アジア地域における英語学習者の英語使用の特徴

藤野沙也加　久野雅樹
電気通信大学大学院 情報理工学研究科
f2430116@gl.cc.uec.ac.jp　hisano@uec.ac.jp

概要

本研究では，アジア地域の英語学習者が書いたエッセイを対象に，各地域特有の言語使用の特徴を明らかにすることを目的とした．ICNALE コーパスの中級レベル（B1_1 および B1_2）の「大学生のアルバイト」に関するエッセイデータを用い，TF-IDF でベクトル化したデータに基づきロジスティック回帰モデルで地域分類を実行した．モデルの正解率は 0.74 であり，混同行列と使用単語の寄与度の分析から，各地域特有の語彙傾向が確認された．これは，文化的・教育的背景が特徴に影響を与えることを示している．

1　はじめに

学習者コーパスは，言語学習者が産出した言語データを収集した特殊なコーパスであり，応用言語学と言語教育の架け橋として注目されている．学習者が使用する外国語は，母語（第一言語）と目標言語（第二言語）の間に位置する中間言語として特徴づけられることから，学習者コーパスは「中間言語コーパス」とも呼ばれる [1]．これを分析することで，言語習得の過程や文化的背景を明らかにすることが期待される．

特に英語学習者コーパスは，国際的なコミュニケーションで需要が高まる英語に特化し，多様な地域や文化的背景を反映したデータを提供する．これにより，学習者の言語使用の特徴を明らかにし，教育や教材開発に役立つ可能性がある．しかし，従来の研究は特定の学習者グループや教材に偏り，学習者を取り巻く多様な条件や文脈が十分に考慮されていない場合が多い [2]．

本研究では，英語学習者コーパスの一つである ICNALE を用い，広範囲かつ一定の条件で収集されたデータを分析し，地域や文化的背景による言語使用の特徴を包括的に捉え，より実践的な教育への応用を目指す．

2　ICNALE について

ICNALE（International Corpus Network of Asian Learners of English）は，神戸大学の石川慎一郎が開発した国際的な学習者コーパスである [2][3]．ICNALE には，アジア地域の 10 か国/地域の大学生（大学院生含む）および英語母語話者による，統制されたトピックに基づくスピーチやエッセイ 10,000 件以上が収録されている．

学習者は L2 語彙サイズテスト（VST）の受験と TOEFL や TOEIC スコアの提示が義務付けられ，CEFR に基づき 4 つの習熟度バンド（A2, B1_1, B1_2, B2+）に分類された．ICNALE は，Spoken Monologue，Spoken Dialogue，Written Essays，Edited Essays の 4 モジュールで構成され，公に利用可能な最大規模の学習者コーパスの 1 つである．

ICNALE の Written Essays は，「大学生のアルバイト」と「レストランでの禁煙」の 2 つのトピックに基づくエッセイで構成されている．地域別・レベル別のエッセイ数を表 1 に示した．

表 1: 地域別・レベル別の Written Essays データ数

	A2	B1_1	B1_2	B2+	合計
CHN	50	232	105	13	400
HKG	1	30	52	17	100
IDN	32	82	83	3	200
JPN	154	179	49	18	400
KOR	75	61	88	76	300
PAK	18	91	88	3	200
PHL	2	11	176	11	200
SIN			134	66	200
THA	119	179	100	2	400
TWN	29	87	61	23	200
ENS					200
合計	480	952	936	232	2800

学習者は CEFR に基づいて習熟度の低い順に A2，B1_1，B1_2，B2+の 4 段階に分類されている．

3　関連研究

Granger (1998) は，コンピュータ学習者コーパスを用いて学習者の第二言語産出を分析する新しい枠組みとして，対象中間言語分析（Contrastive Interlanguage Analysis, CIA）を提案した [4]．この手法は，ネイティブスピーカーと学習者，あるいは異なる学習者グループ間の言語使用を比較し，過剰使用や過少使用といった学習者特有の特徴を明らかにすることを目的としている．また，CIA を通じて得られた知見は，英語教育における指導法や教材設計の改善に応用可能であり，学習者コーパス研究の中心的な方法論としての地位を確立した．

Ogata ら (2014) は，アジアの学習者，特に日本人，中国人，韓国人，台湾人が基本的な前置詞と-ly 副詞を英語母語話者と比較してどのように使用しているかを調査した [5]．まず AntConc を使用して，ICNALE コーパスデータの頻度分析を実行し，最も頻繁に使用される前置詞と-ly 副詞を特定し，対数尤度比を用いて過剰使用，過少使用の上位 10 語を抜き出した．4 つの地域に共通して“besides”，“completely”が過剰使用され，“over”，“toward”，“upon”，“within”，“simply”が過少使用されていた．また 3 つの地域に共通して，“beside”，“near”，“among”，“during”，“gradually”が過剰使用され，“as”，“on”，“throughout”，“without”，“really”，“fully”，“eventually”，“personally”，“probably”が過少使用されていた．

Ishikawa（2016）は，ICNALE コーパスを用いて that 節を伴う報告動詞の使用に関して検討した [6]．その結果，英語母語話者はアジア地域の学習者に比べて使用頻度が高いことが示された．一方，学習者は"think", "believe", "agree", "know"などの特定の報告動詞を多用することが確認された．

Ishikawa（2023）は，アジア地域の英語学習者の語彙使用の特徴を明らかにするため，ICNALE の Written Essays データから対数尤度比を用いて過剰使用，過少使用しているキーワードを明らかにした [2]．具体的には，学習者は"we", "people"など集団的アイデンティティを強調する表現を多用する一方で，英語母語話者は"I"を用いて個人的な視点から主張する傾向があった．また，学習者は"but"を多用する一方，英語母語話者は"and", "as"を多用し，論理の流れが自然でなめらかに展開されていることが分かった．

4　分析

4.1　目的

本研究の目的は，アジア地域の英語学習者が書いたエッセイを対象に，各地域に特有の言語使用の特徴を明らかにすることである．ICNALE コーパスに収録された「大学生のアルバイト」に関するエッセイを用い，学習者の英語産出における単語の使用傾向を定量的に評価した．さらに，地域ごとの特徴を明らかにするため，機械学習モデルを活用して地域間の相違点を可視化することを目指した．

4.2　方法

ICNALE の Written Essays に収録された，B1_1 および B1_2 レベルのエッセイを対象に分析を行った．B1 とは CEFR の 6 段階のスケールで下から 3 番目に位置する区分で，中級レベルに相当する．このレベルを対象とした理由は，地域間の言語使用の差異を効果的に捉えられる中間的な習熟度であり，さらにデータ数が多く統計的信頼性が高いためである．データの処理には，単語の重要度を反映する TF-IDF（Term Frequency-Inverse Document Frequency）を用いてベクトル化を行い，ロジスティック回帰モデルを適用して学習者のエッセイが属する地域を分類した．

モデルの性能評価には 5 分割交差検証を使用した．この手法では，データセットを 5 つのサブセットに分割し，1 つをテストデータ，残り 4 つを訓練データとして学習を行うプロセスを 5 回繰り返した．各分割で得られた結果を統合し，評価指標を算出するとともに，混同行列を作成して分類結果の詳細を分析した．

評価指標には，正解率，適合率，再現率，F1 スコアを用いた．また，ロジスティック回帰モデルの回帰係数を解析することで，単語の分類への寄与度を明らかにした．

4.3　結果と考察

分析の結果から，分類結果の評価指標（表 2）およびモデルの性能を示す混同行列（図 1）を得た．分類全体の正解率は 0.74 であった．

地域ごとの結果を見ると，HKG（香港），KOR（韓国），および TWN（台湾）がそれぞれ高い適

合率（HKG: 0.95, KOR: 1.00, TWN: 0.95）を示した．一方で，再現率は低い値（HKG: 0.48, KOR: 0.33, TWN: 0.13）になっており，特定の地域において分類の偏りが観察された．

表 2: 地域別分類結果の評価指標

地域	適合率	再現率	F1	データ数
CHN	0.63	0.96	0.76	337
HKG	0.95	0.48	0.63	82
IDN	0.88	0.55	0.68	165
JPN	0.78	0.91	0.84	228
KOR	1.00	0.33	0.49	149
PAK	0.90	0.92	0.91	179
PHL	0.79	0.70	0.74	187
SIN	0.86	0.72	0.78	134
THA	0.65	0.91	0.76	279
TWN	0.95	0.13	0.23	148
ENS	0.70	0.84	0.76	200

混同行列（図 1）からは，モデルが CHN（中国），THA（タイ），JPN（日本）の分類において特に高い正答率を示していることが確認できる．一方で，SIN（シンガポール/マレーシア）や KOR（韓国）では，他地域への誤分類が多く，モデルの性能が低い傾向が見られた．

分類指標におけるデータ数（表 2）は，各地域ごとのサンプル数を反映しており，モデル性能の評価に影響を与えている．特に，サンプル数が多い CHN（337 件）や JPN（228 件）においては比較的高い再現率が得られたが，サンプル数が少ない HKG（82 件）や TWN（148 件）では再現率のばらつきが目立つ結果となった．

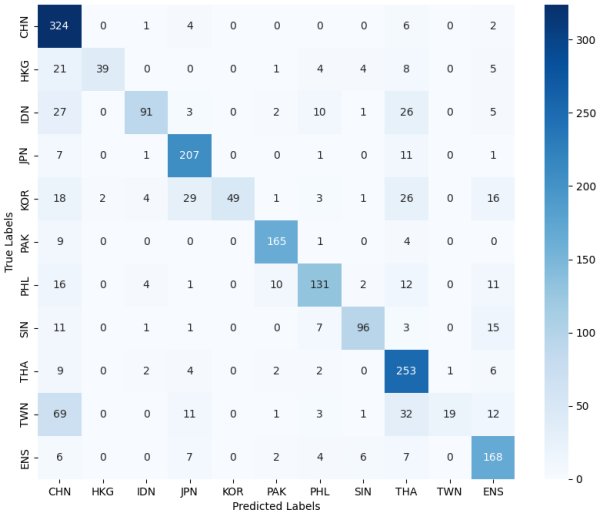

図 1: 交差検証の全 fold を集計した混同行列

全体の正解率は 0.74 と比較的高いものの，一部の地域間で誤分類が見られた．特に注目すべき TWN（台湾）の CHN（中国）への誤分類である．この原因として，両地域が文化的・言語的に近く，類似した語彙や表現を共有しているため，分類が困難になったと考えられる．

KOR（韓国）のエッセイは，適合率が 1.0 と非常に高い一方で，再現率が 0.33 と低い結果を示した．これは，KOR として正しく分類されたエッセイは全て正解しているものの，多くが他の地域（特に CHN，JPN，THA，ENS）に誤分類されたためである．この誤分類の背景には，KOR のデータ数が 149 件と少なく，モデルが韓国特有の特徴を十分に学習できていない可能性が挙げられる．

次に寄与度の分析について結果を示す．ここでは紙面の制約から，日本の近隣地域であり，文化や教育に共通点と相違点が見られるアジア圏（CHN, JPN, KOR）と，比較対象として英語母語話者（ENS）の 4 地域を抜粋し，表 3 に示した．

CHN（中国）では寄与度が高い単語として "society", "us", "we", "our"といった集団的・社会的な視点を表す単語が挙げられた．また"more"も寄与度が高く，"more and more" や "more students" といった形で多用されており，エッセイ内で"more"を用いた強調表現や比較を頻繁に使用する傾向があることが伺える．一方で，"because" や "work" が低寄与語となっており，これらの語が他地域と比較して CHN のエッセイで使用頻度が低い可能性があると考えられる．

JPN（日本）では寄与度が高い単語として"think", "agree", "example" が挙げられた．特に "think" は，日本語の「私は〜と思う」という表現が"I think"に直訳される形で多用され，日本の学習者に特徴的な語として現れている．"agree", "example" は，自分の立場を明確にした後で具体例を展開するという日本の作文指導や議論スタイルを反映していると考えられる．一方で，"the" や "more" といった単語は低寄与語となった．特に "the" は，日本語に冠詞が存在しない影響で，使用頻度が他地域と比べて低い可能性がある．また，"more" は，中国の学習者が強調や比較で多用する一方，日本の学習者では使用頻度が低く，低寄与になると考えられる．

KOR（韓国）では，"tuition" や "money" など，教育費や経済的な背景を反映する単語が高い寄与を示した．一方，"the", "their", "time" など，

他地域でも一般的に使用される単語が低寄与語として挙げられた．全体として，韓国の寄与度は他地域と比較して特定の単語に偏らず，絶対値が小さい傾向が見られた．このため，韓国特有の特徴が分類に十分反映されず，誤分類が多くなった可能性が考えられる．

ENS（英語母語話者）では，"that", "and" などの機能語が高い寄与を示した．これらの単語は，文構造の接続や補足に頻繁に用いられるため，英語母語話者が特に多く使用している可能性が考えられる．一方で，"job", "part" などエッセイのテーマに関連する単語は低寄与語として挙げられた．ENS は，特定の語彙に依存せず多様な語彙を用いて表現しているため寄与度が低くなったと考えられる．

表 3: 日本近隣地域と英語母語話者の分類に寄与した単語

地域	高寄与語	係数	低寄与語	係数
CHN	society	3.23	because	-1.82
	more	2.76	work	-1.62
	us	2.28	money	-1.11
	we	2.15	or	-1.10
	our	2.15	their	-1.06
	taking	2.02	working	-1.06
	can	1.96	would	-1.04
	jobs	1.89	that	-0.99
	take	1.84	they	-0.98
	time	1.72	studies	-0.91
JPN	think	3.01	the	-1.88
	we	2.08	more	-1.70
	college	2.03	in	-1.16
	agree	1.98	their	-1.11
	statement	1.86	find	-0.92
	money	1.85	just	-0.88
	example	1.72	help	-0.87
	with	1.71	some	-0.87
	society	1.69	time	-0.83
	so	1.66	as	-0.82
KOR	tuition	2.27	the	-1.84
	part	1.64	their	-1.44
	money	1.54	with	-1.18
	korea	1.44	them	-1.10
	was	1.29	us	-1.04
	difficult	1.27	for	-1.03
	social	1.25	our	-0.99
	expensive	1.24	may	-0.97
	vacation	1.16	that	-0.96
	economic	1.10	or	-0.92
ENS	that	3.05	job	-2.13
	and	1.64	part	-1.89
	at	1.33	we	-1.68
	would	1.26	time	-1.65
	any	1.15	money	-1.59
	enough	1.08	can	-1.09
	work	1.02	society	-1.06
	graduate	0.98	life	-1.06
	of	0.98	earn	-1.01
	real	0.97	our	-0.99

5 結論

本研究では，ICNALE コーパスを用いて地域分類を行い，分類精度と寄与度を考察した．分類における全体の正解率は 0.74 と高かったが，TWN（台湾）が CHN（中国）に誤分類されるなど，一部地域で誤分類が見られた．これは両地域が文化的・言語的に近いことが影響していると考えられる．また，KOR（韓国）では適合率 1.0 と高い一方で，再現率 0.33 と低く，他地域に誤分類されやすい傾向が確認された．これにはデータ数の少なさや，他地域との特徴の類似性が関与している可能性がある．

また寄与度の分析により，各地域の語彙選択や表現スタイルの違いが明らかになった．CHN は"more"を多用し，数量や比較を重視する表現が特徴的だった．JPN では"think"が日本語の「私は〜と思う」の影響で頻出し，"agree"や"example"は日本の作文指導や議論スタイルを反映して高い寄与を示した．KOR は"tuition"や"money"が寄与したが，寄与度の絶対値が小さく，他地域との類似性が誤分類の要因となった．ENS は"that"や"and"が高い寄与を示し，多様な語彙の活用が特徴的だった．

6 課題

今後の課題として，データ数の偏りを改善するため，特に HKG や TWN などサンプル数が少ない地域のデータへの対応を検討する必要がある．また，もう 1 つのエッセイデータ（「レストランでの禁煙」をテーマとするもの）や他の習熟度レベルのデータ，さらには他のモジュールを追加することで，データの多様性を高め，より包括的な分析が可能になると考えられる．

さらに，モデルの改善には，語彙だけでなく，構文や文法的特徴を新たな特徴量として取り入れることが求められる．構文や文法的特徴は，地域ごとの言語使用における重要な違いを反映するため，これらを適切に活用することで分類精度の向上が期待され，学習者の英語使用の特徴をより精緻に捉えられると考えられる．

参考文献

[1] 石川慎一郎. ベーシックコーパス言語学. ひつじ書房, 2012.

[2] Shin'ichiro Ishikawa. The ICNALE Guide: An introduction to a learner corpus study on Asian learners' L2 English. Routledge, 2023.

[3] 石川慎一郎. ICNALE: The international corpus network of Asian learners of English, (2024-1-7 閲覧). https://language.sakura.ne.jp/icnale/modules.html#0.

[4] Sylviane Granger. The computer learner corpus: A versatile new source of data for SLA research. **Leaner English on computer**, pp. 3–18.

[5] Takashi Ogata and Koichi Kawamura. Asian learners' common overuse/underuse of basic prepositions and -ly adverbs : A study based on the ICNALE. **Learner Corpus Studies in Asia and the World**, No. 2, pp. 349–359, 2014.

[6] Shin'ichiro Ishikawa. Use of that-clauses after reporting verbs in Asian learners' speech and writing: Frequency, verb type, and that-omission. **EPiCSeries in Language and Linguistics**, Vol. 1, pp. 202–215, 2016.

A 付録 (Appendix)

寄与度について，日本近隣地域と英語母語話者以外の地域を表 4 に示す．

表 4: その他の地域の寄与度

地域	高寄与語	係数	低寄与語	係数
HKG	university	5.48	because	-0.49
	students	1.18	many	-0.49
	academic	1.07	work	-0.54
	may	1.04	make	-0.55
	fee	0.91	student	-0.55
	hong	0.80	parents	-0.63
	communication	0.79	we	-0.65
	therefore	0.78	and	-0.76
	kong	0.76	you	-1.22
	time	0.74	college	-2.52
IDN	their	1.97	these	-0.91
	because	1.88	would	-0.95
	can	1.73	think	-0.99
	they	1.53	society	-1.04
	student	1.39	and	-1.07
	must	1.39	may	-1.17
	college	1.37	should	-1.22
	beside	1.36	students	-1.31
	get	1.32	to	-1.41
	teacher	1.28	jobs	-1.42
PAK	their	2.20	university	-1.08
	his	1.87	that	-1.18
	fulfill	1.85	think	-1.19
	jobs	1.74	my	-1.22
	studies	1.65	will	-1.25
	do	1.64	important	-1.26
	country	1.45	you	-1.26
	person	1.37	to	-1.66
	poor	1.34	have	-1.71
	education	1.34	college	-1.74
PHL	having	2.33	get	-0.90
	help	1.82	agree	-0.91
	your	1.80	job	-0.96
	jobs	1.65	important	-0.98
	studies	1.53	more	-0.98
	their	1.51	make	-1.01
	also	1.34	think	-1.13
	really	1.22	many	-1.22
	her	1.18	do	-1.23
	hard	1.18	university	-1.24
SIN	hence	1.76	do	-0.90
	would	1.71	so	-0.94
	up	1.64	but	-0.95
	to	1.50	they	-0.97
	as	1.44	university	-1.06
	be	1.39	have	-1.13
	singapore	1.30	think	-1.19
	able	1.30	can	-1.24
	one	1.15	we	-1.54
	allow	1.14	you	-1.73
	money	2.15	become	-1.06
THA	work	2.02	is	-1.08
	free	1.93	school	-1.12
	do	1.81	our	-1.12
	the	1.68	on	-1.12
	good	1.60	jobs	-1.28
	you	1.47	society	-1.29
	have	1.46	having	-1.32
	etc	1.45	also	-1.52
	use	1.42	of	-1.66
TWN	learn	2.34	help	-0.67
	more	1.70	for	-0.72
	school	1.57	able	-0.72
	how	1.52	do	-0.77
	the	1.25	university	-0.85
	learning	1.24	will	-0.89
	different	1.23	that	-0.91
	experiences	1.12	study	-0.97
	let	1.09	they	-0.98
	my	1.03	of	-1.17