともに相対頻度は低下し、A2 以上のレベルは学年とともに上昇することがわかった。また、品詞ごとの分析から、動詞のみがその他の品詞と比べて異なった語彙レベルの推移があることが明らかになった。さらに、CEFR-J Wordlist の中には国名などの固有的単語や、現代の英語教育現場で頻繁に使用される単語の一部が欠落していた。また、教科書には動詞としての用法を持つ語の名詞用法や、身の回りの具体物の名称が欠落していることがわかった。これらの結果は、CEFR-J Wordlist の改訂に向けて示唆のあるものだと言えるとともに、教科書の内容の改善にも有意義なものである。今後の課題としては、一単語のみの語彙レベル分析だけではなく、複数の語が組み合わさったフレーズを含めた分析などが挙げられる。

［謝辞］
本研究は JSPS 科研費 JP15K16798 および JP16H01935 の助成を受けたものである。

［参考文献］

長谷川修治・中條清美・西垣知佳子(2008).「中・高英語検定教科書語彙の実用性の検証」『日本大学生産工学部研究報告 B』41, 49-56.

村岡亮子(2014).「中学校検定教科書で学習される語彙,学習されない語彙：延べ語数,異なり語数,語彙レンジの視点から」『英検研究助成報告書』22, 182-203.

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

山本五郎(2016).「文部科学省検定済教科書語彙リストに関する研究：中学・高校での指導目安 3,000 語レベルの語彙リストの分析」『広島外国語教育研究』19, 43-55.

[5] 本研究では単語単位でレベル付与を行ったため、according to などのフレーズとして掲載されているものは考察の対象外とした。