博士論文審査及び最終試験の結果

審査委員（主査）　吉冨　朝子　　印

学位申請者　　フ・シャオリン

論　文　名　　*Criterial Features in L2 English of Japanese Learners Based on Complexity, Accuracy, and Fluency*

【審査結果】

吉冨朝子を主査とし，主任指導教員の投野由紀夫，および副査として根岸雅史，アリアン・ボルロンガン，九州大学内田諭（外部委員）から成る審査委員会は，2023 年 12 月 23 日に上記論文の審査ならびに口述による最終試験を行った。その結果，審査委員会は全員一致で，申請者に対し博士（学術）の学位を授与するのが適当であるとの結論に達した。

【論文の概要】

本研究の目的は英語学習者のスピーキング能力の発達レベルを示す言語特徴に関して，複雑性（complexity），正確性（accuracy），流暢性（fluency）の尺度（いわゆる CAF measure）の観点から明らかにしようとしたものである。外国語運用能力の国際標準というべき指標として利用が促進されつつあるヨーロッパ言語共通参照枠（CEFR）を用い，その CEFR レベル（A1 から C2 など）の特徴を示す，いわゆる「基準特性（criterial feature）」に関する研究ということもできる。「基準特性」に関しては，英語のプロファイル資料が次々と公開されつつある（例：English Profile, Core Inventory for General English, Global Scale of English, など）。しかし，これらのプロジェクトが依拠するコーパス資源はほとんどが母語話者コーパスと学習者の作文コーパスに限定されており，学習者のスピーキング力の記述にそのまま当てはめることが難しい部分が多い。

さらに，産出能力の分析には 20 年程前からいわゆる CAF measure と呼ばれる複雑性・正確性・流暢性の尺度がさまざまに提案・研究されてきており，習熟度レベルとの関連も研究の蓄積があるが，これら CAF measure の研究の大部分も第二言語ライティング研究においてがその中心で，スピーキングに当てはめた研究は全般に立ち後れている。

このギャップに対処するため，本研究では，英語学習者のスピーキング力を CEFR レベル基準特性となりえる CAF measure を検討することで明らかにした。さらに，有効と判明した CAF measure を用いてスピーキング力を判定する CEFR レベル予測モデルを構築し，実証的にその効果を日本人英語学習者の口頭対話（dialogue）タスクと単話（monologue）

タスクにおける予測精度で検証を試みた。

論文は，全 7 章，付録を含めて 317 ページから成る。第 1 章で論文全体の背景と目的を述べ，第 2 章で，CEFR の基礎概念，スピーキング能力の記述と基準特性，CAF measure の基礎概念の整理と主要な先行研究について整理し，CEFR およびスピーキングの能力記述における CAF measure の有効性を訴えた上で，研究設問を提示している。第 3 章では研究手法の説明，特にコーパスと CAF measure の選定・抽出の具体を提示している。第 4 章では CAF measure の CEFR レベルごとの記述統計，第 5 章では CEFR 予測モデルの候補変数の選定過程およびモデルの構築と評価，第 6 章では全体の研究結果の考察を行い，第 7 章で総括と今後の展望を述べている。以下に第 3〜5 章を中心に具体的な論文の概要をまとめる。

本研究では，国立研究開発法人情報通信研究機構（NICT）が開発した話し言葉学習者英語コーパス（NICT JLE Corpus）を主要なデータベースとして採用した。このコーパスから，エラータグが付与され，かつ CEFR レベルの A1（n=50），A2（n=50），B1(n=50)および母語話者（NS=Native Speaker ; n=20）の合計 170 の対話文トランスクリプト（各 15 分，42.5 時間分）を分析に利用した。まず先行研究を精査し，主要な CAF measure のうち，複雑性指標 12 種，正確性指標 12 種，流暢性指標 12 種を選定した。選定の基準は発話単位を 100 語，AS-unit，または節（clause）で定義でき，サンプルのテキスト長に影響を受けにくいもの，汎用の分析ツール等で計算できるもの，人手によるアノテーションが必要だが指標として重要だと思われるもの，等である。また正確性指標に関しては NICT JLE Corpus に付与された 47 のエラーカテゴリーを使用し，誤りのない節の比率（error-free clause ratio）および重み付き節の比率（Weighted Clause Ratio: WCR）を用いた。最後に流暢性指標に関しては，NICT JLE Corpus にタグが付与されている，フィラー，繰り返し，自己訂正，ポーズなどの非流暢性を流暢性が増すと同時に，減少する特徴量として 100 語単位の回数で算出した。

これら 36 種の CAF measure に関して，ファイル単位の特徴量の算出後，CEFR レベル別の各特徴量の平均値を求め，A1〜B1 のレベル単位の変化の傾向，および母語話者との比較を可視化して全体的傾向を把握，その後，Kruskal-Wallis 検定および Dunn の多重比較により，CEFR レベル間（CEFR の 3 レベルおよび母語話者）の有意差を個々の CAF measure ごとに検定した。その結果，36 種のうち，27 の特徴量を予測モデルに入れる候補として選んだ。さらに，これら 27 の特徴量をもとに，CEFR レベル推定に貢献する変数を探るためステップワイズ判別分析を行い，27 から 9 つの特徴量に予測変数の絞り込みを行った。最終的に多重共線性の認められる 2 変量を削除した有効な CAF measure として CEFR 予測モデルに組み込まれたのは，3 つの複雑性尺度（文字/単語，節/AS-unit，節の平均長），1 つの正確性尺度（エラーのない節の比率），3 つの非流暢性尺度（繰り返し/100 単語，短いポーズ/100 単語，終節の長いポーズ/100 単語）を含む 7 つの特徴で，交差検証を行ったケ

ース（A1 から B1 まで）の 86.7%が正しく分類され，モデル構築に有用であることがわかった。

最後にこのモデルを元に類似のスピーキング・データを判定できるかを検証した。このために NICT JLE Corpus の元資料になっている Standard Speaking Test (SST) の簡易版である Telephone Standard Speaking Test (TSST)によって得られた発話データをコーパス化したものから 45 の CEFR レベル付きトランスクリプトを選定し，それらに必要な 7 種の特徴量のアノテーションを行い，レベル予測を先の判別分析のモデルを用いて行った。その結果，ほとんどの CAF 特徴が CEFR レベル間で有意な差を示し，熟達度の高い話者ほど優れたパフォーマンスを示すことがわかった。予測モデルの頑健性は TSST コーパスを用いて検証され，73.3%の第二言語学習者を適切な習熟度グループに分類することに成功した。一方，TSST は電話による単話（モノローグ）形式テストであるので，対面インタビュー形式の NICT JLE コーパスに比べて，低熟達度話者の流暢さが目立って高く，それが CEFR レベルの過大評価を引き起こした可能性が否定できない。このことは，今後モデルの調整が必要であることを示唆している。

本研究により，CEFR レベルの異なる第二言語学習者間でスピーキングにおける CAF measure が示す特徴に明確な違いがあることが確認され，いくつかの CAF の特徴量が基準特性として有効であることが確認された。また，特徴量を厳選した CEFR 予測モデルの頑健性が裏付けられ，CAF measure を用いた適用範囲が書き言葉だけでなく，話し言葉それも対話音声と単話音声の両方に拡大された。

第 6 章では，特に理論的示唆として，英語学習者のスピーキング能力に CAF measure を利用する有効性と可能性を述べ，CEFR レベルの個別言語のより精密な記述と理論構築に貢献すると論じ，方法論的示唆としては音声認識の実用化が現実的なものとなっており，スピーキング能力の評価に本論文で注目した客観的な特徴量を用いる有効性を論じ，かつ教育的示唆としては，カリキュラム開発や評価における潜在的な実用性，AI などの活用による評価システムの開発について論じている。また本研究の残された課題として，今回未調査の CAF measure がまだ一定数あること，2 つのコーパス間のタスクやデータ採取方法の不一致による影響，CAF measure の人手によるアノテーションの負担とそれによる比較的小規模サンプルの限界，なども示している。

最後に第 7 章では，本研究の意義を再度総括しつつ，CEFR 尺度に関する貴重な洞察に貢献し，異なる習熟度レベルにおける第二言語学習者のスピーキング能力の発達と評価に関する理解を深め，具体的な特徴量を用いた今後の評価測定の方向性を示したものであると論文の価値を述べて締めくくっている。

【審査の概要及び評価】

審査では，フ・シャオリン氏による博士論文の概要の発表の後，各審査委員から本論文を評価できる点として，以下のことが指摘された。

1. 第二言語ライティングの分野で研究が発展してきた CAF measure と CEFR の基準特性の研究を融合し,未開拓のスピーキング能力記述の分野に取り組んだ意欲的な研究である。
2. 博士論文準備段階に Hu (2020)で Lexical diversity measure の評価を，また Hu (2021)で complexity measure のみを扱った論文を国際学術誌(Asia Pacific Journal of Corpus Linguistics)に発表しており,発話データでどの CAF measure の特徴量を取り出すべきかに関して地道に予備調査を重ねて今回の論文にまとめあげている。
3. スピーキング・データのアノテーションを粘り強く行い,通常の自動分析では十分にわからない AS-unit などの発話単位を用いた各種エラー並びにフィラー等の詳細なレベル別記述・評価を行っており，今後の AI などでの自動化処理にも役立つ基礎データを提供している。
4. 大規模言語モデルなどは判定プロセスがブラックボックス化する傾向が強いが,今回の提案は説明可能な特徴量を重視したモデルを提案しており,教育・学習上のフィードバックや学習方略を立てやすいという利点がある。
5. 単に CAF measure の傾向を記述的に調査しただけでなく,CEFR レベル判定の予測モデルを構築しており，その中で CAF measure の重み付けなどのデータと考察を提供しており，最終モデルにおける clause per AS-unit の説明力が極めて高い点など，新しい興味深い知見が得られている。

このように，フ・シャオリン氏の博士論文が高く評価できるものであると確認された上で，質疑応答では以下のような指摘と，フ氏からの回答があった。

1. 対話データである NICT JLE Corpus で作った予測モデルを単話データである TSST Corpus に当てはめているが，対話と単話の話し言葉の特徴はかなり異なりうることを踏まえると,モデルを別々に作るべきだったのではないか,という指摘に対し,フ氏は,その可能性は否定できないが，TSST は SST をもとに作られているので，比較可能性が高いデータだと判断すると回答した。
2. 予測モデルを自動で作る場合，音声認識の精度が影響しないかという指摘に対しては,現状では難しいと思うが,将来的には音声認識の精度が上がることが期待されると回答した。
3. High-stakes な検定試験の評価モデルを今回のような CAF features で取って代わるべき

と考えるか，という質問に対しては，人間と自動分析の利点を考慮して，ある程度併用するべきである，との考えを示した。

4. 今回の研究成果が，English Profile Programme (EPP)の基準特性の研究に貢献できる点はあるか，という質問に対しては，EPP が具体的な文法と語彙に関する情報であるのに対して，今回の指標は特定の語彙や文法に寄らないものが多い点が異なるので，難しいとの見解であった。
5. 自然言語処理ベースの分類手法(BERT, LLM など)を採用しなかった理由については，最新の手法についてはあまり精通していなかったことを認めた。
6. 予測モデルの精度が低かった場合，具体的にどのような発話が予測通りに判定ができなかったのか具体例を教えて欲しかった。また，判別分析以外に他の手法(Random Forest など)を使わずに，判別分析を選んだ理由が明らかではない，という指摘がなされた。
7. 流暢さの測定法として最も一般的な words per minutes (WPM)を選ばなかった理由を問われた際には，WPM は CEFR レベルの判定に役立つが，NICT JLE では発話の時間が明確に計算できないので，WPM が計算できないと説明した。
8. この論文のタイトルには L2 English とあるが，今回の対象は L2 learner English の中でも，日本語を第一言語とする学習者の英語，すなわち拡張円の英語変種のみを対象としている。L2 English では世界英語の変種すべてを指しうるので，タイトルが曖昧ではないか，という指摘がなされた。フ氏はこの点を認め，将来的にデータが整えば日本語以外の第一言語の学習者英語についても調査してみたいと述べ，博士論文の最終稿ではタイトルを英語学習者の第一言語を特定する形で修正するとした。
9. 今回の予想モデルは比較的 high-stakes なテストをもとにした分析だが，これが実際の授業の中で具体的にどのように活用されるのか，特に文法指導に関して何か提案はあるか，という質問に対しては，将来的に自動分析が可能になると，教室内の生徒の CEFR レベルを判定でき，それに応じたタスクやサポートができる可能性について答えた。
10. CEFR や量的な分析の限界や課題は何だと思うか，という質問に対しては，社会言語学や語用論的側面等，人間でないと評価できない点が明確にあるはずなので，将来的に機械ができることも増えていくとは思うが，機械と人間の役割をより明確にして全体的な評価方法を考える必要があると回答した。
11. 教育的示唆として，具体的に CEFR レベル別に判明した特徴（例：A1-A2 は流暢性が，A2-B1 だと正確性が重要）をどのように指導に活かすのか，より具体的な指導上の工夫の提示，および B2 レベル以上に関する議論もあると良かった，との指摘があった。
12. 36 の CAF measure のうち，どの部分が自動でどこが手動だったのか。また，9 つの測定法の中で Weighted Clause Ratio (WCR)を落とした理由は何かと問われると，AS-unit および clause の認定は手動で行ったこと，また，WCR と error-free clause ratio は異なるものを測っているが，相関が非常に高かったので落としたと回答した。これに対して，

統計的に選定するだけではなく,特徴量の持つ情報に教育的価値があるかについても考慮すべきだとの指摘がなされた。

このように，フ・シャオリン氏は審査委員からの質問や指摘に対して的確に応答するとともに，自らの研究の限界を認識しその改善法についても十分に理解していることが窺える返答をした。

最終試験後の審査委員会での審議においては，委員からの指摘の多くが，フ・シャオリン氏の論文から導かれる今後の研究の発展の可能性について指摘したもので，本調査がもつ学術的意義を高く評価した上での期待の表れとみなすべきものであったことが確認された。

以上の論文評価および最終試験での質疑応答の内容から，本論文は英語教育学および第二言語習得研究の話し言葉の習得や指導と評価に大いに貢献する秀逸なコーパス研究であり，学位申請者が優れた研究者・教育者としての資質を十分に有していると判断された。よって審査委員会は，全員一致で，学位申請者が博士（学術）の学位を授与するにふさわしいとの結論に達した。