The KISTEC: 日本の大学生の発話データに基づく英語学習者話し言葉コーパスの構築

神澤 克徳 [1]　瀬戸口 彩花 [2]　田中 悠介 [3]　近 大志 [4]
小林 雄一郎 [5]　光永 悠彦 [6]　森 真幸 [1]　李 在鎬 [7]
[1] 京都工芸繊維大学　[2] 京都大学大学院　[3] 福岡大学　[4] 京都大学
[5] 日本大学　[6] 名古屋大学　[7] 早稲田大学
{kanzawa, morim}@kit.ac.jp
{guchi.a.chan7, yusuke.tanaka.07, kobayashi0721, jhlee.n}@gmail.com
chika.taishi.6p@kyoto-u.ac.jp　mitsunaga.haruhiko.p9@f.mail.nagoya-u.ac.jp

概要

本稿では，**The KIT Speaking Test Corpus**（以下，**KISTEC** と表記）の設計と仕様について報告する．KISTEC は日本の大学生英語学習者が受験したスピーキングテストの解答音声を基に構築した約 30 万語規模の話し言葉コーパスであり，書き起こしテキストとアノテーションから構成される．各種タグだけでなく，学習者の属性や，全体およびタスクごとのスコアを参照できるため，個人差やタスク特性を考慮した発話の分析を可能とする．KISTEC は L2 話者の非流暢性現象を自動検出する BERT モデルの性能評価にも使われており [1]，NLP 研究への応用も十分に可能である．

1　はじめに

現在の日本の英語教育は，スピーキングやコミュニケーション能力の向上に焦点が当てられている．これらを効果的に指導するには，指導および評価方法の改善やスピーキング練習教材の開発が不可欠である．しかし，そのための学習者の発話データは現状，不足している．この問題解決の一助となるべく，著者らは京都工芸繊維大学で開発・実施された英語スピーキングテストである KIT Speaking Test の解答音声を用いて **The KIT Speaking Test Corpus**（以下，**KISTEC** と表記）を構築し，Web 上で一般公開した [2].

KISTEC は発話に対する各種タグだけでなく，学習者の属性や，全体およびタスクごとのスコアを参照できることが特色の 1 つである．KISTEC の活用により，日本語を母語とする英語学習者の発話特徴の解明が進展し，エビデンスに基づく英語指導や言語テストの開発・改良，スピーキング練習教材の開発，自然言語処理モデルの改良など，英語教育と NLP の両面への貢献が期待できる．

2　先行コーパス

KISTEC に関連する先行コーパスとしては，Standard Speaking Test (SST) の解答音声に基づく NICT Japanese Learners of English (JLE) Corpus [3, 4]，アジア 10 カ国・地域の大学生・大学院生のスピーチやエッセイを収録した The International Corpus Network of Asian Learners of English (ICNALE) [5, 6]，日本の高校生の英語発話を 3 年間にわたって縦断的に記録した The Longitudinal Corpus of Spoken English (LOCSE) [7, 8, 9, 10] が挙げられる．表 1 に，これらのコーパスの概要を示す．

表 1　主な先行コーパス

コーパス名	発話形式	語数	公開
NICT JLE Corpus	対話	約 100 万	有
ICNALE	独話	約 50 万	有
	対話	約 160 万	
LOCSE	独話	約 40 万	無

なお，KISTEC を構築するにあたっては，NICT JLE Corpus を大いに参考にした．ファイル書式やタグは，NICT JLE Corpus に準拠している．

他にも小規模な話し言葉コーパスはいくつか構築されているものの，全体としては依然として少数に留まっている．書き起こしなどの作業に莫大なリソースを要すること，本格的なスピーキング教育が行われていない日本において，大規模なデータ収集

が困難であることが理由に挙げられる．英語学習者による言語習得の基礎研究や，英語教育への応用を進展するには，比較的大規模なコーパスを早急に整備する必要がある．

3 KISTEC の構築

3.1 データ収集

このような背景から，著者らは KIT Speaking Test の解答音声に基づき KISTEC を構築した．KIT Speaking Test は著者の神澤と森を含む京都工芸繊維大学の教員チームが開発したコンピュータ方式の英語スピーキングテストである [11].

テストの概要 KIT Speaking Test は，2018 年に京都工芸繊維大学の 1 年次生全員（当時）を対象に実施した．Ver. 1 から Ver. 3 まで合計 3 バージョンがあり，各受験者は 3 つのうちいずれかを受験した．どのバージョンもタスクタイプは共通しており，3 つのパート，合計 9 問から構成される．パートとタスクの内訳は以下の通りである:

- Part1 (Q1–3): 写真に基づいた想像や比較
- Part2 (Q4–7): 会話の要約および意見陳述
- Part3 (Q8–9): 構造化された意見陳述

各問の解答時間は 45 秒あるいは 60 秒であり，すべての問いにモノローグで解答する．Part3 では，60 秒のリハーサルタイムを設けている．[2] において，実際のテスト問題が出題時に用いた写真や会話などとともに公開されている．

採点基準 各問につき，十分な訓練を受けたネイティブスピーカー (NS) 1 名とノンネイティブスピーカー (NNS) 1 名（フィリピンの英語講師）のペアが採点を行った．採点者は問題ごとに異なり，合計 18 名（NS，NNS とも 9 名）が担当した．採点者に NNS を加えたのは，KIT Speaking Test がリンガフランカ（国際語）としての英語運用能力測定を目的としているためである．採点基準は Task Achievement（TA: 問題で求められているタスクをどの程度達成できたか）と Task Delivery（TD: 自らのスピーチをどの程度効果的に伝えられたか）の 2 つであり，いずれも 0–5 の 6 段階で評価した．2 人の採点に 2 点以上の開きがあった場合は，シニアレイター（採点に熟達した NS）が再採点を行った．1 点差以内の場合は，2 人の平均を取った（採点基準の詳細は付録の表 4 を参照）．

TA と TD は強く相関している ($r = 0.88$)．これは，TD で評価される流暢性が低いと制限時間内に発話できる語数が少なくなってしまい，結果的に TA で評価されるタスクを十分に達成できないためであると考えられる．しかし，TA と TD は根本的に別の評価観点である．例えば，(1) は ‘Some friends from another country are visiting you for one week. Choose a place for them to go and explain why they should go there.’ というタスクに対する解答例である．

(1) I want you go to England because I like um European, and I want to eat England fish and chips. I think I go airplane. Uh, airplane is pi airplane's price is very high, but airplane is aren't good experience. I think England English is many many lucky experience.

この解答の TD の評価は 3 点であるのに対し，TA は 1 点である．これは，海外から来た友人に勧める場所を挙げることが求められているにも関わらず，イングランドに行くよう勧めており，タスクを適切に遂行できていないためである．

採点終了後に，項目反応理論を用いて受験者の能力値を算出し，Overall score (0–100)，TA rank (0–5)，TD rank (0–5) の 3 つの等化スコアを算出した．これらのスコアはバージョン間で比較可能である．KIT Speaking Test ではタスク達成度が重視されるため，Overall score については，TA が 80%，TD が 20%になるように重み付けされている．

スコアの分布 全受験者のうち，同意が得られた 575 名[1)]の解答音声に対して書き起こしとアノテーションを行った．575 名の内訳は，約 97%が日本人で，残り 3%が中国，マレーシア，韓国からの留学生であった．彼らの KIT Speaking Test と直近に受験した TOEIC L&R のスコアは，表 2 のとおりである．

表 2 KIT Speaking Test と TOEIC L&R のスコア

スコア種別	平均（標準偏差）	最高	最低
Overall score	48.22 (10.45)	90	20
TA rank	2.97 (1.42)	5	1
TD rank	2.98 (1.40)	5	1
TOEIC L&R	563.54 (133.15)	985	195

KIT Speaking Test のスコアと TOEIC トータルスコア，リーディングスコア，リスニングスコアとの相

1) そのうちの 1 名は，システムトラブルにより音声が収録できていなかったため，除外した．

関係数はそれぞれ 0.59，0.56，0.54 であり，中程度の相関を示している．

3.2 書き起こし・アノテーション

方法 書き起こしおよびアノテーションは研究代表者の統括のもと，言語学や英語教育学を専攻する大学院生（当時）が行った．書き起こしの負担を軽減すべく，Azure Video Indexer（VI: 現 Azure AI Video Indexer）の Speech to Text 機能を利用した．[2)] 作業は以下の 3 段階から構成される．

1. VI による解答音声の自動書き起こし
2. 作業者による書き起こしの修正とタグの付与
3. 作業者間のクロスチェック

作業に先立ち，NICT JLT Corpus のガイドライン [12] に準拠したマニュアル [13] を作成した．これを参照しつつ，研究代表者と作業者間で適宜相談の機会を設け，作業内容を可能な限り統一させた．

研究倫理 KISTEC の構築にあたっては，京都工芸繊維大学「ヒトを対象とする研究倫理審査」において研究計画の承認を得た．KIT Speaking Test の受験者には，研究計画を説明し，同意を得た者の解答音声のみをコーパス化した．書き起こし・アノテーション作業者とは，作業で知り得た情報を第三者に口外しない，データの管理を徹底するなどの取り決めを交わした．

4 コーパスの特徴

コーパスサイズ 表 3 に KISTEC のコーパスサイズを示す．合計約 30 万語と，2 節に挙げた先行コーパスよりは小規模であるが，同種のコーパスの中では比較的大規模である．

表 3 KISTEC のコーパスサイズ

バージョン	受験者数	総語数	総解答時間
Ver. 1	193	98507	24:55:45
Ver. 2	190	96945	24:32:30
Ver. 3	191	95002	24:40:15
合計	574	290454	74:08:30

発話特徴の分析 発話の書き起こしには繰り返しや自己訂正，フィラーなど，発話特徴を表すタグが付与されている．タグ一覧は付録の表 5 を参照されたい．付与したタグは NICT JLE Corpus に準拠しているが，NICT JLE Corpus はダイアローグであるのに対して KISTEC はモノローグであるため，一部のタグを改変した．アノテーションの詳しい仕様は [13] を参照されたい．

タグの例を 1 つ紹介する．自己訂正タグ (<SC></SC>) は，学習者自身で発話の修正を行ったとみなされる場合に，修正前の表現に付与した．(2) の場合，‘health’ を ‘healthy’ に訂正するとともに ‘body’ を追加し，さらに ‘healthy body’ の前に ‘if I have a’ を追加することで，最終的に ‘if I have a heathy body’ に至ったと考えられる．この場合，‘health’ や ‘healthy body’ は訂正を受けた要素とみなされ，(2) のようにアノテーションされる．自己訂正を分析することで，学習者がどこでつまづき，どのように克服したかという発話の産出過程を明らかにできる．

(2) <SC>health</SC> <SC>healthy body</SC> if I have a healthy body,

KISTEC を用いた発話特徴の分析研究としては，フィラー (e.g., uh, you know) の使用率および種類と習熟度の関係を分析した [14]，フィラーの生起位置と流暢性の関係を分析した [15]，自己訂正を 4 タイプに分類し，他の非流暢性現象（反復，フィラー）との共起関係を分析した [16] などがある．

ヘッダー情報と発話の関連性の分析 KISTEC には受験者ごとに，受験者属性，KIT Speaking Test のスコア，TOEIC スコア，英語学習経験といったヘッダー情報が付与されている．ヘッダー情報の一覧は付録の表 6 を参照されたい．ヘッダー情報のうち，国籍と性別は京都工芸繊維大学の許可を得て，学生本人が大学に申告した情報を使用した．また，<experience 1> から <experience 7> までは，国内で TOEIC を主催する国際ビジネスコミュニケーション協会が TOEIC 試験時に行う英語学習経験などについてのアンケート結果を引用した．

ヘッダー情報を参照することで，KIT Speaking Test や TOEIC のスコアを基に学習者の習熟度を予測し，習熟度と発話内容の関連性を分析することが可能である．また，タスクごとにスコアが付与されているため，個別のパフォーマンスを詳細に分析できる．2 節で紹介した先行コーパスでも，習熟度を示す指標（スピーキングテストスコアや Common European Framework of Reference for Languages (CEFR) レベル [17]）が付与されているが，これらはタスク

2) https://azure.microsoft.com/ja-jp/products/ai-video-indexer

ごとのスコアではなく，学習者ごとのスコアやレベルである．そのため，同一受験者のパフォーマンスがタスクごとに異なっていた場合でも，これを詳細に分析することは困難である．

また，英語学習経験の情報（付録表6の <experience 1> から <experience 7>）を用いることで，英語学習経験と発話内容の関係性を分析できる．例えば，英語圏の滞在の有無がフィラーの使用頻度や種類に与える影響といった分析が可能である．

タスクが発話に与える影響の分析　KISTEC では，コーパスに収録されている発話を引き出したタスクがすべて公開されている．例えば，NICT JLE Corpus は SST というスピーキングテストに基づいているが，民間（商用）のテストであるため，タスクの具体的な内容は公開されていない．また，ICNACE では，タスクは2つ（「大学生のアルバイトの是非」と「レストランにおける全面禁煙の是非」）に限定されている．このため，タスクの違いが発話に与える影響の十分な分析は困難である．

一方 KISTEC は，すべてのタスクを出題時に用いた写真や会話などとともに公開している．また，3.1 で述べたように，タスクのバリエーションも豊かである．このため，タスクタイプが異なる場合 (e.g., 意見陳述 vs 会話要約)，また，タスクタイプが同じであっても問題が異なる場合（e.g., 想像に関するタスク (Question 1, 2) のバージョン間比較）において，学習者の発話にどのような影響があるかを分析可能である．さらに，リハーサルタイム（準備時間）の有無が発話に与える影響についても分析できる．

NLP への応用可能性　KISTEC は，NLP への貢献も期待される．英語教育関連では，スピーキング練習教材やツールの開発への応用が見込まれる．KISTEC を活用して，日本語を母語とする英語学習者特有の誤用や間違いを特定することで，それらに焦点を当てた問題を自動生成するツールを開発できる可能性がある．また，習熟度に応じた最適な問題の生成も可能である．さらに，日本語を母語とする英語学習者に特有の言い回しや表現，および発話特徴を自然言語処理モデルに取り入れることで，L2 英語話者の発話（またはライティング）をよりよく理解するためのモデルや，彼らにとってより分かりやすい表現を生成するモデルの構築も期待される．例えば，[1] は，KISTEC を用いて L2 話者の非流暢性現象を自動検出する BERT モデルの性能評価を実施している．また，著者らは未公開ながら，コーパスの元となる音声データも保有している．この音声データを活用することで，将来的には音声的特徴を取り入れたモデルの構築も可能である．KISTEC が英語教育研究と NLP の交流を促進するきっかけとなることが期待される．

5　おわりに

本稿では，The KIT Speaking Test Corpus (KISTEC) の構築プロセスや設計特徴を紹介した．日本語を母語とする英語学習者話し言葉コーパスは数が少なく，多くの発話特徴が未解明に留まっている．KISTEC を利用した分析を行うことで，基礎的な知見の蓄積が期待される．また，英語教育の指導法やカリキュラムの改善，スピーキング練習ツールなどの教材開発，英語スピーキングテストを含むテスト開発や妥当性研究，自然言語処理モデルの改良などへの応用も期待される．

しかし，KISTEC には課題もある．1 点目は，対象となる学習者が京都工芸繊維大学の 1 年次生（当時）のみであるため，習熟度にあまり差がない．KIT Speaking Test や TOEIC のスコアにはある程度のばらつきが見られるものの，大半の学生の英語スピーキング力は CEFR の A2 から B1 に収まると考えられる．そのため，特に習熟度の高い学習者（CEFR B2 以上）の発話特徴を調査するには不十分である．

2 点目は，KISTEC がモノローグ形式の英語スピーキングテストの解答音声に基づいている点である．このため，自然発話とは異なる特徴を含む可能性が否定できない．KIT Speaking Test などのスピーキングテストは，心理的負担が比較的高く，応答時間も限られる条件下で行われるため，特殊なコンテクストに基づく発話といえる．また，モノローグとダイアローグで発話の特徴が異なる部分もある（e.g., フィラーの使用）．したがって，KISTEC をモノローグ形式のスピーキングテストの発話として分析することには問題ないが，自然発話として分析する際には注意が必要である．

これらの課題については，より習熟度が高い／低い英語学習者の発話や，自然発話を収録した他のコーパスを援用し，分析を進めていくことが有効である．また，1 点目に関連して，著者らは NS や英語習熟度の高い学習者を対象とした KISTEC の対照コーパスの構築を進めている．この対照コーパスが完成すれば，より幅広い習熟度にわたる英語発話の特徴を連続的に分析することが可能となる．

謝辞

本研究は JSPS 科研費 22K00736, 19K00849 の助成を受けたものです．KIT Speaking Test の開発・運営に関わった方，および，KISTEC の書き起こし・アノテーション作業者に心より感謝いたします．

参考文献

[1] Lucy Skidmore and Roger K. Moore. Bert models for spoken learner english disfluency detection. **Proceedings of SLaTE 2023**, pp. 91–92, 2023.

[2] Katsunori Kanzawa, Yuichiro Kobayashi, Jaeho Lee, Haruhiko Mitsunaga, Masayuki Mori, Yusuke Tanaka, and Taishi Chika. The KIT Speaking Test Corpus, n.d. https://kitstcorpus.jp.

[3] 国立研究開発法人情報通信研究機構（NICT）. 日本人 1200 人による英語コーパス・The NICT JLE (Japanese Learner English) Corpus, n.d. https://alaginrc.nict.go.jp/nict_jle/index.html.

[4] 和泉絵美, 内元清貴, 井佐原均. 日本人 1200 人の英語スピーキングコーパス. アルク, 2004.

[5] Shin'ichiro Ishikawa. The International Corpus Network of Asian Learners of English., n.d. https://language.sakura.ne.jp/icnale/.

[6] Shin'ichiro Ishikawa. **The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English**. Routledge, 2023.

[7] Mariko Abe. LOCSE Reserach Project, n.d. https://sites.google.com/view/locse/.

[8] Mariko Abe and Yusuke Kondo. Constructing a longitudinal learner corpus to track L2 spoken English. **Journal of Modern Languages**, Vol. 29, No. 1, pp. 23–44, 2019.

[9] Yuichiro Kobayashi, Mariko Abe, and Yusuke Kondo. Exploring L2 spoken developmental measures: Which linguistic features can predict the number of words? **English Corpus Studies**, Vol. 29, pp. 1–18, 2022.

[10] Mariko Abe, Yuichiro Kobayashi, and Yusuke Kondo. Capturing chronological variation in L2 speech through lexical measurements and regression analysis. **Applied Corpus Linguistics**, Vol. 4, No. 3, 2024. Article 100105.

[11] 神澤克徳, 羽藤由美. CBT スピーキングテストの舞台裏，どこがどう難しいのか？ KIT Speaking Test の実践より. JACET 関西支部紀要, Vol. 23, pp. 96–120, 2021.

[12] The NICT JLE Corpus 書き起こし・基本タグ付与ガイドライン ver.2.1.3,（2024-12 閲覧）. https://alaginrc.nict.go.jp/nict_jle/src/readme_transcription.pdf.

[13] KIT Speaking Test Corpus 書き起こし・タグ付与ガイドライン Ver. 1,（2024-12 閲覧）. https://kitstcorpus.jp/wp-content/uploads/2022/04/manual-1.pdf.

[14] 田中悠介, 瀬戸口彩花, 近大志, 神澤克徳. 日本語を母語とする英語学習者が使用するフィラーの分析：習熟度との関連性および英語母語話者との比較から. 英語コーパス学会大会予稿集 2023, pp. 13–18, 2023.

[15] 田中悠介, 瀬戸口彩花, 近大志, 神澤克徳. 日本人英語学習者の発話におけるフィラーの生起位置と習熟度の関係性. 英語コーパス学会大会予稿集 2024, pp. 73–78, 2024.

[16] 近大志, 瀬戸口彩花, 田中悠介, 神澤克徳. 英語学習者の発話にみられる非流暢性に関する考察：自己訂正と反復・フィラーの関係性. 言語処理学会第 30 回年次大会発表論文集, pp. 3166–3170, 2024.

[17] 吉島茂, 大橋理枝（訳・編）. 外国語の学習，教授，評価のためのヨーロッパ共通参照枠 追補版. 朝日出版社, 2014. https://www.goethe.de/resources/files/pdf191/cefr31.pdf.

A 付録

表 4 採点基準

Score	Task Achievement (80% weighting)	Task Delivery (20% weighting)
5	The task is achieved, being developed with a satisfactory level of detail.	The delivery is mostly confident. Given time is well used without obvious problems with delivery such as intrusive pauses, hesitations, or repetitions.
4	The task is mostly achieved, with some supporting detail in places.	Given time is quite well used despite some problems with delivery such as slow rate of speech, pauses, hesitations, or repetitions.
3	The task is minimally or partially achieved, being supported with some basic detail.	General meaning comes across, but given time is not effectively used because of problems with delivery such as slow rate of speech, pauses, hesitations, or repetitions.
2	The task is addressed, but there is no or very little supporting detail.	The speaker keeps trying, but problems with delivery (e.g., slow rate of speech, pauses, hesitations or repetitions) allow a very limited amount of meaning to be conveyed.
1	The task remains essentially unachieved, though there may be some relevant words.	The speaker gives up trying, or problems with delivery (e.g., slow rate of speech, pauses, hesitations, repetitions) are fatal to meaning coming across.
0	There is no relevant contribution (e.g., content is entirely unconnected to topic).	The speaker does not start the task (e.g. s/he is silent, utters only fillers, or just says, 'I don't know').

表 5 タグ一覧

タグ	用途	タグ	用途
<F></F>	フィラー・あいづち・感動詞	<CO></CO>	途中で中断した発話
<R></R>	繰り返し（聞き取りに自信がある）	<?></?>	聞き取りに自信がない語
<R?></R?>	繰り返し（聞き取りに自信がない）	<??></??>	全く聞き取り不可能な語
<SC></SC>	自己訂正（聞き取りに自信がある）	<H pn = "X"></H>	固有名詞・差別用語など
<SC?></SC?>	自己訂正（聞き取りに自信がない）	<JP></JP>	日本語
<TO></TO>	タイムアウト	<.></.>	2 秒〜3 秒のポーズ
<RE></RE>	レコーディングエラー	<..></..>	3 秒以上のポーズ
<nvs></nvs>	非言語音	<laughter></laughter>	笑いながらの発話

表 6 ヘッダー情報一覧

ヘッダー情報	説明
<grade>	学年
<nationality>	国籍
<sex>	性別
<version>	KIT Speaking Test のテストバージョン
<total_score>	KIT Speaking Test のトータルスコア
<ta_rank>	KIT Speaking Test の TA ランク
<td_rank>	KIT Speaking Test の TD ランク
<toeic_score>	TOEIC のトータルスコア
<toeic_rscore>	TOEIC のリーディングセクションのスコア
<toeic_lscore>	TOEIC のリスニングセクションのスコア
<experience1>	何年間，英語を学習していますか.
<experience2>	次のうち，最も重要視する／していた英語の技能はどれですか.
<experience3>	日常生活において英語を使用する割合はどのくらいですか.
<experience4>	次の英語技能のうち，最もよく使用するものはどれですか.
<experience5>	英語のやりとりに苦労する頻度はどのくらいですか.
<experience6>	英語を主言語とする国に滞在したことがありますか.
<experience7>	英語圏に滞在した主な目的は何でしたか.