Moderate Use of Fillers Can Enhance Fluency Assessments

Yusuke Tanaka
Kumamoto Gakuen University
Ayaka Setoguchi
Meiwa Elementary School, Miyakonojo
Taishi Chika
Kyoto University
Katsunori Kanzawa
Kyoto Institute of Technology

Reference Data:

Tanaka, Y., Setoguchi, A., Chika, T., & Kanzawa, K. (2025). Moderate use of fillers can enhance fluency assessments. In B. Lacy, R. P. Lege, & M. Swanson (Eds.), *Moving JALT Into the Future: Opportunity, Diversity, and Excellence*. JALT. https://doi.org/10.37546/JALTPCP2024-12

There is a perspective that categorizes the use of fillers by language learners as either an indicator of fluency or disfluency. The researchers explored how learners' filler usage affects assessments of Task Delivery (fluency evaluation) and Task Achievement using the KIT Speaking Test Corpus. The findings showed that task achievement evaluations consistently declined as filler usage increased. However, the effect on task delivery was more nuanced. Specifically, when the frequency of filler use was up to 4.1 per 100 words, an increase in usage correspondingly enhanced task delivery evaluations. Beyond this threshold, however, evaluations started to decline, and when exceeding approximately 12 fillers per 100 words, the impact on fluency assessments turned negative. These results indicate that categorizing fillers solely as indicators of fluency or disfluency is overly simplistic. While excessive use of fillers may hinder fluency, moderate use can actually enhance it.

学習者によるフィラーの使用を流暢性の指標とする立場と非流暢性の指標とする立場が存在する。本研究ではKIT Speaking Test Corpusを用いて，学習者が使用するフィラーがタスク達成度とタスク伝達度（＝流暢性）の評価に与える影響を分析した。その結果，タスク達成度の評価はフィラーの使用頻度の増加とともに低下することが確認された。一方，タスク伝達度への影響はより複雑で，100語あたりのフィラー使用頻度が4.1語までであればフィラーの増加とともに評価に与える影響は向上するが，これを超えると評価に対する影響は減少に転じ，約12語を超えると評価にネガティブな影響を与え始めることが明らかになった。この結果は，フィラーの使用が流暢性の指標であるか非流暢性の指標であるかという二元的な見方に問題がある可能性を示している。すなわち，フィラーは過剰に使用すれば流暢性を損なう一方で，適度に用いられた場合には流暢性を高める効果があると考えられる。

In current Japanese English education, the focus is on improving speaking and communication skills (MEXT, 2018). Therefore, identifying the factors that affect assessments of speaking skills is crucial. One effective method for assessing the speaking abilities of language learners involves evaluating fluency (Ellis, 2003). Disfluencies, phenomena that negatively impact fluency, play a crucial role in these assessments. Examples of disfluencies include repetitions, self-corrections, pauses, and fillers (Biber et al., 1999). This study specifically focuses on fillers, which are defined as "vocalized hesitation, including sounds as (English) *eh*, *uh*, *uhm*, *er* and the like" (Eklund, 2004, p. 207).

The focus on fillers stems from their controversial nature. Although generally classified as a type of disfluency, several studies have demonstrated that fillers are used more frequently by fluent learners (Cenoz, 1998; Iwashita et al., 2008; Kosmala & Morgenstern, 2019; Préfontaine & Kormos, 2016; Rieger, 2003). Cenoz (1998), for example, reported a higher frequency of fillers among more advanced L2 learners in a study involving 15 Spanish (L1) undergraduate learners of English (L2) in Spain, suggesting that fillers can indicate fluency.

In contrast, other studies have linked fillers to disfluency (de Jong, 2016; Götz, 2019; Kahng, 2014; Kosmala & Crible, 2021; Lennon, 1990; Tanaka et al., 2023; Williams & Korko, 2019). For example, Lennon (1990) recruited four German (L1) students visiting

the UK for a six-month stay. An analysis of their English (L2) speaking data at the beginning and end of their stay revealed a reduction in the frequency of fillers, which Lennon interpreted as evidence of improved fluency. These studies suggested that fillers may be seen as indicators of disfluency.

Given the conflicting views, there is no consensus on whether fillers indicate fluency or disfluency. However, the binary classification of fillers as strictly signs of either fluency or disfluency may be overly simplistic. Tottie (2014) advocated a balanced approach, particularly focusing on the two common fillers *uh* and *um*. Tottie suggested that these fillers, unless overused or used in the wrong context, can facilitate comprehension by allowing listeners additional time to absorb the speaker's message, thereby enhancing fluency. This advantage of disfluency was demonstrated by the faster comprehension of corrected words in utterances like *Move to the yel- uh, orange square*, where *uh* aided in processing the speaker's intended message (Brennan & Schober, 2001). Similarly, Lickley (2015) argued that achieving levels of disfluency comparable to those of native speakers is necessary for learners to reach proficiency akin to native speakers. Based on these insights, it was hypothesized that fillers can enhance fluency when used at rates similar to those of native speakers, although excessive use may impair fluency. This nuanced understanding may explain the varied impact of fillers on fluency assessments across different studies.

To the best of our knowledge, however, no empirical study has yet demonstrated with experimental data or quantitative analyses that moderate use of fillers enhances the fluency of L2 learners. Furthermore, the point at which filler usage transitions from being perceived as a marker of fluency to an indicator of disfluency remains unclear. This study sought to address these gaps by analyzing a learner corpus of spoken English, as detailed in the next section.

Before proceeding, it is essential to define the specific type of fluency under investigation in this study. According to Segalowitz (2010), the construct of fluency comprises three distinct types: cognitive fluency, utterance fluency, and perceived fluency. Cognitive fluency involves the efficient coordination and integration of cognitive processes essential for speech production. Utterance fluency refers to the observable characteristics of speech, including timing, pauses, hesitations, and repairs. Perceived fluency relates to the judgments listeners make based on their perception of these utterance characteristics. Our analysis focused on fluency assessments conducted by both native and non-native English-speaking raters, thus specifically focusing on perceived fluency.

Methodology

KIT Speaking Test Corpus

This study utilized the KIT Speaking Test Corpus (KISTEC) (Kanzawa et al., n.d.) to analyze the use of fillers by Japanese learners of English. The corpus includes transcriptions from audio responses to the KIT Speaking Test, a computer-based English speaking test developed and conducted by the faculty at the Kyoto Institute of Technology for first-year undergraduates in 2018. The test comprises three versions, each containing nine questions, such as: "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" (Kanzawa et al., n.d.). The response time allotted is either 45 or 60 seconds, depending on the question. The responses are in the form of monologues.

An example response to this question is provided below (Example 1). KISTEC utilizes transcription tags to denote specific speech characteristics. For instance, the tag <F></F> indicates a filler, <SC></SC> denotes self-corrections, and <R></R> signifies repeated words.

Example 1:

> I want you go to England because I like <F>um</F> European, and I want to eat England fish and chips. I think I go airplane. <F>Uh</F>, <SC>airplane is pi</SC> airplane's <?>price</?> is very high, but airplane is <??></??> <?>aren't</?> good experience. I think England English is <R>many</R> many <?>lucky</?> experience. (Kanzawa et al., n.d.)

Although several corpora of spoken English by Japanese learners exist, such as the NICT Japanese Learners of English (JLE) Corpus (Izumi et al., 2004) and the Longitudinal Corpus of Spoken English (Abe & Kondo, 2019), KISTEC was chosen for two main reasons. Firstly, the presence of filler tags in KISTEC aids in the straightforward identification and extraction of fillers. Secondly, KISTEC assigns scores to each examinee for each question, categorized into Task Achievement (TA) and Task Delivery (TD), both graded on a scale from 0 to 5. This scoring system allows for detailed analysis of the relationship between filler usage rates and fluency assessments, as reflected in the TD scores. The rating scales for these scores are detailed in Table 1.

Table 1

The Rating Scales of Task Achievement and Task Delivery

Score	Task Achievement	Task Delivery
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 unconnected to the topic).	The speaker does not start the task (e.g. s/he is silent, utters only fillers, or just says, 'I don't know').

Each monologue was evaluated by a pair of trained raters, consisting of one native English speaker and one non-native English speaker (a Filipino English instructor). The raters differed for each question, with a total of 18 individuals (9 native speakers and 9 non-native speakers) participating in the assessments. If the two raters' scores differed by one point, the average was taken, resulting in possible scores ranging from 0 to 5 in 0.5-point increments (11 levels in total). If the difference between their scores exceeded one point, a senior rater (a native English speaker proficient in rating the KIT Speaking Test) re-assessed the response. For instance, the earlier response (Example 1) received a TD score of 3 but a TA score of 1. This evaluation reflected the task requirement, which specifically instructed the examinee to recommend a location within their own country for friends from abroad to visit and explain why. However, suggesting travel to England does not meet this criterion, resulting in a low TA score.

Procedures

The primary objective of this study was to explore the relationship between filler usage and fluency assessments, as reflected in TD scores. Additionally, the relationship between filler usage and TA scores was examined for comparative purposes.

KISTEC includes information on examinees' nationalities. For this 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 nine questions, providing a total of 5,022 responses for analysis. One response was excluded due to the absence of spoken words, resulting in a final dataset of 5,021 responses comprising 273,158 words, of which 26,174 are fillers.

Initially, the analysis involved calculating the filler usage rates for these 5,021 responses by dividing the number of fillers in each response by the total number of words. To examine the relationship between these filler usage rates (ranging from 0 to 1) and speaking test scores (TA or TD, graded from 0 to 5 in 0.5 increments), generalized additive model (GAM) analyses were employed. We chose GAMs, facilitated by R (R Core Team, 2024) and the mgcv package (Wood, 2011), because they allow for the modeling of non-linear relationships between dependent and independent variables, thus enabling us to capture the complex patterns we hypothesized. In these analyses, the filler usage rate served as the independent variable, while the speaking test scores (TA and TD) were the dependent variables.

Hypothesis

To assess whether moderate use of fillers can positively affect fluency evaluations, it was critical to define what *moderate* filler usage entails. This study suggests that such usage should align with the average filler rates observed among native speakers, which show significant variability in prior research by Bortfeld et al. (2001), Shriberg (1994), and Tanaka et al. (2023). Given that average filler usage rates can vary under different conditions, the findings from Tanaka et al. (2023) are especially pertinent. This study analyzed data from 38 native English speakers who participated in the KIT Speaking

Test, revealing an average filler usage rate of 3.47%. This rate is used as a benchmark, suggesting that filler usage up to approximately 3.5% could enhance fluency assessments without negative effects. However, exceeding this threshold may lead to perceptions of filler usage shifting from enhancing fluency to indicating disfluency.

In contrast, TA scores are expected to decline steadily as filler usage rates increase. This is because minimizing filler usage maximizes the amount of substantive content that can be conveyed, which is crucial for addressing the task requirements. Given the limited response times of 45 or 60 seconds per question, even a small increase in filler usage reduces the opportunity to provide meaningful information. Consequently, any increase in filler usage is likely to result in lower TA scores.

Results

The results from the GAM analyses for TA and TD are detailed in Tables 2 and 3, respectively. The data showed that the rate of filler usage significantly affects the assessments for both TA and TD ($p < .01$).

Table 2

Results of the GAM Analysis for TA

	Estimate	Std. Error	t-value	*p*-value
(Intercept)	2.865	0.011	258.6	< 2e-16
	Estimated df	Reference df	F-value	*p*-value
Filler usage rates	3.554	4.439	62.62	< 2e-16

Table 3

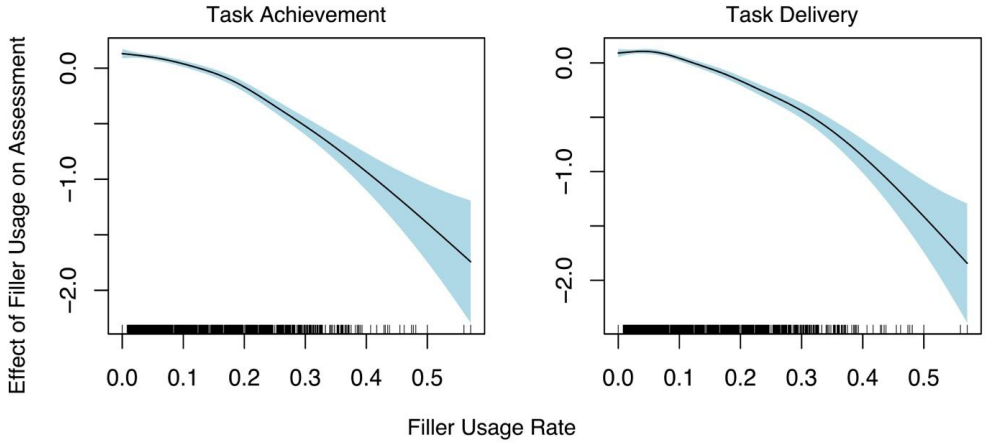
Results of the GAM Analysis for TD

	Estimate	Std. Error	t-value	*p*-value
(Intercept)	2.890	0.009	313	< 2e-16
	Estimated df	Reference df	F-value	*p*-value
Filler usage rates	4.438	5.445	60.1	< 2e-16

The results of the analyses are illustrated in Figure 1, which demonstrated that the assessments of TA and TD generally decrease as filler usage rates increase.

Figure 1

The Relationships Between Filler Usage Rates and Assessments (TA/TD)



To further examine the relationships between filler usage rates and the assessments of TA and TD, differential coefficients were calculated for every 5% increase in filler usage rates, which are shown in Table 4. Positive coefficients indicated a beneficial impact on fluency assessments, while negative coefficients suggested detrimental effects. The magnitude of these coefficients quantified the extent of their impact.

Table 4

Differential Coefficients by Filler Usage Rate

	Differential Coefficient	
Filler Usage Rate	TA	TD
0%	0.131	0.091
5%	0.098	0.105
10%	0.039	0.043
15%	−0.043	−0.052
20%	−0.169	−0.164
25%	−0.341	−0.296
30%	−0.522	−0.439
35%	−0.717	−0.626
40%	−0.932	−0.857
45%	−1.159	−1.124
50%	−1.396	−1.414

The analysis revealed a non-linear relationship between filler usage rates and their impact on both TA and TD. For TA, a positive effect was observed at low rates (0%–10%), which diminished and became negative between 10% and 15%. Beyond 15%, the negative impact sharply intensified, especially at higher rates (30%–50%). This pattern indicated that as filler usage rates increase, the impact on TA scores consistently declines, intensifying as the usage rate increases.

Similarly, the impact of filler usage rates on TD also exhibited a non-linear relationship, though more complex in nature. At lower rates (0%–10%), TD demonstrated a slight positive effect, with the peak impact occurring at 5% (0.105). To precisely determine the filler usage rate where the effect on TD scores is maximized, a local maximum was calculated. This analysis revealed that the optimal filler usage rate for TD is 4.1%, with a maximum coefficient of 0.107. However, the positive effects diminished rapidly beyond a 15% usage rate, with coefficients declining steadily to −1.414 at 50%.

The turning points where the smoothed effect shifts from positive to negative were calculated to be approximately 12.28% for TD and 12.56% for TA. These findings suggested that filler usage rates below approximately 12.5% may have a relatively positive impact on assessments, while rates at or above this threshold are associated with negative outcomes for both TA and TD.

Discussion

The results confirmed the hypothesis that moderate filler usage, up to approximately 3.5%, does not negatively impact fluency assessments and may, in fact, enhance them. For TD, a slight improvement was observed with filler usage rates up to 4.1%, peaking at this percentage. These findings aligned with Tottie's (2014) argument that fillers aid comprehension when not overused, and with Lickley's (2015) suggestion that achieving native-like levels of disfluency signified high proficiency in a second language. The reason moderate use of fillers enhances perceived fluency may lie in the nature of learners' responses. Although these responses are typically simple in terms of grammar and vocabulary, suggesting that they might not require extra processing time provided by fillers, this study indicated that the additional time is actually crucial. Learners' responses often contain grammatical errors and incorrect vocabulary, complicating comprehension. Thus, fillers provide listeners with the necessary time to process responses that are laden with errors, thereby aiding understanding and enhancing the overall fluency impression.

Our findings also corroborate the hypothesis that excessive filler usage detracts from fluency. The impact of fillers on fluency assessments shifted from positive to negative when usage exceeds a critical threshold of approximately 12.5%. Beyond this point, fillers negatively affected perceptions of fluency. These observations challenge the simplistic view of fillers as mere markers of disfluency, illustrating instead that fillers play a complex role in spoken language. At optimal levels, fillers can enhance fluency, but their overuse can compromise fluency.

Another key finding from this study is that the negative impacts of fillers on fluency assessments become more pronounced as filler usage rates increase. This could be attributed to raters becoming increasingly sensitive to fillers as they encounter them more frequently, which exacerbated their perceived impact.

It is essential to determine whether the relationship between filler usage rates and fluency assessments is causal or merely correlational. The results might suggest that examinees who achieve high scores use fewer fillers due to their high proficiency,

indicating no causal relationship between filler usage rates and fluency assessments. However, if filler usage were merely correlated with high proficiency among test takers, explaining why it peaks at 4.1% would be challenging. A plausible explanation is that this rate is moderate, closely aligning with the average filler usage rate of native speakers (3.47%), thus supporting a causal relationship. However, it is difficult to prove that this relationship is not merely correlational but causal by using a corpus analysis alone. Conducting an experimental study where only the filler usage rates are manipulated in responses would be beneficial to examine whether responses with a moderate filler usage rate are evaluated most favorably, based on this study's findings.

Furthermore, it is important to emphasize that the concept of fluency in this study specifically refers to perceived fluency, as previously discussed. There is no guarantee that other types of fluency, such as cognitive and utterance fluency, are similarly affected by moderate use of fillers. It is possible that moderate use of fillers could enhance the efficiency of cognitive processes involved in speech production by providing speakers with additional time to organize their thoughts, potentially improving cognitive fluency. However, this study does not provide evidence to confirm such effects.

Regarding TA scores, the hypothesis was similarly validated. Unlike TD scores, TA scores consistently decreased as filler usage increased, starting at lower usage rates than those observed for TD. This trend suggested that fillers may dilute the delivery of substantive content, crucial for task achievement. In timed speaking tests, fillers occupy time that could otherwise be used to convey meaningful information, resulting in reduced TA scores. Therefore, although moderate filler use may be permissible in fluency assessments, it must be judiciously controlled in speaking tests where examinees are required to answer questions within a specified time frame.

These findings might also have educational implications. Among the 5,021 responses analyzed, 656 (13.1%) contained no fillers. Given that even native speakers typically use fillers to some extent, this suggests that some examinees might intentionally avoid using fillers, likely to prevent giving the impression of disfluency. Therefore, educators should not discourage the use of fillers entirely, unless in timed speaking tests where succinct and efficient communication is crucial. Knowing that fillers do not negatively impact fluency assessments unless overused can allow learners to speak in a more relaxed manner, particularly for those who are too afraid of using fillers.

Conclusion

This study focused on the relationship between filler usage and perceived fluency among Japanese learners of English using the KISTEC. The findings highlighted a nuanced role for fillers in fluency assessments, supporting the hypothesis that moderate filler usage enhanced fluency, while excessive use undermined it. Specifically, the beneficial effects of fillers on fluency evaluations grow with their use up to a threshold of 4.1%. However, beyond this point, the positive impact diminishes, and filler use becomes detrimental when it exceeds 12.5%. These results challenged the binary view of fillers as either indicators of fluency or disfluency. Instead, they illustrated that judicious use of fillers can emulate native-like disfluency patterns, thus enhancing fluency by providing listeners with additional processing time, while overuse negatively impacts both fluency and overall task performance. These findings offered practical implications for educators, emphasizing the need to reduce excessive filler use while teaching learners to use fillers moderately to improve fluency and listener comprehension.

Despite these insights, this study has several limitations. First, the data were exclusively drawn from KISTEC, which focuses solely on Japanese learners of English. Moreover, the English proficiency levels in KISTEC are relatively limited, likely ranging from A1 to B2 on the Common European Framework of Reference for Languages (CEFR) scale (Council of Europe, 2001). These factors limit the generalizability of the findings to other linguistic and cultural contexts, as well as to learners with different proficiency levels. Second, the analysis primarily examined filler frequency, without considering other important factors such as filler duration (short vs. long), types (e.g., vowel fillers like *uh* vs. lexical fillers like *well*), or placement (e.g., clause boundaries vs. within clauses). Notably, the influence of fillers on fluency assessments varies depending on their placement (Chika et al., 2024). Third, the study analyzed responses to an English speaking test, which are monologues. This raises questions about the applicability of the findings to dialogues, where fillers serve additional functions, such as maintaining or yielding the floor (Williams, 2023)—functions irrelevant in monologues. Consequently, it remains unclear whether moderate filler use similarly enhances fluency assessments in conversational contexts.

Future research should address these limitations. Expanding the analysis to include learners from diverse linguistic and cultural backgrounds, as well as more proficient learners (C1 and C2 on the CEFR scale), would enhance the generalizability of the findings and provide a broader understanding of filler usage in second language acquisition. In addition, examining the interplay of filler duration, type, and placement could uncover how these factors collectively influence perceived fluency. Finally,

investigating the relationship between fillers and other dimensions of fluency, such as cognitive and utterance fluency, and assessing whether similar results are observed in dialogues would contribute to a more holistic understanding of fillers' roles in fluency assessments. Such research could inform more effective teaching strategies and assessment frameworks, ultimately helping learners become more fluent L2 speakers.

Notes

1. An internationally recognized guideline used to classify foreign language skills. It describes and assesses language proficiency across six levels, ranging from A1 for beginners to C2 for highly proficient speakers.

Bio Data

Yusuke Tanaka is a lecturer at Kumamoto Gakuen University. His research areas include language comprehension, corpus linguistics, cognitive linguistics, and English education. He is involved in the project to develop the KIT Speaking Test Corpus. Using this resource, he is researching the impact of filler usage on fluency assessments among Japanese learners of English. <yusuke.tanaka.07@gmail.com>

Ayaka Setoguchi is an elementary school teacher in Miyazaki Prefecture, Japan. Her research interests include cognitive linguistics, second language acquisition, and L2 education. Recently, she has been exploring the effects of teaching methods applying cognitive linguistic theory on grammar instruction. <guchi.a.chan7@gmail.com>

Taishi Chika is a part-time lecturer at Kyoto University in Japan. His current research interests include theoretical linguistics (word-formation theory and lexical semantics). He is also a PhD candidate at the Graduate School of Human and Environmental Studies, Kyoto University. <nanou7614@gmail.com>

Katsunori Kanzawa is an associate professor at the Faculty of Arts and Sciences, Kyoto Institute of Technology (KIT), Japan. His research interests include language testing, corpus linguistics, and cognitive linguistics. He has been a project leader of the KIT Speaking Test since 2022. He has also built a corpus with his colleagues based on the audio responses of the KIT Speaking Test, which is called the KIT Speaking Test Corpus (KISTEC). <kanzawa@kit.ac.jp>

References

Abe, M., & Kondo, Y. (2019). Constructing a longitudinal learner corpus to track L2 spoken English. *Journal of Modern Languages*, *29*, 23–44. https://doi.org/10.22452/jml.vol29no1.2

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Grammar of spoken and written English*. Longman.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, *44*, 123–147. https://doi.org/10.1177/00238309010440020101

Brennan, S. E., & Schober, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, *44*, 274–296. https://doi.org/10.1006/jmla.2000.2753

Cenoz, J. (1998). *Pauses and communication strategies in second language speech*.

Chika, T., Setoguchi, A., Tanaka, Y., & Kanzawa, K. (2024). 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 [Analyzing disfluency in English learners' speech: The relationship of self-correction with repetition and fillers]. *Proceedings of the 30th Annual Meeting of the Association for Natural Language Processing*, 3166–3170.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

de Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*, 113–132. https://doi.org/10.1515/iral-2016-9993

Eklund, R. (2004). Disfluency in Swedish human–human and human–machine travel booking dialogues (Doctoral dissertation, Linköping University Electronic Press).

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Götz, S. (2019). Filled pauses across proficiency levels, L1s and learning context variables: A multivariate exploration of the Trinity Lancaster Corpus Sample. *International Journal of Learner Corpus Research*, *5*, 159–180. https://doi.org/10.1075/ijlcr.17018.got

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*, 24–49. https://doi.org/10.1093/applin/amm017

Izumi, E., Uchimoto, K., & Isahara, H. (2004). 日本人1200人の英語スピーキングコーパス [Corpus of spoken English by 1,200 Japanese individuals]. ALC.

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, *64*, 809–854. https://doi.org/10.1111/lang.12084

Kanzawa, K., Kobayashi, Y., Lee, J., Mitsunaga, H., Mori, M., Tanaka, Y., & Chika, T. (n.d.). *The KIT Speaking Test Corpus (KISTEC)*. https://kitstcorpus.jp

Kosmala, L., & Crible, L. (2021). The dual status of filled pauses: Evidence from genre, proficiency and co-occurrence. *Language and Speech*, *65*, 216–239. https://doi.org/10.1177/00238309211010862

Kosmala, L., & Morgenstern, A. (2019). Should 'uh' and 'um' be categorized as markers of disfluency? The use of fillers in a challenging conversational context. In L. Degand, G. Gilquin, L. Meurant, & A. C. Simon (Eds.), *Fluency and disfluency across languages and language varieties* (pp. 67–90). Presses universitaires de Louvain.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*, 387–417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Lickley, R. J. (2015). Fluency and disfluency. In M. A. Redford (Ed.), *The handbook of speech production* (pp. 445–474). Wiley-Blackwell.

MEXT (2018). *The course of study*. Retrieved January 12, 2025, from https://www.mext.go.jp/content/20230120-mxt_kyoiku02-100002604_03.pdf.

Préfontaine, Y., & Kormos, J. (2016). A qualitative analysis of perceptions of fluency in second language French. *International Review of Applied Linguistics in Language Teaching*, *54*, 151–169.

R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rieger, C. L. (2003). Disfluencies and hesitation strategies in oral L2 tests. In R. Eklund (Ed.), *Gothenburg papers in theoretical linguistics 90*, pp. 41–44.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Shriberg, E. (1994). Preliminaries to a theory of speech disfluencies (Doctoral dissertation, University of California, Berkeley).

Tanaka, Y., Setoguchi, A., Chika, T., & Kanzawa, K. (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*, 13–18.

Tottie, G. (2014). On the use of *uh* and *um* in American English. *Functions of Language*, *21*, 6–29. https://doi.org/10.1075/fol.21.1.02tot

Williams, S. A. (2023). *Disfluency and proficiency in second language speech production*. Palgrave Macmillan. https://doi.org/10.1007/978-3-031-12488-4

Williams, S. A., & Korko, M. (2019). Pause behavior within reformulations and the proficiency level of second language learners of English. *Applied Psycholinguistics*, *40*, 723–742. https://doi.org/10.1017/S0142716418000802

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *73*, 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x