

詳細なメタデータを含む英語学習者話し言葉コーパスの構築： 発話特徴の包括的な分析に向けて

神澤 克徳 (京都工芸繊維大学) *

田中 悠介 (熊本学園大学)

近 大志 (京都大学)

瀬戸口 彩花 (都城市立明和小学校)

小林 雄一郎 (日本大学)

光永 悠彦 (名古屋大学)

森 真幸 (京都工芸繊維大学)

李 在鎬 (早稲田大学)

Constructing a Spoken Corpus of English Learners with Detailed Metadata: Towards a Comprehensive Analysis of Speech Features

Katsunori Kanzawa (Kyoto Institute of Technology)

Yusuke Tanaka (Kumamoto Gakuen University)

Taishi Chika (Kyoto University)

Ayaka Setoguchi (Meiwa Elementary School, Miyakonojo)

Yuichiro Kobayashi (Nihon University)

Haruhiko Mitsunaga (Nagoya University)

Masayuki Mori (Kyoto Institute of Technology)

Jaeho Lee (Waseda University)

Abstract

This study⁽¹⁾ presents an overview of the Kyoto Institute of Technology (KIT) Speaking Test Corpus (hereafter, the KISTEC), constructed by the authors, and illustrates its utility with a sample analysis. The KISTEC is an approximately 300,000-word corpus of spoken English produced by learners, based on responses elicited by the KIT Speaking Test administered to all first-year students at KIT. The transcriptions are annotated with tags capturing speech features and are accompanied by metadata, including learner attributes; speaking-test scores (overall and by task); and TOEIC L&R scores. All task prompts used to elicit responses are publicly available. As speaking is strongly emphasized in English education in Japan, the linguistic characteristics and developmental trajectories

* kanzawa[at]kit.ac.jp

⁽¹⁾ This study is a reconstruction of Kanzawa et al. (2025) and Tanaka et al. (to appear).

of Japanese-speaking learners of English need to be clarified. Using the KISTEC enables comprehensive analyses of learner utterances in consideration of individual differences and task characteristics. We expect the KISTEC to yield new insights into Japanese learners' English-speaking abilities, contributing to English education and related fields.

1. Introduction

Improving communication skills, especially speaking skills, is currently emphasized in Japanese English education. To support effective teaching and learning, it is imperative to clarify the linguistic characteristics and developmental trajectories of Japanese-speaking learners of English. However, learner spoken data remain scarce. To address this gap, we constructed the Kyoyo Institute of Technology (KIT) Speaking Test Corpus (hereafter, the KISTEC) from audio recordings of the KIT Speaking Test—an English speaking test developed and administered at KIT—and made it publicly available online (Kanzawa et al. n.d.). We also released a new search interface for the KISTEC (Tanaka et al. n.d.).

The most significant feature of the KISTEC is that it provides detailed metadata alongside transcribed utterances. The transcriptions are annotated with tags representing speech features and are accompanied by metadata, including learner attributes; speaking-test scores (overall and by task); and TOEIC L&R scores. Furthermore, all tasks that served as prompts for the responses are made publicly available. The use of the KISTEC facilitates comprehensive analyses of learner utterances that account for individual differences and task characteristics. We expect the KISTEC to yield new insights into the English speaking abilities of Japanese-speaking learners of English, contributing to English education and related fields.

2. Previous Corpora

Corpora related to the KISTEC include the NICT Japanese Learners of English (JLE) Corpus, based on audio recordings of the Standard Speaking Test (SST) (National Institute of Information and Communications Technology n.d.a, Izumi et al. 2004); the International Corpus Network of Asian Learners of English (ICNALE), which contains speeches and essays from university and graduate students across 10 countries and regions in Asia (Ishikawa n.d., 2023); and the Longitudinal Corpus of Spoken English (LOCSE), which is made through longitudinally recording English utterances from Japanese high school students over three years (Abe n.d., Abe and Kondo 2019, Kobayashi et al. 2022, Abe et al. 2024). Table 1 provides an overview of these corpora.

Table 1 Representative Preceding Corpora

Corpus Name	Speech Type	Word Count (Approx.)	Availability
NICT JLE Corpus	Dialogue	1,000,000	Available
ICNALE	Monologue	500,000	Available
	Dialogue	1,600,000	
LOCSE	Monologue	400,000	Unavailable

In constructing the KISTEC, we referred to the NICT JLE Corpus, adopting its file formats and tagging scheme.

Although other smaller learner spoken corpora have been developed, they remain relatively scarce probably because of the substantial resources required for transcription and annotation and because of the difficulty of large-scale data collection in Japan, where comprehensive speaking education has only been implemented. To advance foundational research on the English acquisition of Japanese-speaking learners and its applications to English education in Japan, establishing a relatively large-scale corpus is an urgent priority.

3. Construction of the KISTEC

3.1 Data Collection

Against this backdrop, we constructed the KISTEC based on audio recordings from the KIT Speaking Test, which is a semi-direct computer-based English speaking test developed by a team of faculty members at KIT, including the authors of the study, Kanzawa and Mori (Kanzawa and Hato 2021).

3.1.1 Overview of the Test

The KIT Speaking Test was conducted in 2018 for all first-year students at Kyoto Institute of Technology. The test had three versions (Ver. 1–3), and each examinee took one version. All versions shared the same task types and consisted of three parts, totaling nine questions. The breakdown of parts and questions is as follows:

- Part 1 (Q1–3): Imagination and comparison based on photographs
- Part 2 (Q4–7): Summary of conversations and expression of opinions
- Part 3 (Q8–9): Structured expression of opinions

The response time for each question was either 45 or 60(s), with all responses in monologue form. In Part 3, a rehearsal time of 60(s) was provided. The actual test questions, along with the photographs and conversations used for the test, are publicly available at Kanzawa et al. (n.d.).

3.1.2 Scoring Criteria

Each question was scored by a pair of one native speaker (NS), an English lecturer at a Japanese university, and one non-native speaker (NNS), an English instructor from

the Philippines, both of whom received appropriate training. Raters varied by question, totaling 18 individuals (9 NSs and 9 NNSs). The inclusion of NNS raters reflected the the objective of the KIT speaking test: to assess participants' proficiency as a lingua franca.

The scoring rubric comprises two components: Task Achievement (TA: the extent to which the question's task was fulfilled) and Task Delivery (TD: the effectiveness of delivery, i.e., fluency), each rated on a 6-point scale (0–5). If the two raters' scores differed by two points or more, a senior rater (an NS experienced in scoring) conducted a reassessment. If the difference was one point, the two scores were averaged. For detailed scoring criteria, see Table 2.

Table 2 Scoring Criteria

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').

TA and TD are strongly correlated ($r = 0.88$). We attribute this correlation to the fact that low fluency, evaluated under TD, reduces the number of words produced within the time limit, which in turn constrains the participants' ability to achieve the task evaluated under TA. Nevertheless, TA and TD represent fundamentally different

evaluative perspectives. For example, (1) shows a response to the task “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.”

- (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.

In this response, the TA score is 1 whereas the TD score is 3. This discrepancy arises because, although the task requires identifying a place to recommend to friends visiting from abroad, the response frames the choice as going to England, indicating insufficient task fulfillment.

After scoring was completed, the item response theory was used to estimate examinees’ proficiency, resulting in three equalized scores: Overall score (0–100), TA rank (0–5), and TD rank (0–5). These scores are comparable across versions. Given the emphasis on task achievement in the KIT Speaking Test, the Overall score is calculated with TA weighted at 80% and TD at 20%.

3.1.3 Score Distribution

We transcribed and annotated the audio responses of 575 consenting examinees⁽²⁾. Of these 575 examinees, about 97% were Japanese, and the remaining 3% were international students from China, Malaysia, and South Korea. Table 3 shows the examinees’ KIT Speaking Test scores and their most recent TOEIC L&R scores.

Table 3 Scores from the KIT Speaking Test and TOEIC L&R

Score Type	Mean	SD	Max.	Min.
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

The correlation coefficients between the KIT Speaking Test scores and the TOEIC L&R overall, Reading, and Listening scores are 0.59, 0.56, and 0.54, respectively, indicating moderate correlations.

⁽²⁾ One examinee was excluded due to a system error that prevented audio recording.

3.2 Transcription and Annotation

3.2.1 Method

Transcription and annotation were conducted by graduate students specializing in linguistics and English education under the supervision of the principal investigator. To reduce workload, a Speech-to-Text tool (Azure Video Indexer: VI, currently Azure AI Video Indexer) was used.⁽³⁾ The workflow comprised three stages:

1. Automatic transcription of the audio responses with VI
2. Manual correction of the auto-transcriptions and tagging by annotators
3. Cross-checking across annotators

Prior to commencement of the work, a manual was prepared (Tanaka and Kanzawa n.d.) in accordance with the guideline of the NICT JLT Corpus (National Institute of Information and Communications Technology n.d.b). Guided by this manual, the principal investigator and the annotators held meetings to ensure maximum consistency in their work.

3.2.2 Research Ethics

In constructing the KISTEC, approval for the research plan was obtained from KIT's "Ethics Review for Research Involving Human Subjects." Examinees in the KIT Speaking Test were informed of the research plan, and only the audio responses of those who provided consent were included in the corpus. Annotators were bound by confidentiality agreements regarding any information obtained during their work and adhered to strict data management protocols.

4. Characteristics of the Corpus

4.1 Corpus Size

Table 4 summarizes the size of the KISTEC. At approximately 300,000 words, the KISTEC is smaller than the representative corpora reviewed in Section 2, yet relatively large among comparable corpora of this type.

Table 4 Corpus Size of the KISTEC

Version	N of Examinees	Total Words	Total Response Time
Ver. 1	193	98,507	24:55:45
Ver. 2	190	96,945	24:32:30
Ver. 3	191	95,002	24:40:15
Total	574	290,454	74:08:30

⁽³⁾ <https://azure.microsoft.com/en-us/products/ai-video-indexer>

4.2 Analysis of Speech Features

Transcriptions include tags that capture speech features, such as repetition, self-correction, and fillers. Table 5 shows a list of tags. The tagset is based on the NICT JLE Corpus; however, because the NICT JLE is dialogue-based whereas the KISTEC is monologue-based, some tags were adapted. For detailed annotation specifications, see Tanaka and Kanzawa (n.d.).

Table 5 Tagset

Tags	Meaning
<F></F>	Filler
<CO></CO>	Cutoff (suspended speech)
<R></R>	Repetition
<R?></R?>	Repetition (not confident in listening)
<SC></SC>	Self-correction
<SC?></SC?>	Self-correction (not confident in listening)
<?></?>	Not confident in listening
<??></??>	Completely inaudible
<H pn = "X"></H>	Proper nouns, discriminatory terms, etc.
<JP></JP>	Japanese
<TO></TO>	Timeout
<RE></RE>	Recording error
<.></.>	Pause (2–3 sec.)
<..></..>	Pause (3 or more sec.)
<nvs></nvs>	Non-verbal sound
<laughter></laughter>	Laughing while speaking

One example is the self-correction tag (<SC></SC>), which is applied to expressions that the learner appears to have corrected. In example (2), the learner corrects “health” to “healthy” and adds “body,” yielding the final expression “if I have a healthy body.” In this instance, “health” and “healthy body” are treated as the corrected elements and are annotated as shown in (2). By analyzing self-corrections, we could clarify the speech-production process, revealing where learners struggle and how they overcome these challenges.

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

In Section 5, we will introduce a sample analysis of utilizing the tagged speech features, specifically focusing on the filler tag.

4.3 Analysis of the Relationship between Header Information and Speech

The KISTEC includes header information for each examinee, such as learner attributes, KIT Speaking Test scores, TOEIC L&R scores, and English learning experience.

A list of the header fields is provided in Table 6. Nationality and gender are based on information that students self-reported to KIT, used with the institution’s permission. In addition, <experience 1> through <experience 7> are derived from a survey on English learning experiences conducted by the Institute for International Business Communication, which administers the TOEIC in Japan.

Table 6 Header Information

Header information	Meaning
<grade>	University year
<nationality>	Nationality
<sex>	Sex (1 for male, 2 for female)
<version>	Version of the speaking test
<total_score>	Score in the speaking test (0–100)
<ta_rank>	Task Achievement rank in the speaking test (0–5)
<td_rank>	Task Delivery rank in the speaking test (0–5)
<ta1–9_score>	Task Achievement raw score for each response (0–5)
<td1–9_score>	Task Delivery raw score for each response (0–5)
<toEIC_score>	TOEIC L&R score (10–990)
<toEIC_rscore>	Score of Reading Section in TOEIC (5–495)
<toEIC_lscore>	Score of Listening Section in TOEIC (5–495)
<experience1>	How many years have you studied English?
<experience2>	Which of the following language skill(s) is/are the most important for you?
<experience3>	What percentage do you use English in your daily life?
<experience4>	Which of the following language skills do you use the most?
<experience5>	How often does your lack of English proficiency prevent your communication?
<experience6>	Have you ever stayed in a country where English is the primary language?
<experience7>	What was the purpose of your stay in a country where English is the primary language?

The KIT Speaking Test and TOEIC L&R scores serve as predictors of learner proficiency, enabling analyses of the relationship between proficiency and speech content. Furthermore, because scores are assigned at the task level, individual performance can be examined in detail. In the corpora reviewed in Section 2, proficiency indicators (e.g., speaking test scores and Common European Framework of Reference for Languages (CEFR) levels (Yoshijima and Ohashi 2014)) are provided, but these are overall scores or levels at the learner level rather than the task level. Consequently, detailed analyses are difficult especially when an examinee’s performance varies across tasks.

In addition, information on English learning experiences (<experience 1> through <experience 7>) allows us to analyze the relationship between these experiences and

speech content. For example, we can examine whether stays in English-speaking countries affect the frequency and the types of used fillers.

4.4 Analysis of the Influence of Tasks on Speech

In the KISTEC, all tasks used to elicit the recorded speech are publicly available. In the other corpora reviewed in Section 2, specific tasks were not disclosed because the tests were proprietary (commercial), or the number of tasks was limited. Consequently, it is difficult to conduct sufficiently detailed analyses of task effects on speech.

By contrast, the KISTEC provides all tasks together with the photos and conversations used for administration. Furthermore, as discussed in Section 3.1.1, task variety is substantial. This enables analyses of how different task types (e.g., opinion statements vs. conversation summaries) and different prompts within the same task type (e.g., cross-version comparisons of imaginative tasks in Questions 1 and 2) affect learners' speech. In addition, the impact of the presence or absence of rehearsal (preparation) time on speech can be analyzed.

5. Analysis Example

5.1 Purpose

In this section, based on Tanaka et al. (to appear), we present a sample of the analysis of KISTEC to examine the relationship between the rate of filler use and perceived fluency, defined as listeners' judgments of fluency (Segalowitz 2010).

Evaluating fluency is an effective way to assess language learners' speaking abilities (Ellis 2003). Disfluencies—phenomena that negatively affect fluency—play a crucial role in such assessments. One type of disfluency is fillers (Biber et al. 1999). Fillers include sounds such as *eh*, *uh*, *um*, *er*, and the like (Eklund 2004: 207). The focus on fillers stems from their controversial status. Although these fillers are generally considered to constitute a type of disfluency, several studies have shown that more fluent learners more frequently use them (Cenoz 1998, Iwashita et al. 2008, Kosmala and Morgenstern 2019, Préfontaine and Kormos 2016, Rieger 2003). However, a binary classification of fillers as either markers of fluency or disfluency may be overly simplistic. Tottie (2014) argues that fillers aid comprehension when not overused, and Lickley (2015) suggests that achieving native-like levels of disfluency signifies high second-language proficiency.

Based on these insights, Tanaka et al. (to appear) hypothesized that fillers may enhance fluency when used at rates comparable to those of native speakers, whereas excessive use may impair fluency. More specifically, Tanaka et al. (2023) analyzed data from 38 native English speakers who participated in the KIT Speaking Test and found an average filler usage rate of 3.47%. This rate was used as a benchmark, suggesting that filler usage of up to approximately 3.5% could enhance fluency assessments without negative effects.

However, rates exceeding this threshold may cause perceptions of filler usage to shift from enhancing fluency to signaling disfluency.

5.2 Procedures

To test this hypothesis, Tanaka et al. (to appear) used the KISTEC to examine the relationship between filler usage and fluency assessments as reflected in the TD score. The KISTEC includes information on examinees' nationalities. In their study, which focuses on English learners who are Japanese, 558 examinees identified as Japanese nationals were selected for analysis from a total of 574 examinees. Each examinee responded to 9 questions, yielding 5,022 responses for analysis. One response was excluded owing to absence of spoken words, resulting in a final dataset of 5,021 responses comprising 273,158 words, of which 26,174 are fillers.

Tanaka et al. (to appear) first computed the filler usage rate for each of the 5,021 responses as the number of fillers divided by the total number of words. They then examined the relationship between these filler usage rates (0–1) and TD score (0–5 in 0.5 increments) using a generalized additive model, in which the filler usage rate served as the independent variable and TD score as the dependent variable.

5.3 Results

The results of the analysis are illustrated in Figure 1, which shows that TD scores generally decrease as filler usage rates increase.

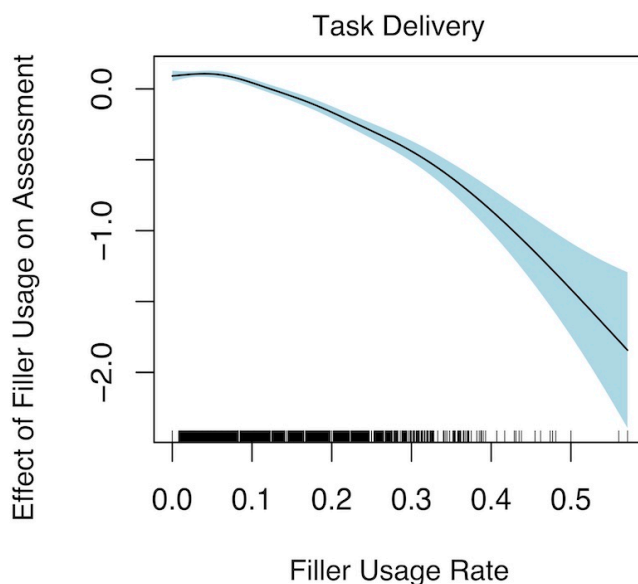


Figure 1 Relationship between Filler Usage Rates and TD Scores

To examine the relationship in greater detail, Tanaka et al. (to appear) calculated differential coefficients for every 5% increase in the filler usage rate (i.e., five-percentage-

point increments); these are reported in Table 7. Positive coefficients indicate a beneficial effect on fluency assessments, whereas negative coefficients indicate detrimental effects. The magnitude of the coefficients quantifies the size of these effects.

Table 7 Differential Coefficients by Filler Usage Rate

Filler Usage Rate	Differential Coefficient of TD
0%	0.091
5%	0.105
10%	0.043
15%	-0.052
20%	-0.164
25%	-0.296
30%	-0.439
35%	-0.626
40%	-0.857
45%	-1.124
50%	-1.414

The relationship between the filler-usage rate and TD was non-linear. At lower rates (0–10%), the smoothed effect on TD was slightly positive, with a peak at 5% (0.105). To precisely identify the rate at which the effect on TD scores is maximized, Tanaka et al. (to appear) computed a local maximum; this analysis showed an optimal filler usage rate of 4.1% with a maximum coefficient of 0.107. However, the positive effects diminished rapidly exceeding approximately 15%, with coefficients declining steadily to -1.414 at 50%. The turning point at which the smoothed effect shifts from positive to negative was approximately 12.28%. These findings suggest that filler-usage rates below roughly 12% may have a relatively positive impact on assessments, whereas rates at or above this threshold are associated with negative outcomes for TD.

5.4 Discussion

The results support the hypothesis that moderate filler use—up to approximately 4%—does not adversely affect fluency assessments and may even enhance them, consistent with the arguments of Tottie (2014) and Lickley (2015). These findings also corroborate the hypothesis that excessive filler use detracts from fluency: the effect shifts from positive to negative once usage exceeds a critical threshold of roughly 12.5%, beyond which fillers negatively affect perceived fluency. These observations challenge the simplistic view of fillers as mere markers of disfluency and instead underscore their complex role in spoken language: at optimal levels, fillers can enhance fluency, whereas their overuse compromises it.

6. Conclusion

This study presented an overview of the KISTEC and its features and introduced a specific sample of analysis. Spoken corpora for Japanese-speaking learners of English remain scarce, and many of their speech features are still unexplored. By conducting analyses based on the KISTEC, we expect the accumulation of foundational knowledge. Furthermore, we anticipated that KISTEC could be applied for improving teaching methods and curricula in English education, developing instructional materials such as speaking practice tools, and conducting development and validation studies of English speaking tests.

The KISTEC also has limitations. The first limitation is that the target learners are limited to first-year students at KIT, resulting in restricted variation in proficiency levels. Despite some variance in KIT Speaking Test and TOEIC L&R scores, the English speaking abilities of most students are believed to fall within CEFR Levels from A2 to B1. Therefore, the KISTEC is insufficient for investigating the speaking features of learners with particularly high proficiency (CEFR B2 and above). The second limitation is that the KISTEC is based on audio recordings from English speaking test in a monologue format. As a result, it may exhibit characteristics that differ from natural speech. Speaking tests are conducted under conditions of relatively high psychological burden and limited response time, making them context-specific forms of speech. In addition, speech features differ between monologues and dialogues (e.g., the use of fillers). Thus, while it is appropriate to analyze the KISTEC as a monologue-format speaking test, caution is warranted when it is used as a proxy for natural speech.

To address these limitations, using the KISTEC in conjunction with other corpora that include speech from highly proficient and less proficient English learners and natural speech is effective. Related to this point, we are constructing a contrastive corpus for the KISTEC that includes data from NSs and learners with high English proficiency. Once completed, this corpus will enable comprehensive analyses of English speech features across a wider range of proficiency levels.

Acknowledgments

This research was supported by JSPS Grants-in-Aid for Scientific Research (22K00736, 19K00849). We are grateful to all those involved in the development and administration of the KIT Speaking Test and to the annotators who performed the transcription and annotation of the KISTEC.

References

Mariko Abe, and Yusuke Kondo (2019). “Constructing a Longitudinal Learner Corpus to

- Track L2 Spoken English.” *Journal of Modern Languages*, 29:1, pp. 23–44.
- Mariko Abe, Yuichiro Kobayashi, and Yusuke Kondo (2024). “Capturing Chronological Variation in L2 Speech through Lexical Measurements and Regression Analysis.” *Applied Corpus Linguistics*, 4:3, p. 100105.
- Mariko Abe (n.d.). *LOCSE Reserach Project*. <https://sites.google.com/view/locse/>
- Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan (1999). *Grammar of Spoken and Written English*. Longman.
- Jasone Cenoz (1998). *Pauses and Communication Strategies in Second Language Speech*.
- Robert Eklund (2004). “Disfluency in Swedish Human–Human and Human–Machine Travel Booking Dialogues.” Doctoral dissertation, Linköping University. Linköping University Electronic Press.
- Rod Ellis (2003). *Task-Based Language Learning and Teaching*. Oxford University Press.
- Shin’ichiro Ishikawa (2023). *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners’ L2 English*. Routledge.
- Shin’ichiro Ishikawa (n.d.). *The International Corpus Network of Asian Learners of English*. <https://language.sakura.ne.jp/icnale/>
- Noriko Iwashita, Annie Brown, Tim McNamara, and Susan O’Hagan (2008). “Assessed Levels of Second Language Speaking Proficiency: How Distinct?” *Applied Linguistics*, 29:1, pp. 24–49.
- Emi Izumi, Kiyotaka Uchimoto, and Hitoshi Isahara (2004). *Nihonjin 1200-Nin No Eigo Supikingu Kōpasu* [English Speaking Corpus of 1,200 Japanese Learners]. ALC.
- Katsunori Kanzawa, and Yumi Hato (2021). “CBT Supikingu Tesuto No Butaiura, Doko Ga Dō Muzukashii No Ka? KIT Speaking Test No Jissen Yori [Behind the Scenes of a Newly-Developed CBT Speaking Test: What Is Difficult and Why Is It Difficult?].” *JACET Kansai Journal*, 23, pp. 96–120.
- Katsunori Kanzawa, Yuichiro Kobayashi, Jaeho Lee, Haruhiko Mitsunaga, Masayuki Mori, Yusuke Tanaka, and Taishi Chika (n.d.). *The KIT Speaking Test Corpus*. <https://kitstcorpus.jp>
- Katsunori Kanzawa, Ayaka Setoguchi, Yusuke Tanaka, Taishi Chika, Yuichiro Kobayashi, Haruhiko Mitsunaga, Masayuki Mori, and Jaeho Lee (2025). “The KISTEC: Nihon No Daigakusei No Hatsuwa Dēta Ni Motozuku Eigo Gakushūsha Hanashi Kotoba Kōpasu No Kōchiku [The KISTEC: Constructing a Spoken Learner Corpus of English Based on Speech Data from Japanese University Students].” *Proceedings of the Thirty-First Annual Meeting of the Association for Natural Language Processing*, pp. 83–88.
- Yuichiro Kobayashi, Mariko Abe, and Yusuke Kondo (2022). “Exploring L2 Spoken Developmental Measures: Which Linguistic Features Can Predict the Number of Words?” *English Corpus Studies*, 29, pp. 1–18.
- Ludivine Kosmala, and Aliyah Morgenstern (2019). “Should ‘Uh’ and ‘Um’ Be Catego-

- rized as Markers of Disfluency? The Use of Fillers in a Challenging Conversational Context.” L. Degand, G. Gilquin, L. Meurant, and A. C. Simon (Eds.), *Fluency and Disfluency across Languages and Language Varieties*. Presses universitaires de Louvain. pp. 67–90.
- Robin J. Lickley (2015). “Fluency and Disfluency.” M. A. Redford (Ed.), *The Handbook of Speech Production*. Wiley-Blackwell. pp. 445–474.
- National Institute of Information and Communications Technology (n.d.a). *Nihonjin 1200-Nin Ni Yoru Eigo Kōpasu: The NICT JLE (Japanese Learner English) Corpus* [The NICT JLE (Japanese Learner English) Corpus: An English Corpus Produced by 1,200 Japanese Learners]. https://alaginrc.nict_jle/index.html
- National Institute of Information and Communications Technology (n.d.b). *The NICT JLE Corpus Kakiokoshi Kihon Tagu Fuyo Gaidorain, Ver.2.1.3* [The NICT JLE Corpus Transcription and Basic Tagging Guidelines, Ver. 2.1.3]. Retrieved December 2024, from https://alaginrc.nict.go.jp/nict_jle/src/readme_transcription.pdf
- Yannick Préfontaine, and Judit Kormos (2016). “A Qualitative Analysis of Perceptions of Fluency in Second Language French.” *International Review of Applied Linguistics in Language Teaching*, 54:2, pp. 151–169.
- Caroline L. Rieger (2003). “Disfluencies and Hesitation Strategies in Oral L2 Tests.” Robert Eklund (Ed.), *Gothenburg Papers in Theoretical Linguistics 90*. pp. 41–44.
- Norman Segalowitz (2010). *Cognitive Bases of Second Language Fluency*. Routledge.
- Yusuke Tanaka, and Katsunori Kanzawa (n.d.). *KIT Speaking Test Corpus Kakiokoshi Tagu Fuyo Gaidorain, Ver. 1* [KIT Speaking Test Corpus Transcription and Tagging Guidelines, Ver. 1]. Retrieved December 2024, from <https://kitstcorpus.jp/wp-content/uploads/2022/04/manual-1.pdf>
- Yusuke Tanaka, Ayaka Setoguchi, Taishi Chika, and Katsunori Kanzawa (2023). “Analysis of Fillers Used by English Learners Whose Native Language Is Japanese: The Relationship with Proficiency and Comparison with Native English Speakers.” *Proceedings of the JAECs Conference 2023*, pp. 13–18.
- Yusuke Tanaka, Ayaka Setoguchi, Taishi Chika, and Katsunori Kanzawa (n.d.). *Search Interface for the KIT Speaking Test Corpus*. <https://www.kistecsearch.org/>
- Yusuke Tanaka, Ayaka Setoguchi, Taishi Chika, and Katsunori Kanzawa (to appear). “Moderate Use of Fillers Can Enhance Fluency Assessments.” B. Lacy, M. Swanson, and P. Lege (Eds.), *Moving JALT into the Future: Opportunity, Diversity, and Excellence*. JALT.
- Gunnel Tottie (2014). “On the Use of *Uh* and *Um* in American English.” *Functions of Language*, 21:1, pp. 6–29.
- Shigeru Yoshijima, and Rie Ohashi (2014). *Gaikokugo No Gakushū, Kyōju, Hyōka No Tame No Yōroppa Kyōtsū Sanshō Waku Tsuihoban* [Common European Framework

of Reference for Languages: Learning, Teaching, Assessment. Companion Volume].
Asahi Press.