九州大学学術情報リポジトリ
Kyushu University Institutional Repository

CEFR-J Text Profileとは何か

内田, 諭
九州大学大学院言語文化研究院 : 准教授

https://hdl.handle.net/2324/1932324

出版情報：英語教育. 67 (3), 2018-06. 大修館書店
バージョン：
権利関係：

CEFR-J Text Profile とは何か

CEFR-J Text Profile は、CEFR レベル別の海外の EFL コースブックのコーパスを基に、各 CEFR レベルにおける語彙および構文の特徴を明らかにすることを目的に作成されたものである。A1, A2, B1, B2, C（C2 レベルのデータが少ないため 1 つに統合）の 5 段階で、A1 レベルの語がどの程度の割合で出現するか、文の平均的な長さはどの程度か、一文あたりに名詞句はどの程度出現するか、などについての基本統計量（平均・分散など）を示している。これを参照することで、例えば「B1 レベルのテキストの語彙レベルの分布」などを知ることができ、後述するアプリケーションを使えば、手元にある英文のレベルをテキスト特徴からおおよそ知ることが可能となる。

Text Profile の指標

CEFR-J Text Profile が提供する指標には大きく語彙に関するものと構文に関するものがある。語彙指標としては、例えば各レベルの単語 [1]の割合（A1_per, A2_per など）や単語の長さの平均（avg_word_length）、単語の種類の平均（type）などがある。表 1 を見ると、例えば B1 レベルの語の割合はテキストのレベルが上がるに連れて上昇し、A1 レベルでは 4%であるのに対して、B1 レベルでは倍近くの 7.6%になることがわかる。また、上位レベルの語（C1、C2）の割合は、CEFR レベルが上がっても出現率は高くない（1%以下）ことがわかり、C1、C2 レベルの語の割合はテキストレベルの差としてあまり重要ではないことが読み取れる。

一方、構文指標には、文あたりの動詞句数の平均（avg_VP）や関係節数の平均（avg_REL）、文の木構造の深さの平均（avg_max_depth）などが含まれている [2]。これらの構文指標は B1 レベルで顕著に高くなる傾向が読み取れる。このことは B1 レベルからより複雑な文構造が出現する傾向にあることが示唆される。

各素性のレベル弁別性

CEFR-J Text Profile に含まれている素性がどの程度 CEFR レベルの弁別に寄与しているかということを明らかにするため、統計的手法を用いて実験を行った [3]。その結果、A1 と A2 レベルの弁別には A1_per、A2_per などの語彙指標、A2 と B1 の弁別には avg_VP、avg_max_depth など構文指標が強く貢献することがわかった。また、B1 と B2 の弁別には語彙指標および構文指標の双方が寄与するということが明らかになった。このことから「A2 レベルのテキストは A1 レベルと比べて特に語彙的な違いがある」、「B1 レベル以上のテキストでは構文的な違いが出てくる」などの傾向

[1] CEFR-J Wordlist（東京外国語大学投野由紀夫研究室作成）と English Vocabulary Profile（詳細は内田 2017 を参照）をベースに作成したリストに基づきレベルを付与。

[2] 句構造解析には Enju (http://www.nactem.ac.uk/enju/)を利用。

[3] 詳細は水嶋 海都・荒瀬 由紀・内田 諭 (2016)「CEFR 準拠教科書における語彙・構文の特徴分析とレベル自動分類」『言語処理学会第 22 回年次大会発表論文集』pp.789-792.を参照。

が読み取れる。

CEFR-J Text Profile のアプリケーション

前述のような語彙・構文指標を利用することで、テキストのレベルを判定することができる可能性がある。そこで筆者はオンラインのテキストレベル判定システム CVLA（CEFR-based Vocabulary Level Analyzer）を開発し、公開している[4]。ただし、現時点ではベータ版であるため、解析にはキーワードの入力を求めている（ETM201806 でお試し頂けます）。CVLA では入力文章の単語を CEFR レベルごとに色分けして表示し、Text Profile の指標を複合的に用いた独自の 4 つの指標を基にテキストレベルを判定する。これらの指標には、語彙指標として AvrDiff（A1=1, A2=2, B1=3, B2=4 としたときの内容語の難易度平均）と BperA（A レベルの内容語に対する B レベルの内容語の割合）、構文指標として ARI（リーダビリティ指標）と VperSent（文あたりの動詞の数の平均）が含まれている。例えば、平成 29 年度のセンター試験の英語（筆記）第 5 問（物語文）は ARI=1.57 (PreA1 相当), VperSent=2.00 (A2.2 相当), AvrDiff=1.51 (B1.1 相当), BperA=0.16 (B1.2 相当)となり、総合すると文章レベルは A2.1 という判定になる。レベルの判定は暫定的であり、今後さらなるチューニングが必要ではあるが、この文章は文構造は読みやすいが、語彙レベルは B1 レベルになっているということが読み取れる。

CEFR-J Text Profile の意義と今後

CEFR-J Text Profile を参照することで語彙や文構造などの側面から CEFR の各レベルのテキストのプロファイルが明らかになる。この情報は、ターゲットのレベルに合わせた教材や試験の作成に大いに役に立つだろう。例えば、現行の教科書ではそれぞれ語彙の使用基準が異なるという問題があるが（投野 2016）、CEFR-J Text Profile は語彙統制のための一つの基準になりえる。また、民間の英語能力試験の難易度を CEFR レベルに換算する指標の一つとしても活用できるだろう。今後はより精度の高いレベル判定を達成するため、CEFR-J Grammar Profile の情報と組み合わせてテキストレベルの判定を行うシステムの開発を行っていきたいと考えている。

内田諭(2017)「English Vocabulary Profile を語彙指導に活用する」『英語教育』2017 年 2 月号, pp32-34.

投野由紀夫(2016)「教科書語彙の「調理法」と「品質管理」：中学校改訂版教科書の語彙レベルと語数」『英語教育』2016 年 2 月号, pp.17-19.

[4] http://dd.kyushu-u.ac.jp/~uchida/cvla.html

表 1 各 CEFR レベル別の単語の割合

	Features	A1_per	A2_per	B1_per	B2_per	C1_per	C2_per
Text Level	A1	0.858	0.084	0.040	0.011	0.001	0.001
	A2	0.800	0.119	0.058	0.018	0.001	0.001
	B1	0.746	0.147	0.076	0.025	0.002	0.001
	B2	0.692	0.165	0.096	0.039	0.004	0.003
	C	0.669	0.171	0.105	0.044	0.005	0.004