CEFR レベル別英語教科書における基準特性の重要度：CVLA の指標を用いて

畔元、里沙子
九州大学大学院言語文化研究院 ： 教授

内田, 諭
九州大学大学院言語文化研究院 ： 准教授

https://doi.org/10.15017/4773109

出版情報：言語文化論究. 48, pp.35-47, 2022-03-17. Faculty of Languages and Cultures, Kyushu University
バージョン：
権利関係：

CEFR レベル別英語教科書における基準特性の重要度
――CVLA の指標を用いて――

畔元里沙子・内田　　諭

１．はじめに

現在までに、外国語習得における適切な教材選択を目的として、インプット英文の自動レベル判定の研究が盛んに行われてきた。代表的なレベル判定の手法としては、テキストのリーダビリティを客観的な基準から計算し数値化することであり、１文あたりの平均単語数や１単語あたりの平均音節数から計算される Flesch Reading Ease Formula や Flesch-Kincaid Grade Level が代表的である。Flesch Reading Ease Formula は 0 から100までの数値で結果が出力され、数値が大きければ大きいほど対象のテキストが簡単であることを示す。対して、Flesch-Kincaid Graded Level の計算結果は米国の学年制度で表される。例えばスコアが8.2であれば、対象の英文は米国の平均的な８年生（日本の中学２年生）が理解できるレベルであるといえる。他にも、100単語あたりの平均文字数と100単語あたりの平均文数から計算される Coleman-Liau Index（Coleman & Liau, 1975）、文字数、単語数、１文あたりの平均単語数から計算される Automated Readability Index（Senter & Smith, 1967）等がある。水本（2013）はこれらの従来から使用されているリーダビリティ指標や Type-Token Ratio（総語数に対する異なり語の数）等といった表面的な特徴に加え、英文全体で文や句の結束性を示す指標を計算することができる Coh-Metrix（McNamara et al., 2010）を使用し、「語が具体的でイメージしやすいか」や「因果・論理関係を示す接続語がどの程度含まれているか」といった英文の意味的内容にまで踏み込んだ指標を基に教材の難易度推定を試みている。日本人英語学習者に向けては、SSS 英語多読研究会では13,000冊以上の多読教材に対して独自の基準で0.0から9.9までの100レベルに分類し、その基準を YL（Yomiyasusa Level）と名付けた（古川他、2010）。

しかし、これらの難易度推定システムが算出する推定結果の多くは、米国の学年制度や独自の単位で示されており、英語学習者にとって自身が選択すべきテキストレベルを正確に把握することが難しいことに加え、指導者にとっても学習者のレベルと照らし合わせて適切な教材を選択することが困難であると考えられる。そこで近年では、英文の難易度評価の結果を、ヨーロッパを中心に世界中で広く使用されている CEFR（Common European Framework of Reference for Language）レベルに落とし込む研究が多く行われるようになった。例えば、Smith & Turner（2016）は、「英文を読む能力」と「文の複雑さ」の両方を独自の単位で測ることができる Lexile Measure と CEFR レベルの関係を明確にしようと試みた。その結果、教材の作者が独自に設定した CEFR レベルと Lexile Measure の関係、ならびに TOEFL や TOEIC の点数から予想される学習者の CEFR レベルと Lexile Measure との関係を表にまとめている。しかし、Lexile Measure は教材の難易度の判定だけではなく学習者がどの程度のレベルの教材まで理解することができるかを測る目的で開発されたものであり、学習

者がどの程度のことを行うことができるかの指標である CEFR とは性質が異なることを Smith & Turner（2016）は自身で指摘している上、内田・根岸（2021）はそもそも Lexile 値の算出方法が公開されておらず、学術的な検証を行うことができないという欠点を挙げている。

他には、文の数、語数、異語数、音節、TTR、平均文長、文字数などといったテキストの表面的特徴に加えて、Flesch Reading Ease、Flesch-Kincaid Grade、Gunning Fox Index 等のリーダビリティ指標等から CEFR レベルを推定する Text Inspector が存在するが、Lexile Measure と同様、内田・根岸（2021）から CEFR レベル判定に関するアルゴリズムが公開されていないという指摘を受けている。

このような状況を受け、Uchida & Negishi（2018）は CVLA（CEFR-based Vocabulary Level Analyzer）を開発した。これは入力テキストの CEFR-J レベル（CEFR を日本の英語教育に適用させた枠組み）を推定するオンラインアプリケーションである。推定方法としては、CEFR レベルが付与された教材コーパス（Coursebook Corpus）を作成し、そのデータを基に ARI（リーダビリティ指標）、VperSent（１文に含まれる動詞の平均数）、AvrDiff（CEFR-J Wordlist[1]基づく平均語彙レベル）、BperA（A レベルの内容語に対する B レベルの内容語の割合）という４つの指標を、CEFR レベルごとに分類した各サブコーパスでの平均値を算出し、基準を設定するという手法がとられている。テキストを入力すると利用者は４つの指標の値とそれに準ずる４つの CEFR レベルを得ることができ、その４つの平均レベルが最終的なテキストレベルとして与えられる。CVLA の最大の特徴としては CEFR レベルの推定までの過程が明確であることが挙げられる。CVLA は CEFR レベルの推定に使用されている尺度が４つと少ないながら、内田・根岸（2021）によると、テキストの CEFR レベルを約53％の精度で推定する。隣接レベルを正解として扱った場合、約80％の精度でレベルを推定することができており、これらの４つの指標を用いた推定が安定していることが分かる。

しかし、内田・根岸（2021）が指摘している通り、レベル別の正答率をみると、A1 が最も高い（100％）ことに対して C2が最も低く（約21％）、レベルごとに正答率にばらつきがあることも明らかである。内田・根岸（2021）は C1、C2 の正答率が低いことの原因に CVLA が依拠している教材コーパスのデータが B2までであり、C1、C2 に関しては回帰直線を伸ばして推定していることを挙げているが、B1（約59％）、B2（約54％）の正答率の低さや、結果のばらつきに関しては原因を言及していない。正答率がレベルによって大きく異なることの原因として、隣接したレベルの分類を行う場合に、各レベル間で最も効果的な指標が異なることが挙げられる。つまり、A1 から A2、A2 から B1、B1 から B2、B2 から C1 にレベルが上がる際に最も影響力が強い指標が異なる可能性を CVLA は考慮していないと考えられる。

本研究は、テキストの難易度が上がる際にどの指標が最も影響力が強いかを各隣接レベル間で明らかにすることを目的とし、予備実験として単一分類モデルにおけるレベル間でのテキストの分類の成否を実証的に検証した後に、各隣接レベル間で分類モデルを新たに作成し、各指標の変数重要度を比較し分析を行う。

２．CEFR 準拠テキスト

2.1　データ

本研究を行うにあたり、自作の CEFR 準拠教材コーパスの一部を使用した。このコーパスは2014年から2020年の間に Pearson、Oxford University Press、Cambridge University Press、Cengage から出版された、CEFR レベルが付与されている英語教材のうち、A1から C1レベルのリーディングパー

トのみをコーパス化したものである。C2レベルの教材は数に限りがあるため、データに含まなかった。また、このコーパスには「A1-A2」といった横断的なレベルが付与されているテキストも含まれているが、本研究ではそのようなデータは除き、「A1」「A2」「B1」「B2」「C1」と明記されているもののみを使用した。なお、CVLA の開発の際に使用されたデータと重複はない。表１は本研究で使用したコーパスの中での教材数、文章数[2]、単語数をレベルごとにまとめたものである。

表１　使用したサブコーパスのレベル別の教材数、文章数、単語数

CEFR レベル	教材数	文章数	単語数
A1	7	102	17,779
A2	7	131	33,312
B1	8	156	60,220
B2	7	125	64,932
C1	5	110	79,829

2.2　レベル別コーパスにおける４指標の分析

前述の CEFR レベル付きの文章に対して、CVLA を使用して ARI、VperSent、AvrDiff、BperA の４つの指標を各テキストに付与した。表２は CVLA における各 CEFR レベルにおける各指標の平均値、表３は本研究にて作成した CEFR 準拠教科書コーパスにおける各指標のレベルごとの平均値とレベル間での増加量および増加率を示している。平均値をみると、すべての指標において各レベル間で値の増加が見られ、この４指標は明確にテキストのレベル上昇の特徴をとらえた有効な指標であることを示している。増加量を詳しく見ると、レベル間で同量ずつ数値が増加しているのではなく、各レベル間、各指標で異なる値の増加をたどっていることが分かる。これは、レベル間の差は同じではないことを示す。例えば、ARI の増加量を見ると A1 から A2 にテキストレベルが上がる際には約2.75の値の増加が見られるが、A2か B1では約1.59、B1から B2では約1.76、B2から C1では約1.24と、ARI の値の増加量がレベル間で同量ではないことが分かり、ARI の観点からみると A1から A2が最もレベルの上がり具合が大きいことが分かる。同様に考えると、VperSent では A2から B1が、AvrDiff では A1から A2が、BperA では B2から C1が最も数値の上がり具合が大きい。

表２　CVLA における CEFR レベルごとの各指標の平均値

CEFR	ARI	VperSent	AvrDiff	BperA
A1	5.73	1.49	1.31	0.08
A2	7.03	1.82	1.41	0.12
B1	10.00	2.37	1.57	0.18
B2	12.33	2.88	1.71	0.26

表３　自作コーパスにおける各指標のレベルごとの概要

CEFR	ARI			VperSent			AvrDiff			BperA			増加率合計
	平均値	増加値	増加率	平均値	増加値	増加率	平均値	増加値	増加率	平均値	増加値	増加率	
A1	2.50			1.55			1.29			0.08			−
A2	5.25	2.75	37.50	2.15	0.60	33.15	1.45	0.16	30.05	0.13	0.05	19.47	120.17
B1	6.84	1.59	21.66	2.76	0.61	33.57	1.58	0.13	24.04	0.19	0.06	22.44	101.71
B2	8.60	1.76	23.96	3.02	0.26	14.46	1.70	0.12	23.02	0.26	0.07	25.89	87.33
C1	9.84	1.24	16.88	3.36	0.34	18.82	1.83	0.12	22.89	0.35	0.09	32.20	90.78

次に、総合的なレベル間の差を考えるために 4 指標すべての増加率に注目する。増加率は（増加値 /A1から C1までの増加量）*100で計算されており、A1から C1までの増加量を基とするとき、各レベル間での増加量は何％を占めるのかを表している。よって、A1 から C1 までの各指標の増加率を合計すると 100 となる。ここで 4 指標のレベル間の増加率の合計を計算すると、A1-A2 間では約120.17、A2-B1間では約101.71、B1-B2間では約87.33、B2-C1間では約90.78となる。このことから、CVLA における CEFR レベル判定と同様に 4 指標がテキストレベルに与える影響が等しいと仮定した場合、A1-A2間のレベル差が最も大きく、続いて A2-B1間、B2-C1間、そして B1-B2間が最もレベル間の差が小さいと判断できる。

次に、各指標の値のばらつきを明らかにするために、各テキストに付与された 4 指標に関して、レベルごとに分散分析を行った。表 4 は各指標の平均値、第一四分位数、中間値、第三四分位数、第一四分位数から第三四分位数のレンジをまとめたものであり、図 1 は表 4 を基に箱ひげ図を表している。また、表 5 はテキストレベル間で各指標の値に有意な差があるかどうかを検定するために一元配置分散分析（R ver. 4.1.0, aov 関数を利用）を行った結果を表したものである。

表4　自作コーパスにおける各指標の基本統計量

ARI	平均値	第一四分位数	中間値	第三四分位数	レンジ
A1（n=102）	2.50	0.96	2.52	4.35	3.40
A2（n=131）	5.25	3.55	5.25	6.73	3.18
B1（n=156）	6.88	5.25	7.00	8.49	3.25
B2（n=125）	8.60	6.93	8.49	10.15	3.22
C1（n=110）	9.84	8.20	10.16	11.16	2.96

VperSent	平均値	第一四分位数	中間値	第三四分位数	レンジ
A1（n=102）	1.55	1.25	1.49	1.82	0.57
A2（n=131）	2.15	1.75	2.12	2.38	0.63
B1（n=156）	2.73	2.33	2.74	3.09	0.76
B2（n=125）	3.02	2.67	3.07	3.40	0.73
C1（n=110）	3.36	2.97	3.32	3.72	0.74

BperA	平均値	第一四分位数	中間値	第三四分位数	レンジ
A1（n=102）	0.08	0.03	0.06	0.10	0.07
A2（n=131）	0.13	0.07	0.12	0.16	0.09
B1（n=156）	0.19	0.11	0.18	0.24	0.13
B2（n=125）	0.26	0.20	0.24	0.32	0.12
C1（n=110）	0.35	0.28	0.33	0.44	0.16

AvrDiff	平均値	第一四分位数	中間値	第三四分位数	レンジ
A1（n=102）	1.29	1.18	1.25	1.37	0.19
A2（n=131）	1.45	1.35	1.44	1.55	0.21
B1（n=156）	1.58	1.44	1.56	1.70	0.26
B2（n=125）	1.70	1.58	1.69	1.81	0.23
C1（n=110）	1.83	1.73	1.82	1.94	0.21

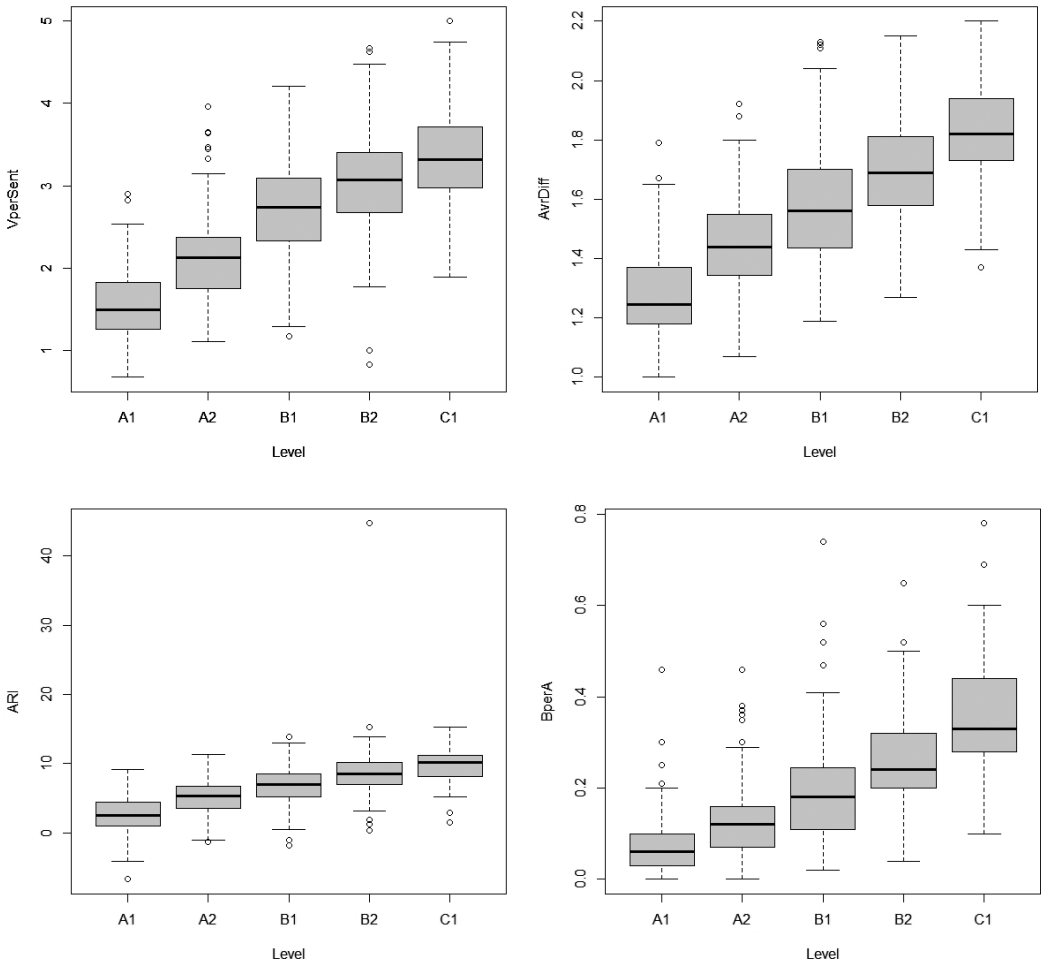

図１　各指標のレベル別箱ひげ図

表５　各指標の値とテキストレベルにおける分散分析の結果

	平方和	平方平均	F 値	p 値
ARI	3615.0	903.7	102.1	< 0.01
VperSent	227.5	56.9	171.9	< 0.01
AvrDiff	19.5	4.9	173.1	< 0.01
BperA	5.2	1.3	134.0	< 0.01

箱ひげ図の結果から、４指標とも同じレベルであってもある程度のばらつきが認められる。これは、複数の出版社が独自の判断で教材に CEFR レベルを付与していることに起因すると考えられる。第一四分位数から第三四分位数までの幅の中に、中央値に近い約半数のテキストが含まれるので、今後 CEFR レベルに準ずる新たなテキストを作成する際には、テキストの４指標の値を、表４を参考に第一四分位数から第三四分位数の中におさまるように調整することが望ましいことが示唆される。また、分散分析では全ての４指標の値のばらつきが偶然によるものではなくテキストレベル間での差が有意であることを認める結果となった（p<0.01）。[3] よって、この４指標はテキストレベルを測るうえで有効であると言える。

３．単一分類モデルを使用した A1から C1のレベル判定

本章では、CVLA に使用されている４つの指標を利用した A1から C1のテキスト分類モデルを、決定木を利用して作成した後にその精度を測定する。

3.1 手法

まず、A1から C1のテキストを訓練データと評価データに７：３の割合（訓練データ 436個、評価データ 188個）で R（ver. 4.1.0）の rsample パッケージを利用してランダムに分類した。その後、訓練データを使用して、４つの指標によるテキストの CEFR レベル予測の決定木を rpart パッケージ（デフォルト）を使用して作成した。その決定木の精度を測るために評価データに対して決定木に基づいて CEFR レベルを予測し、その結果がテキストの CEFR レベルと一致しているかの分類精度を測定した。図２は実験の概要を示している。

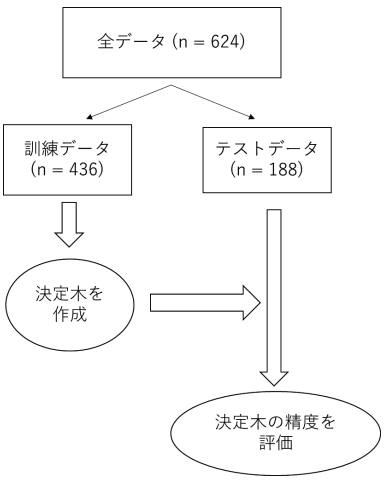

図２　実験の概要

3.2 結果と考察

図３は訓練データ（n=436）に基づいて作成した決定木を図式化したものである。最も上の接点は全データを、最下層のグラフの x 軸は分類を、y 軸はその割合を示している。rpart パッケージが計算したそれぞれの指標の変数重要度は大きいものから VperSent=32、BperA=27、AvrDiff=25、ARI=16となっている。これは、436個のテキストを A1から C1までに分類する際、VperSent の値が最も寄与していることを表している。

決定木の変数重要度から、各 CEFR レベル間で重要となるのは VperSent（１文に含まれる動詞の数の平均）であることが分かる。図３からも明らかなように、この指標は全データの分類に関わっている。１文に含まれる動詞の数が増えるとより文構造が複雑になり、使用される文法項目が増えることを示している。これは A1から B2にかけて隣同士のレベルで分類する際にも特に有効な指標となっており、B2までは文構造の複雑さは有意に上がり続けていることが分かる。しかし、B1、B2と C1を分類する際（図３⑬）では BperA が使用されており、B2までは文構造や出現する文法項目の増加が伺えるが、B2から C1にレベルが上がる際には文構造の複雑さよりも語彙レベルに変化が生じることが分かる。VperSent に対し、ARI（文字数や１文の単語数などから推定されるリーダビ

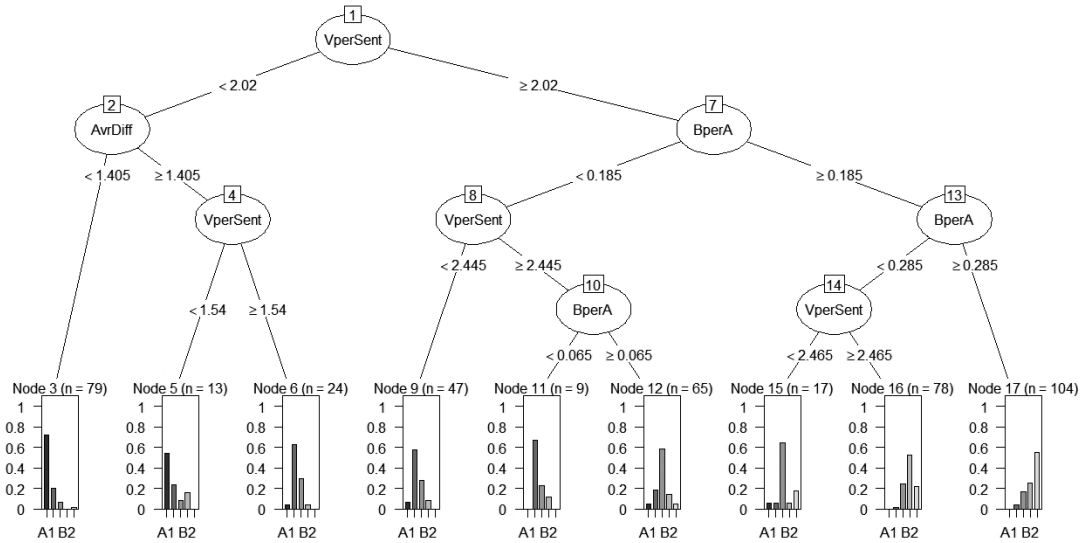

図３　全データのレベル分類における決定木[4]

リティ指標）は決定木の中で一度も出現しておらず、今回の分類モデルでは重要な役割を果たしているとは言い難い。

表６は評価データ（n=188）を上記の決定木を用いて R の predict 関数を使用して CEFR レベルを予測した結果と、教材に付与されている CEFR レベルの関係を表している。この結果から、上記の決定木は全体として約55％の精度にとどまり、これは同じ４指標を用いた CVLA（内田・根岸 , 2021）の精度（約53％）と近いものとなった。レベルごとの正答率を見ると、A1が約73％、A2が約61％、B1が約37％、B2が約38％、C1が約79％という結果になった。

表６　分類モデルによる予測 CEFR レベルと教材の CEFR レベル

		教材の CEFR レベル					
		A1	A2	B1	B2	C1	計
予測 CEFR レベル	A1	22（73％）	11（24％）	1（ 2％）	1（ 3％）	0	35
	A2	8（27％）	28（61％）	10（23％）	3（ 8％）	0	49
	B1	0	5（11％）	16（37％）	8（20％）	1（ 3％）	30
	B2	0	1（ 2％）	10（23％）	15（38％）	5（17％）	31
	C1	0	1（ 2％）	6（14％）	13（33％）	23（79％）	43
	計	30	46	43	40	29	188

表７は CVLA を使用した精度分析（内田・根岸 , 2021）と本研究でのレベル別の精度を比較したものである。本研究と内田・根岸（2021）で使用されているテキストデータの数や１つのテキストの平均語数が異なるため、単に精度を比較することはできないが、両者とも B1、B2 レベルのテキストの正答率が落ち込んでいることが共通している。このことから、A レベルや C レベルと比較して、特に B レベルの判定が CVLA と本研究が使用している４指標では困難であることが分かる。ま

た、表6から明らかなようにBレベルの誤判定は「1つ上」あるいは「1つ下」のレベルとするものが多く、このレベルでは特に隣接レベルとの判別が難しいことが読み取れる。

表7　内田・根岸（2021）と本研究での正答率の比較

CEFR レベル	内田・根岸（2021）	本研究
A1	100%（1/1）	73%（22/30）
A2	70%（7/10）	61%（28/46）
B1	59%（10/17）	37%（16/43）
B2	54%（7/13）	38%（15/40）
C1	67%（14/21）	79%（23/29）
C2	21%（4/19）	—

4．隣接レベル間での分類

前節で作成した分類モデルから、A1とA2、A2とB1、B1とB2、B2とC1の分類において重要になる指標が異なることが分かった。これは、学習者がA1からA2、A2からB1と教材のレベルを上げる際に向上が求められる英語読解能力の種類が異なることを示している。各レベル間でそれぞれどの指標が重要となるかを明らかにするため、隣接したレベル間の分類における4指標の影響力を分析する。

4.1　手法

隣接したレベル間での分類を行う際の各指標の影響力を比較するために、隣接したレベルのデータを取り出し、訓練データ（x）と評価データ（y）が7：3になるように分割する（A1-A2: x = 163, y = 70; A2-B1: x = 200, y = 87; B1-B2: x = 196, y = 85; B2-C1: x = 164, y = 71）。訓練データから決定木を作成し、評価データにて分類モデルの精度を測定する。また、4指標の変数重要度の違いを分析する。

4.2　結果と考察

表8は隣接したレベルのテキストデータ分類の決定木を作成した際の各指標の変数重要度を表している。決定木の精度はA1-A2間では約83%、A2-B1間では約74%、B1-B2間では約64%、B2-C1間では約68%となった。表8通り、各レベル間で分類に有効な指標が異なることから、各レベル間でテキストのレベルが上がる際に学習者にとって必要となる能力が異なることがわかる。精度を比較

表8　各レベル間における各指標の変数重要度

	分類精度	VperSent	AvrDiff	BperA	ARI
A1-A2	83%	39	29	16	15
A2-B1	74%	43	21	20	16
B1-B2	64%	25	30	30	16
B2-C1	68%	15	21	21	43

すると、B1-B2間の分類精度が最も低く、これは3節での分析と同様にB1-B2間のレベルの分類を行う際はこの4指標のみでは十分でないことを示している。それに対し、A1-A2間の分類精度は3節の分析同様に高く、レベルが低いテキストを分類する際にはこの4指標で十分であると言える。

A1とA2の分類にて変数重要度が最も高いものはVperSentであり、続いてAvrDiff、そしてBperAとARIは重要度が低い。VperSentが最も高いということは、A1とA2では文の構造が有意に複雑になっていることを示す。例えば、A1のテキストは“She has a brother and a sister.”（*American Headway Starter*, Oxford University Press）や“In Japan we eat rice with many meals.”（*American Headway Starter*, Oxford University Press）といった1文に動詞が1つの単純な文章が多くを占め、“They are showing old movies all month.”（*Four Corners 1*, Cambridge University Press）といった動詞が2つ（are, showing）の進行形の文章が散見される程度である。しかし、A2になると、“In September it's our family party again and we all plan to meet in Izmir as usual.”（*Empower A2*, Cambridge English）といったように接続詞を使用して1文の中に動詞を多数使用する文が見られる。また、plan to meetやwant to learnなどといった、不定詞を使用した文章が多く見られ、A2で出現する文法項目が増えることが予測できる。反対に、BperAの変数重要度は4レベル間で最も低値を示している。BperAは、Aレベルの内容語に対するBレベルの内容語の割合を示すものなので、A1からA2というAレベル内でのレベル移行においては影響力を持たないと推測することができる。

A2とB1の分類にて変数重要度の順位はA1からA2と同じであり、これはA2からB1に上がる際にはA1からA2に上がる際と同様に文法項目の習得が重要になることを示す。詳しくテキストを見ると、A2レベルでは“The market is a great place to find bargains, and prices are generally low.”（*Four Corners 2*, Cambridge University Press）というように、接続詞や不定詞等を用いているものの、現在完了形や受動態はほとんど見られない。しかし、B1レベルでは“Good logos have been built up so they are recognizable.”（*New Language Leader Intermediate*, Pearson）というように、現在完了形や受動態などといった新出の文法項目も増え、文法的な複雑さが増している。

B1からB2の分類の変数重要度では、AvrDiffとBperAが最も高い値を示している。AvrDiffは単語のCEFRレベル、BperAはAレベルの内容語に対するBレベルの内容語の割合を示す指標であり、両者とも語彙レベルに関する変数である。これは今までは文法項目がレベル移行の際に重要であったことに対し、B1からB2にレベルが上がる際は語彙の知識を増やすことがA1-A2間やA2-B1間と比べると重要となることが分かる。しかし、語彙に関する指標の変数重要度が4指標の中で最も高いとは言え、重要度の値は30と低い。B1-B2の分類精度が64％と低いことと合わせて考えると、この4指標のみでのB1-B2間の分類の限界を示唆していると考えられる。

B2とC1の分類における変数重要度で最高値を示したのはARIである。ARIは、1単語に含まれる文字数と1文に含まれる単語数から計算される数値である。これは文法項目や語彙レベルを考慮した他の3指標とは異なり、テキストの表面的な特徴をとらえたものとなっており、ARIの値が上がるということは、文が長くなることを表す。このARIが変数重要度で最高値を示すということは、B2からC1にテキストレベルが上がる際に、1文が長くなるという表面的な特徴が大きく変化することを示す。実際、C1のテキストの中には“Britain's longest 'clutter-free' street was opened today with the aim of making cars and people co-exist harmoniously - without the need for hectoring signs and protective steel barriers.”（*Empower C1*, Cambridge English）や、“From Rafael Nadal who lines up his water bottles before each match, to Tiger Woods who always wore the colour red for the final round of golf tournaments—many of sport's biggest stars believe in the power of rituals to bring them luck.”（*Empower C1*,

Cambridge English）といった長い文が多々見られ、ダッシュ、コロン、セミコロン等を使用した文も多い。ARI とは反対に、他の指標の変数重要度は落ち込んでいる。これは、動詞を含む文法的な特徴や語彙レベルに関しては B2 と C1 のテキストでは大きな差が見られないことを示す。よって、B2 から C1 にテキストレベルが上がる際には、文法や語彙などの知識量よりも、より長い 1 文の情報を処理する能力が求められていることが分かる。

図 4 は、表 8 における各指標の値をレベルごとの違いが読み取れるようにグラフに表したものである。このグラフから、A1-A2間、A2-B1間では各指標の変数重要度が似た値を示しているが、B1-B2 では前述したとおり BperA の変数重要度は上昇しているものの、全体的に変数重要度は下がり、一方で B2-C1 では今まで最低値を示していた ARI が最高値を示していることが読み取れる。これは、A1-A2間と A2-B1間でのテキストレベルの上昇は同じ因子によって引き起こされていることに対し、B1-B2間、B2-C1間ではそれぞれ全く異なる因子が関係していることを示唆している。このことから、学習者は B1までは新出文法項目の履修という比較的目標がわかりやすい努力を行えばテキストを処理することができるが、B1以降では同じ努力の方向では処理することができるテキストレベルを上げることに困難が生じることは明らかである。B1-B2間では、分類精度が低いこと、また、図 4 からわかる通り 4 指標の中で突出した重要度を持つ変数がないことから、B1-B2間をより正確に区別するためには、新たな指標を増やす必要があることが示唆される。B2-C1間では、ARI の指標が突出していることから、B2の学習者は今までとは異なり、C1のテキストを理解するためには、より多くの文章を読むトレーニングをすること、英語に慣れること、そして長い文の情報処理能力を上げることが求められていると推測される。これらは B2のテキスト理解までには比較的必要とされていない能力であり、B2 の学習者のレベルを C1 に向上するには他のレベルと異なるアプローチが必要となることが示唆される。しかし、B2-C1 間の分類精度が 70％に満たないことから、ARI 指標の他に有効な分類因子を探索する余地がある。

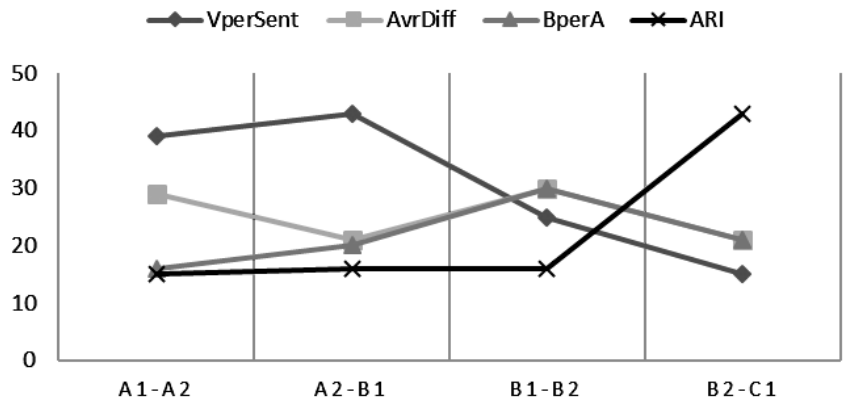

図４　各指標の変数重要度の推移

５．まとめ

本研究は、テキストの難易度が上がる際にどの指標がどの程度有効なのかを各隣接レベル間で明らかにすることを目的として、はじめに CEFR 準拠教科書コーパスの各文章に CVLA を利用して

BperA、VperSent、ARI、AvrDiff の 4 指標の値を文章ごとに算出し、それらの値のレベル別の平均値、増加率等を分析した。その結果、4 指標すべてがレベル間で有意差を示し、テキストレベルを推定する際に有効であることが示された。また、各値の第一四分位数、平均値、第三四分位数を算出したことにより、ターゲットレベルが決まっている新たなテキストを作成する場合に参考とすべき各指標の値が提示できた。その後、CEFR 準拠教科書コーパスに含まれる文章を訓練データと評価データに分け、訓練データを用いて A1 から C1 までのテキストレベルの分類モデルを作成した。その結果、テキストレベルを推定する際に VperSent、BperA、AvrDiff、ARI の順で寄与することが明らかになった。しかし、この分類モデルの精度は全体として約55％に留まり、特に B1や B2のレベルの精度が低かった。総合的に見ると VperSent が最も重要な変数であるものの、A1と A2、A2と B1、B1 と B2、B2 と C1 の分類においてそれぞれ重要になる指標は異なることが示唆された。そこで、各隣接レベル間（A1-A2, A2-B1, B1-B2, B2-C1）を判定する分類モデルを 4 つ作成し、4 指標の変数重要度から、教科書のレベルが上がる際にテキストにどのような変化があるのかを考察した。その結果、A1から A2、A2から B1では文の構造が複雑になっており、文法事項が重要になること、B1から B2では語彙の重要性が増すこと、B2から C1では長い 1 文の情報を処理する能力の重要性が高くなることが分かった。このことは、段階ごとに読解のレベルを上げるために求められる能力が異なること示しており、A1から C1まで同じような学習方法を続けるのではなく、特に B1以降では学習方法を大きく見直す必要があることが示唆される。

本研究の課題として、各レベル間のテキストの違いとして CVLA の 4 指標のみを使用していることが挙げられる。特に B1-B2の分類モデルの精度は約64％、B2-C1間では68％と低く、この 4 指標のみでは分類に限界があることがわかった。よって、今後は CVLA の 4 指標に加え、他の指標も加えて分析を行うことによって、B1以降のレベルの上昇に伴うテキストの分類精度を上げることができると考える。

注

1 『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2021年 9 月ダウンロード）
2 内容がひとまとまりのものを 1 文章と表記する。
3 本節ではこれらの指標が CEFR レベル判別に有効であるかどうかを示すことを目的としているため、ここでは多重比較は実施しないこととする。
4 図の大きさの関係で省略されているが、最下層のレベル判定結果のヒストグラムは左から A1, A2, B1, B2, C1となっている。

参 考 文 献

Coleman, M., & Liau, T. L.（1975）. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60（2）, 283.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C.（2010）. Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes,* 47（4）, 292- 330.

Senter, R. J., & Smith, E. A.（1967）. *Automated readability index.* AMRL-TR-6620. Aerospace Medical

Division, Wright Patterson AFB, Ohio.

Smith, M., & Turner, J. (2016). The Common European Framework of Reference for Languages (CEFR) and The Lexile® Framework for Reading. Retrieved from https://metametricsinc.com/wp-content/uploads/2018/01/CEFR_1.pdf.

Uchida, S., & Negishi, M. (2018). Assigning CEFR-J levels to English texts based on textual features. In *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference*. 463-468.

内田諭・根岸雅史 (2021).「英語読解教材の CEFR レベルの推定 ── CVLA の妥当性評価 ──」『Journal of Corpus-based Lexicology Studies』3, 1-14.

古川昭夫・神田みなみ・黛道子・佐藤まりあ・西澤一・宮下いづみ・畑中貴美 (2010).『英語多読完全ブックガイド』東京 : コスモピア.

水本篤 (2013).「英文解析プログラムから得られる各種指標を使ったテキスト難易度の推定 ── 教材作成への適用可能性 ──」『外国語教育メディア学会 (LET) 関西支部 メソドロジー研究部会報告論集』3, 141-150.

The Importance of Criterial Features for CEFR-based textbooks: A Case Study Using CVLA

Risako AZEMOTO and Satoru UCHIDA

This study aims to reveal the importance of criterial features for the CEFR levels using CVLA, an online application for estimating the CEFR level of a given text. The indicators used in this study are BperA, VperSent, ARI, and AvrDiff, which are employed in CVLA. The data used in this study is a corpus that consists of CEFR-based English textbooks. The values of the four indicators of each text were calculated using CVLA, and then a decision tree model that predicts the CEFR level of each text was created to identify which indicator has the strongest influence on the prediction. Our results show that VperSent marked the highest variable importance, which means that it contributed the most to predicting the CEFR level of a text. However, the prediction accuracy was about 55%, which is not sufficiently reliable. Therefore, additional decision tree models were made for the adjacent levels (A1-A2, A2-B1, B1-B2, and B2-C1) to examine the most important indicator for distinguishing the neighboring levels. Consequently, VperSent was the strongest on A1-A2 and A2-B2 decision trees, BperA and AvrDiff on B1-B2, and ARI on B2-C1. This means that when the text level goes up from A1 to A2 and A2 to B1, the constructions of the sentences become more complex. The results also show that the distinction between B1 and B2 levels relies on the vocabulary level. Lastly, it was shown that the difference between B2 and C1 lies mainly in the length of sentences.