D-005

リーディング用英語学習教材の難易度推定手法の検討
Estimating Readability Level of English Text for English-as-Foreign-Language (EFL) Learners

髙橋 里紗[1)]　来住 伸子[1)]
Risa Takahashi　Nobuko Kishi

1　はじめに

英語を母国語としない英語学習者にとって, 適した難易度で興味深い内容のリーディング教材を利用することは非常に重要である. 外国語としての英語学習者が判断する英語テキストの難易度と, ネイティブスピーカーにとっての難易度は大きく異なるが, これまでの研究の多くは, ネイティブスピーカーが判断したテキストの難易度を評価している. 本研究では, リーディング用英語学習教材を用いて, 英語が母国語ではない英語学習者にとって適切な難易度を推定する手法を検討する.

英語リーディング教材として, TED Talks[1] は内容も表現も英語の学びにおいて有用だと考えられている. 本研究では, TED Talks を用いた英語学習教材に掲載されている TED Talks のスクリプトを対象に, Flesch-Kincaid Readability Ease という古典的な手法で, リーダビリティを推定することを試みる. 推定した難易度は, 外国語の運用能力を同一の基準で測ることが出来る国際標準 CEFR(Common European Framework of Reference for Languages)[2] を, 日本の英語教育に合わせて構築した CEFR-J[3] レベルの語彙リストなどを用いて検討する. また, 英語学習教材のレベル分けと比較検討する.

2　先行研究・関連研究

2.1　TED Talks に関する研究

長谷部 [4] は, TED コーパスを自在に検索し, その結果を様々な形式で表示することができるシステム「TED Corpus Search Engine(TCSE)」を実装した. TCSE の機能の 1 つに, トークの相対的な難易度指標の表示がある. この難易度指標には Flesch-Kincaid Readability Ease を採用している.

2.2　難易度に関する研究

Xia ら [5] は, 第二言語 (L2) としての英語学習者向けに作成された CEFR グレードのテキストのデータセットを収集し, ネイティブと L2 学習者の両方に対するテキストの読みやすさ評価について調査した. 言語学習者の読みやすさに合わせたテキストを含むデータセットを収集し, より大規模な既存のネイティブコーパスで学習したモデルを学習者のテキストの読みにくさを推定する際に適応させる方法を検討した. その結果, 読みやすさの推定は線形 SVM を用いて, ネイティブデータと L2 データで最良の性能を達成するシステムを開発した.

内田ら [6] は, 入力テキストから算出されるテキスト特徴量に基づいて, 英語テキストに CEFR-J レベルを付与するシステムである「CVLA」の構築を試みた. 文構造と語彙に関連するテキスト特徴量から構築した 4 つの回帰モデルを用いて, 英語パッセージのレベルを推定した. CVLA の推定値は概ね妥当なものであり，400 語程度の入力があれば，大きくずれたレベルを判定することは少ないことが明らかとなった.

3　方法

3.1　使用データ

3.1.1　対象とするデータ

本研究では,『21st Century Reading』[7] に掲載されている 40 件の TED Talks のスクリプトを対象とする.『21st Century Reading』は, TED と提携をして作られたリーディングテキストである. テキストの難易度に応じて 4 段階のレベルに分かれており, 各レベルは表 1 のように CEFR レベルに対応している. 以下, このレベルを「21st レベル」と呼ぶ. 文部科学省による, 各資格・検定試験と CEFR との対照表 [8] を表 2 に示す.

対象の TED Talks のスクリプトは, TED Talks のホームページから収集し, NLTK[9] を用いて解析する.

表 1　『21st Century Reading』[7] と CEFR レベルの対応

教材の難易度	CEFR レベル
Level1	B1
Level2	B1-B2
Level3	B2
Level4	B2-C1

表 2　各資格・検定試験と CEFR との対照表 (出典：文部科学省)[8]

CEFR	実用英語技能検定	IELTS	TOEIC L&R/ TOEIC S&W
A1	3 級-準 2 級	-	320-620
A2	準 2 級-2 級	-	625-1145
B1	2 級-準 1 級	4.0-5.0	1150-1555
B2	準 1 級-1 級	5.5-6.5	1560-1840
C1	1 級	7.0-8.0	1845-1990
C2	-	8.5-9.0	-

3.1.2　語彙リスト

本研究では学習者向けの語彙リストとして CEFR-J Wordlist[10] と Academic Word List[11] の 2 種類を使用する.

CEFR-J Wordlist とは, 投野由紀夫らの『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』プロジェクトの一環で構築された語彙リストである. 語彙のレベルは A1, A2, B1, B2 の 4 段階である. なお, CEFR-J Wordlist には, 語彙の見出し語, 品詞, CEFR-J レベルが記載されている. 本研究では, ”a.m” と”A.M”のような同一字母の語は, 品詞と CEFR-J レベルが一致している場合は, 同じ単語と見なす.

Academic Word List(以下 AWL) とは, ニュージーランドの言語学者が外国人学生のために開発した, 大学レベルで使用する語の頻出単語リストである.

表 3 に 2 種の語彙リストの語彙数を示す.

1)　津田塾大学 https://www.tsuda.ac.jp/

表 3　語彙リストの語彙数

	CEFR-J A1	CEFR-J A2	CEFR-J B1	CEFR-J B2	AWL
語彙数	1194	1443	2486	2859	570

3.2　Flesch-Kincaid Readability Ease の計算

Flesch-Kincaid Readability Ease（Flesch, R., 1948, 以下 FK-RE) は有名なリーダビリティの指標の 1 つである. この値が大きいほど読みにくい文章だと考えられる. FK-RE は次の式で計算する.

$$206.835 - 1.015\frac{totalwords}{totalsentences} - 84.6\frac{totalsyllables}{totalwords}$$

先行研究である長谷部 [4] による TED Talks スクリプトの FK-RE と, 本研究で算出した FK-RE を比較し, FK-RE と 21st レベルの対応関係を図示する.

4　結果

4.1　語彙分布と 21st レベルの対応

5 種類の語彙リスト (CEFR-J Wordlist A1, A2, B1, B2, AWL) のうち, 教材で使用されている語彙の数を図 1 に示す. CEFR-J Wordlist A1 レベルの語はどのレベルでもよく使われていることがわかる. また, Level4 においては, どの語彙リストの語も多く使われている.

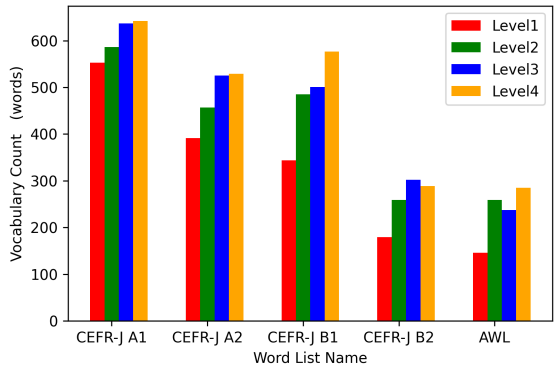

図 1　語彙の分布

4.2　FK-RE と 21st レベルの対応

先行研究 [4] の FK-RE を縦軸, 本研究の FK-RE を横軸とした散布図を図 2 に示す. 各点は 21st レベルごとに色分けをしている. 本研究の FK-RE は, 先行研究 [4] の FK-RE とほぼ一致していることがわかる. FK-RE の計算に関して, 先行研究と同様の値が得られた. 各点の分布から, FK-RE による難易度と 21st レベルは, 必ずしも一致しなかった. これは, FK-RE は単語数や文章数, 音節数から算出しており, 語彙の難易度や文法構造は考慮していないためだと考えられる.

5　まとめ・今後の課題

英語学習者の読みやすさに焦点を当て, 外国語として英語を学ぶ人のために選ばれた 40 個の TED Talks のスクリプトを対象に, Flesch-Kincaid Readability(FK-RE) などの手法で, 難易度を推定することを試みた. 推定した難易度は, 英語学習教材のレベル分けや CEFR レベルと比較検討した結果, 必ずしも一致せず, FK-RE だけではレベルの分類ができないことがわかった. その原因として, FK-RE は単語数などの基本的な文章の特徴から計算した値であるため, 語彙の難易度や文法構造を考慮できていないことが考えられる. 今後は, 他の文章の特徴量 (文の長さ, 文構造など) を追加して, 機械学習で文章の難易度を推定するなど, より良い推定方法を検討する.

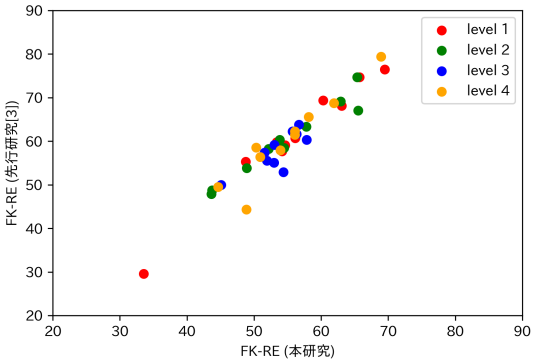

図 2　FK-RE を使った散布図

参考文献

[1]TED Talks. https://www.ted.com/talks

[2]CEFR. https://www.coe.int/en/web/common-european-framework-reference-languages

[3]CEFR-J. https://www.cefr-j.org/

[4]長谷部陽一郎.(2017) ”TED Corpus Search Engine: TED Talks を研究と教育に活用するためのプラットフォーム”. 英語コーパス学会第 43 回大会シンポジウム, pp.159-172.

[5]Menglin Xia, Ekaterina Kochmar and Ted Briscoe. (2016) ”Text Readability Assessment for Second Language Learners”.Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. pp.12-22.

[6]Satoru Uchida and Masashi Negishi.(2018) ”Assigning CEFR-J Levels to English Texts Based on Textual Features”. Proceedings of Asia Pacific Corpus Linguistics Conference. 4, pp.463-467

[7]Robin Longshaw, et al. *21st Century Reading*. National Geographic Learning, Cengage Learning Company. https://eltngl.com/sites/21st-century-reading/student

[8]文部科学省.「各資格・検定試験と CEFR との対照表」. https://www.mext.go.jp/b_menu/houdou/30/03/__icsFiles/afieldfile/2019/01/15/1402610_1.pdf

[9]NLTK. https://www.nltk.org/

[10]『CEFR-J Wordlist Version 1.6』 東京外国語大学投野由紀夫研究室.（http://www.cefr-j.org/download.html より 2022 年 5 月ダウンロード）

[11]The Academic Word List. https://www.wgtn.ac.nz/lals/resources/academicwordlist/links