

## 科学研究費助成事業（科学研究費補助金）研究成果報告書

平成25年 6月 6日現在

機関番号：34418

研究種目：基盤研究（B）

研究期間：2010～2012

課題番号：22300299

研究課題名（和文） 第二言語学習者コーパスを利用した  
読解能力・作文能力の自動評価法に関する研究

研究課題名（英文） Learner Corpus-based Research on Automatic Evaluation Methods for  
Reading Ability and Writing Ability

研究代表者

小谷 克則 (KOTANI KATSUNORI)

関西外国語大学・外国語学部・准教授

研究者番号：30440994

研究成果の概要（和文）：

本研究では、英語学習者の作文、発音、読解、聴解において生じた言語データを収集した第二言語学習者コーパスを構築した。また、その第二言語学習者コーパスに基づき、英語学習者の作文、発音、読解、聴解における各言語技能に対する自動評価法を開発した。さらに、ある英文の各技能における難易度に対する自動評価法も開発した。これらの自動評価法を検証した結果、複数技能を対象とする第二言語学習者コーパスとその自動評価法の有効性が示された。

研究成果の概要（英文）：

The present study constructed a learner corpus by collecting linguistic data generated in learners' writing, pronunciation, reading comprehension, and listening comprehension. This learner corpus was subsequently used for developing automatic evaluation methods for learners' linguistic skills of writing, pronunciation, reading comprehension and listening comprehension. In addition, automatic evaluation methods were developed to measure the difficulty of sentences for each linguistic skill. Experimental results for the validation of these automatic evaluation methods showed the efficacy of our learner corpus and automatic evaluation methods for multiple linguistic skills.

交付決定額

(金額単位：円)

	直接経費	間接経費	合計
2010年度	6,700,000	2,010,000	8,710,000
2011年度	2,900,000	870,000	3,770,000
2012年度	2,700,000	810,000	3,510,000
年度			
年度			
総計	12,300,000	3,690,000	15,990,000

研究分野：総合領域

科研費の分科・細目：教育工学

キーワード：学習者コーパス、自動評価法、作文、発音、読解、聴解

## 1. 研究開始当初の背景

第二言語研究においては、情報処理技術などの向上に伴い、学習者の会話や作文をデータベース化したコーパス（以下、学習者コー

パス）が構築され、学習者の言語使用を客観的に分析することが可能となった（和泉他 2004、Wen et al. 2008、Granger et al. 2009、安田他 2009）。既存の学習者コーパスでは会

話や作文において学習者が生成した文や音声を収集してきた。これらの学習者コーパスを用いることにより、学習者の作文や会話において頻出する誤用を始め、様々な言語的特徴が明らかにされてきた。また、学習者コーパスのデータ量が学習者特有の言語的特徴を統計的に検出可能な量である場合、学習者コーパスを用いてある言語的特徴を自動的に抽出する手法の開発が可能となる。さらに、言語的特徴として誤用のように、学習者の技能を直接的に示す指標を用いることで、学習者の技能を自動的に評価する手法（自動評価法）の開発が可能となる。

研究開始当時は、作文や会話を対象とする学習者コーパス研究が多数ある一方、読解や聴解を対象とする学習者コーパス研究は皆無に等しく、読解を対象とする学習者コーパス研究として本研究の先行研究がある程度であった (Kotani et al. 2005)。その後、Meurers et al. (2010) の読解を扱う研究が発表されたが、ドイツ語学習者コーパスであり、英語学習者コーパスにおける状況は変わりないといえる。

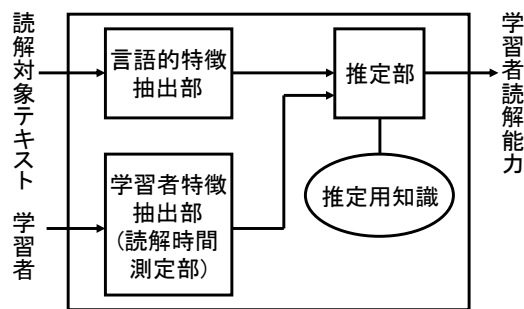


図1：読解自動評価法

本研究の先行研究において構築された読解を対象とする学習者コーパス研究の概要は以下の通りであった。まず、学習者の読解をテキスト単位よりも細かい文単位で観察することを可能とするために文単位で読解時間を計測する読解過程記録システムを構築した (小谷他 2004)。このシステムを用いて、英語学習者(102名)が英文テキスト(66テキスト)の読解に要した文読解時間を計測し、各文の言語的特徴を自然言語処理技術により抽出した。そして、これらの文読解時間と言語的特徴を各文に注記することで読解を対象とする学習者コーパスを構築した。この学習者コーパスを用いて読解自動評価法(図1)を開発した。この読解自動評価法において、学習者の読解能力を推定するための知識(推定用知識)は、学習者読解コーパスに基づいて構築されたものである。読解自動評価法は、学習者が対象テキストを読むのに要する読解時間(学習者特徴)を測定すると同時に、対象テキストの読解における言語的難し

さ(言語的特徴)を自然言語処理技術で抽出する。そして、学習者がどの程度の難しさのテキストをどの程度の時間で読んだかというデータと推定用知識を照合して学習者の読解能力を評価する。

## 2. 研究の目的

上記の先行研究において課題として明らかになったことは、学習者コーパスの対象技能の拡張や自動評価法の更なる有効性の向上であった。

本研究の主たる目的は、既存の読解対象とする学習者コーパスにおいて収集されたデータを検証し改良を行うこと、改良された学習者コーパスを用いて自動評価法を構築することであった。

具体的な目的として、次の三点が挙げられる。

- (1) 学習者コーパスの対象技能の拡張
- (2) 収集するデータの改良
- (3) 自動評価法の改良

まず、学習者コーパスとして収集すべき技能の拡張としては、既存の学習者コーパスでは対象とされてこなかった聴解を対象とすることに加え、作文と読解のように複数の言語技能を対象とする必要性がある。対象となる言語技能は、作文、発音、読解、聴解の四技能である。

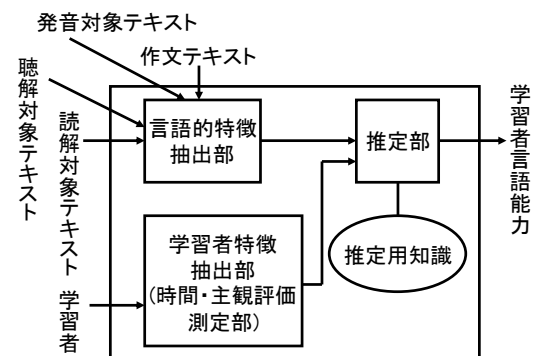


図2：言語技能自動評価法

次に、学習者コーパスとして収集すべきデータを検証すると、既存の読解を対象とする学習者コーパスでは文単位で収集されるデータは読解時間であったが、聴解を観察するデータとして聴解時間を利用することが不可能であるため、時間以外の要素を収集することが改良点として挙げられる。また、複数の領域における言語技能を対象とすることから、収集するデータは聴解においてのみ収集可能ではなく、汎用的に収集可能であることが要件となる。各技能において収集されるデータは言語使用を通じて学習者が生成する文(作文)、音声(発音)、理解度(読解、

聴解)といった結果データに加え、これらの言語使用の過程を示す速度(作文、発音、読解)と主観評価(作文、発音、聴解、読解)といった過程データも加えられる。

そして、新たに構築された学習者コーパスに基づいて、複数の言語技能の観点から学習者の言語技能を評価する自動評価法を開発する必要がある(図2)。また、自動評価法として、言語技能以外にある文の難易度を多角的に評価する自動評価法を開発する必要性もある。

### 3. 研究の方法

本研究は、まず学習者コーパスの対象技能の拡張を検討し、収集するデータの改良を図り、学習者コーパスを構築した。次に、構築された学習者コーパスに基づき自動評価法を開発した。

本研究の学習者コーパスの主たる目的は、学習者の複数の技能を対象としたデータを収集することで、ある学習者の言語使用を形態論・統語論・音韻論・音声学・意味論・語用論といった一般的な言語学の分析を用いて多角的に分析することにあつた。そこで、これらの分析が可能となるような次の5種類の言語使用を課題として設定した。

- (1) 簡単な質問に対する回答としての作文
- (2) 一連のイベントの描写における作文
- (3) 新聞記事の黙読による読解
- (4) 新聞記事の聴き取りによる聴解
- (5) 新聞記事の音読による発音

各課題における言語使用は、作文、読解、聴解、発音の技能を対象とし、以下の分析が可能となる。(1)において学習者が生成した作文を対象に、語彙規則や文法規則に従って適切に文を生成出来ているかと言った形態論・統語論的分析が可能となる。

(2)において生成された作文を対象とする場合、形態論・統語論的分析はもちろん、代名詞や文脈に応じて適切に文を生成出来ているかと言った意味論・語用論的分析も可能となる。

(3)において学習者が読んだ文とその理解を対象に、学習者が読解対象の文の品詞や統語構造を適切に解析できているかと言った形態論・統語論的分析に加え、文脈や状況も踏まえて適切に理解できているかと言った意味論・語用論的分析も可能となる。

(4)においては、学習者が聞いた文とその理解を対象に、読解と同様の形態論・統語論的分析と意味論・語用論的分析に加え、音素や音韻パターンを適切に聴き取れているかと言った音韻論・音声学的分析も可能となる。

(5)においては、学習者が発した文とその音声を対象に、学習者が適切に音素や音韻パ

ターンを再生できるかと言った音韻論・音声学的分析が可能となる。本研究では、音韻論・音声学に特化するため、音読対象の英文は(3)の課題として事前に読解を行った文を用いた。

本研究はこの学習者コーパスに基づき、学習者の技能、及び英文の難易度を対象として自動評価法を開発した。自動評価法の開発において、学習者コーパスの言語データ、その言語データの分析結果、学習者の特徴を素性として重回帰分析や機械学習法といった統計的手法を用いた。評価対象は、技能の場合は学習者であり、難易度の場合は文とテキストの二種類であつた。

### 4. 研究成果

3節で述べた課題を大学生以上の日本語を母語とする英語学習者90人に課し、データ収集を行った。学習者の平均年齢は21.5歳(標準偏差:2.6, 19-40歳)であり、平均学習歴は123.2ヶ月(標準偏差:36.8, 53-243ヶ月)であつた。学習者は一般的な英語テストとして知られるTOEIC(Test of English for International Communications)のスコアに応じて初級(280-495点)・中級(500-735点)・上級(740-985点)に三区別された。学習者の平均TOEICスコアは633.8点(標準偏差:198.3, 280-985点)であつた。また、参照データとして英語母語話者30名からもデータを収集した。今後の課題として、学習者や母語話者の規模の拡充が挙げられる。

表1に各課題(1:簡単な質問に対する回答としての作文、2:一連のイベントの描写における作文、3:読解、4:聴解、5:発音)において得られたデータの文数と単語数、及びデータの要素が示されている。この学習者コーパスは言語資源協会を通じて配布を行う予定である。

	1	2	3	4	5
文	2,166	1,841	80	80	80
単語	14,562	14,663	1,389	1,407	1,389
時間	✓	✓	✓		✓
主観評価	✓	✓	✓	✓	✓
理解度			✓	✓	

表1:学習者コーパスデータ

学習者に課題として与えた英文データと学習者が生成した作文データに対して形態論・統語論的分析を行い、品詞情報や文法情報を追記した。また、作文データに対しては語彙や文法の誤りを示す形態論・統語論的分析と状況や文脈との齟齬を示す意味論・語用論的分析を行った。今後の課題として、発音データに対する音韻論・音声学的分析が挙げ

られる。

本研究により開発された主たる自動評価法は以下の四点である。まず、TOEIC スコアによる自動評価法。この評価法は学習者コーパスにおける全てのデータを基に学習者の総合的な英語運用能力を推定できる。

次に、各学習者の作文評価スコアによる自動評価法。この評価法は作文課題におけるデータに基づき学習者の作文能力を推定できる。

そして、各テキスト、あるいは各文の読解難易度、いわゆるリーダビリティの自動評価法。この評価法は読解課題におけるデータに基づき学習者にとっての読解難易度を推定できる。

最後に、各テキスト、あるいは各文の聴解難易度を対象とする自動評価法。この評価法は聴解課題におけるデータに基づき学習者にとっての聴解難易度を推定できる。

#### [参考文献]

- Granger, S., Dagneaux, E., Meunier, F., and Paquot, M. 2009. International Corpus of Learner English, Version 2. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.
  - 和泉絵美, 内元清貴, 井佐原均(編) 2004. 日本人 1200 人の英語スピーキングコーパス. 東京: アルク.
  - 小谷克則, 吉見毅彦, 九津見毅, 佐田いち子, 井佐原均 2004 読解過程検定ツール OPERATION2 の評価, 情報科学技術レターズ, Vol. 4, pp. 251--254.
  - Kotani, K., Yoshimi, T., Kutsumi, T., Sata, I. and Isahara, H. 2005. Toward a Unified Evaluation Method for Multiple Reading Support Systems: A Reading Speed-based Procedure, Proceedings of Second International Joint Conference on Natural Language Processing (IJCNLP-05), pp. 244--249.
  - Meurers, D., Ott, N., and Ziai, R. 2010. Compiling a task-based corpus for the analysis of learner language in context. In Featherston, S., and Stolterfoht, B. (Eds.) Proceedings of Linguistic Evidence 2010, pp. 214--217.
  - Wen, Q., Liang, M., and Yan, X. 2008. Spoken and Written Corpus of Chinese Learners (SWECCCL) 2.0. Beijing, China: Foreign Language Teaching and Research Press.
  - 安田圭志, 喜多村圭祐, 山本誠一, 柳田益造 2009. 多重タグ付き英語学習者コーパスの開発と英語能力自動測定への応用, 自然言語処理 16(4), pp. 47--63.
5. 主な発表論文等  
(研究代表者、研究分担者及び連携研究者には下線)
- [雑誌論文] (計 16 件)
- ① Kotani, K., Yoshimi, T., Nanjo, H. and Isahara, H.: Corpus Materials for Constructing Learner Corpus Compiling Speaking, Writing, Listening, and Reading Data, International Journal of Computational Linguistics and Applications, 2013, 査読有 (ACCEPTED).
  - ② Kotani, K., Yoshimi, T. and Isahara, H.: Application of Reading Data in an Integrated Learner Corpus, Procedia Social and Behavioral Sciences, 2013, 査読有 (ACCEPTED).
  - ③ Kotani, K., Yoshimi, T. and Uchida, M.: Automatic Classification of Texts Written by Learners of English as a Foreign Language based on Linguistic Features and Learner Features, Proceedings of International Technology, Education and Development Conference (INTED), pp. 6305--6314, 2013, 査読有.
  - ④ 小谷克則, 吉見毅彦, 小谷志穂, 井佐原均: 統合的日本手話学習者コーパスの開発構想, 手話学研究, Vol. 21, pp. 25--43, 2012, 査読有.
  - ⑤ Yoshimi, T., Kotani, K. and Isahara, H.: Use of a New Set of Linguistic Features to Improve Automatic Assessment of Text Readability, US-China Education Review B, Vol. 2, No. 1, pp. 55--62, 2012, 査読有.
  - ⑥ Kotani, K., Yoshimi, T., Nanjo, H. and Isahara, H.: Applicability of Readability Formulae to the Measurement of Sentence-level Readability, Proceedings of International Conference of Education, Research and Innovation, pp. 6023--6031, 2012, 査読有.
  - ⑦ Kotani, K., Yoshimi, T., Nanjo, H. and Isahara, H.: Statistical Analysis of a Learner Corpus Integrating Linguistic Output and Language Processing Data in Speaking, Listening, Writing, and Reading, G. R. S. Weir and S. Ishikawa (Eds.), Corpora and Language Technologies in Teaching, Learning and Research, University of Strathclyde Press, pp. 21--28, 2011, 査読有.
  - ⑧ 小谷克則, 吉見毅彦, 井佐原均: 読解効率に基づく英文読解能力測定テストの開発とその信頼性・妥当性の検証, 情報処理

- 学会論文誌, Vol.52, No.4, pp.1843--1851, 2011, 査読有.
- ⑨ Kotani, K. and Yoshimi, T.: A Scoring Method for Second Language Writing Based on Word Alignment, *Procedia-Social and Behavioral Sciences*, No.27, pp.42--49, 2011, 査読有.
- ⑩ Kotani, K., Yoshimi, T. and Isahara, H.: A Machine Learning Approach to Measurement of Text Readability for EFL Learners Using Various Linguistic Features, *US-China Education Review B*, Vol.1, No.6, pp.767--777, 2011, 査読有.
- ⑪ Kotani, K., Yoshimi, T., Nanjo, H. and Isahara, H.: Compiling Learner Corpus Data of Linguistic Output and Language Processing in Speaking, Listening, Writing, and Reading, *Proceedings of The 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pp.1418--1422, 2011, 査読有.
- ⑫ Kotani, K. and Yoshimi, T.: Second Language Writing Classification System Based on Word Alignment Distribution, *Themes in Science and Technology Education: Special Issue on ICT in Language Learning*, Vol.3, No.1&2, pp.223--238, 2010, 査読有.
- ⑬ Kotani, K., and Yoshimi, T.: A Machine Learning-Based Evaluation Method for Machine Translation, *Lecture Notes in Computer Science*, Vol.6040, Springer, Berlin, pp.351--356, 2010, 査読有.
- ⑭ Kotani, K., Yoshimi, T. and Isahara, H.: Validation of a Reading Speed Test for Japanese Learners of English as a Foreign Language, *G. R. S. Weir and S. Ishikawa (Eds.), Corpus, ICT, and Language Education*, University of Strathclyde Press, pp.215--224, 2010, 査読有.
- ⑮ Kotani, K., Yoshimi, T. and Isahara, H.: A Prediction Model of Foreign Language Reading Proficiency based on Reading Time and Text Complexity, *US-China Education Review*, Vol.17, No.10, pp.1--9, 2010, 査読有.
- ⑯ Kotani, K., Yoshimi, T. and Isahara, H.: Classification of Language Learners' Sentences into Native Speaker-like or Non-native speaker-like Sentences Using Learner Sentences and Machine Translation Sentences as Training Data, *Proceedings of International Conference of Education, Research and Innovation*, pp.2672--2677, 2010, 査読有.
- [学会発表] (計14件)
- ① Kotani, K., Yoshimi, T., Nanjo, H., and Isahara, H.: An Application of Reading Part of the Integrated Learner Corpus, 5th International Conference on Corpus Linguistics, Mar.14-16, 2013, Alicante, Spain.
- ② 上田翔太, 南條浩輝, 吉見毅彦, 小谷克則: 英語学習者の習熟度を考慮した文単位の聴解難易度予測式の構築, 言語処理学会第19回年次大会, 2013年3月12-15日, 名古屋大学.
- ③ Kotani, K., Yoshimi, T. and Uchida, M.: Automatic Classification of Texts Written by Learners of English as a Foreign Language based on Linguistic Features and Learner Features, *International Technology, Education and Development Conference (INTED)*, Mar.4-6, 2013, Valencia, Spain.
- ④ Kotani, K., Yoshimi, T., Nanjo, H. and Isahara, H.: Applicability of Readability Formulae to the Measurement of Sentence-level Readability, *International Conference of Education, Research and Innovation (ICERI)*, Nov.19-21, 2012, Madrid, Spain.
- ⑤ 奥村将成, 吉見毅彦, 南條浩輝, 小谷克則: 英文読解能力測定モデルへの熟語難易度の導入, FIT2012 第11回情報科学技術フォーラム, 2012年9月4-6日, 法政大学.
- ⑥ 上田翔太, 南條浩輝, 吉見毅彦, 小谷克則: 英語学習支援のための聴解コーパスの有効性の検証, FIT2012 第11回情報科学技術フォーラム, 2012年9月4-6日, 法政大学.
- ⑦ 岡本隆志, 南條浩輝, 吉見毅彦, 小谷克則: 英文読解速度テストに用いたテキストの難易度を考慮したテスト有効性の検証, FIT2012 第11回情報科学技術フォーラム, 2012年9月4-6日, 法政大学.
- ⑧ Kotani, K., Yoshimi, T., Nanjo, H. and Isahara, H.: Corpus Materials for Constructing Learner Corpus Compiling Speaking, Writing, Listening, and Reading, 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2012), Mar.12, 2012, New Delhi, India.
- ⑨ Kotani, K., Yoshimi, T., Nanjo, H. and Isahara, H.: Statistical Analysis of a Learner Corpus Integrating Linguistic Output and Language Processing Data,

7th International Conference Corpora and Language Technologies in Teaching, Learning and Research, Sep.11, 2011, Khon Kaen, Thailand.

- ⑩ Kotani, K. and Yoshimi, T.: A Scoring Method for Second Language Writing Based on Word Alignment, 12th Pacific Association for Computational Linguistic Conference (PACLING 2011), Jul.19, 2011, Kuala Lumpur, Malaysia.
- ⑪ Kotani, K., Yoshimi, T., Nanjo, H. and Isahara, H.: Compiling Learner Corpus Data of Linguistic Output and language Processing in Speaking, Listening, The 5th International Joint Conference on Natural Language Processing (IJCNLP2011), Nov. 11, 2011.
- ⑫ Kotani, K. and Yoshimi, T.: Classification of language learners' sentences into native speaker-like or non-native speaker-like sentences using learner sentences and machine translation sentences as training data, International Conference of Education, Research and Innovation(ICERI2010), Nov.15-17, 2010, Madrid, Spain.
- ⑬ Kotani, K., Yoshimi, T. and Isahara, H.: Validation of a reading speed test for Japanese Learners of English as a Foreign Language, The Teaching and Learning of Language International Conference 2010 Kyoto Conference, Sep.21-23, 2010, Kyoto, Japan.
- ⑭ Kotani, K. and Yoshimi, T.: A Machine Learning-Based Evaluation Method for Machine Translation, The 6th Hellenic Conference on Artificial Intelligence (SETN 2010), May.4-7, 2010, Athens, Greece.

[その他]

ホームページ等

<http://nlp.i.ryukoku.ac.jp/>

## 6. 研究組織

### (1) 研究代表者

小谷 克則 (KOTANI KATSUNORI)

関西外国語大学・外国語学部・准教授

研究者番号：30440994

### (2) 研究分担者

吉見 毅彦 (YOSHIMI TAKEHIKO)

龍谷大学・理工学部・准教授

研究者番号：50368031

井佐原 均 (ISAHARA HITOSHI)

豊橋技術科学大学・情報メディア基盤センター・教授

研究者番号：20358881

南條 浩輝 (NANJO HIROAKI)

龍谷大学・理工学部・助教

研究者番号：50388162

### (3) 連携研究者

該当なし