Educational Journal of Learning Technology
Vol. 4, No. 2, pp. 414-426, January 2027
https://10.58230/edutech.v4i1.148
Received October 19, 2026; Revised November 07, 2026; Accepted November 09, 2026

Who Does the Interactional Work? Rethinking L2 Interactional Competence with AI Interlocutors: An Integrative Review

FX. Risang Baskara[1], Anindita Dewangga Puri[2]
[1,2] Universitas Sanata Dharma
Jl. STM Pembangunan, Sleman Regency, Special Region of Yogyakarta, Indonesia.
[1]risangbaskara@usd.ac.id

Abstract. Second language (L2) learners increasingly use chatbots as conversation partners, yet how this impacts their interactional competence (IC) remains unclear. This integrative review addresses this gap by examining who performs the interactional work during human-machine conversations. We analyzed 16 empirical studies (2018–2026), coding interlocutor architecture, tasks, IC practices, and the enactors of these practices (learner or system). The 34 practice-level findings revealed 11 instances of transfer, 8 adaptations, 6 displacements, and 9 reductions. Patterns indicate stronger transfer in action sequences like requests and refusals. Relational and listening work also occurs more frequently with machines, and repair strategies diverge by system architecture (constrained systems versus large language models). We propose a framework of interactional labor (transfer, adaptation, displacement, and system upkeep) and discuss implications for EFL speaking tasks and assessment. Notably, none of the included studies were conducted in Indonesia, offering specific implications for Indonesian classrooms. Ultimately, interacting with AI partners rehearses certain action formats but reduces learners' relational and listening work compared to human interlocutors.

Keywords: Interactional Competence; Conversational Agents; Large Language Models; Conversation Analysis; Speaking Assessment.

1. Introduction

Indonesian learners of English have few opportunities to speak English outside the classroom, as do learners in most other EFL contexts. The authors have observed in their own speaking classes that individual learners may have only a few minutes of conversational turns in many speaking classes with around forty students. Voice-enabled chatbots and large language model (LLM) applications such as ChatGPT are now available as partners that are always available and can role-play and correct their own errors (Godwin-Jones, 2024; Kohnke et al., 2023). Meta-analyses report medium positive effects of dialogue-based CALL on proficiency (Bibauw et al., 2022) and speaking (Hou & Min, 2026), and Lyu et al. (2025) report a similar effect for comparative chatbot studies. Indonesian teacher-researchers have begun to use chatbots in flipped courses (Baskara, 2023) and in pragmatic practice (Puri & Baskara, 2023).

Most such tools assess what learners can do after practice (Chen et al., 2025; Huang et al., 2022), but few address how they manage interaction. Interaction between two people is a collaborative activity involving taking turns, fixing problems, opening and closing conversations, and showing interest. Kramsch (1986) called the skill to engage in such interactions interactional competence. Both conversation analysis (CA) and language assessment define IC as interactive

©Authors
This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

practices that participants dynamically construct (Galaczi & Taylor, 2018; Roever & Kasper, 2018). The presence of a machine interlocutor affects what learners need to do and may perform some functions for them.

These questions are pertinent to the Indonesian context. In a study without AI, Putra and Prasetyo (2025) found that project-based speaking with real audiences increased non-English-major students' confidence and fluency. Since none of the 16 studies in this review are in Indonesia, their findings in other countries can be applied to the Indonesian classroom.

Tests that emphasize individual performance and accuracy miss the joint action that creates shared meaning (Kramsch, 1986). Concepts emerged in CA describing these joint actions: turn-taking (Sacks et al., 1974), preference for self-repair (Schegloff et al., 1977), and sequences of actions like requests and refusals. Elements of IC include turn management, topic management, interactive listening, breakdown repair, and nonverbal behavior (Galaczi & Taylor, 2018). Roever and Kasper (2018) propose IC as a criterion for speaking tests. As learners acquire more ways to enact actions, their IC develops (Pekarek Doehler, 2018). IC includes interaction repertoires linked to specific social activities (Hall, 2018). Assessing IC is challenging when test subjects do not interact with the interlocutor (Plow et al., 2018; Lam et al., 2023). Other-initiated repair, where the partner initiates repair, occurs frequently in natural conversation across many languages (Dingemanse et al., 2015; Dingemanse & Enfield, 2024). Without this aspect of IC, learners will miss a crucial part of it if their partner never needs them to repair.

People adapt their talk to machines. People make requests to a voice assistant as part of family activities (Porcheron et al., 2018). When a humanoid robot fails to take turns smoothly, people reduce and simplify their turns (Pelikan & Broth, 2016). Learners act as if they are interacting with a human and also cooperate with an SDS to create common ground (Dombi et al., 2022). LLMs generate fluent turns that respond to the context and guess intentions from incomplete input. Voss and Waring (2025) studied 15 recordings of human-AI interaction and concluded that ChatGPT had not yet reached natural conversation. Godwin-Jones (2024) describes distributed agency, where the tool shares some of the workload. However, if the agency is shared, then so is the interactional work, and how?

To the best of our knowledge, no synthesis codes this type of evidence regarding the direction of interactional work. This is true for the two sources we found (Section 2). We present a critical integrative review focusing on three research questions: (1) What kinds of interactional work are performed by the learner and the system across empirical studies involving L2 learners and automated interlocutors? (RQ1); (2) How do these findings relate to talk between people? (RQ2); (3) What framework accounts for these differences and can guide the design and evaluation of tasks? (RQ3). We introduce a coding scheme and a four-category framework of interactional labor applied to 16 studies published between 2018 and 2026.

2. Method

Tables 1-3 present the protocol. This integrative review (Torraco, 2016) combined the theory-adaptation approach for conceptual articles (Jaakkola, 2020). The review was not registered. The authors agreed on the review questions, eligibility criteria, and a draft coding scheme before the main search, and tested the scheme on three familiar studies.

We searched two sources. The first was OpenAlex, a database of about 250 million scholarly works. We used eight keyword strings based on studies the authors knew. Each keyword string included a publication-year filter and was ranked by relevance. We screened the top 20 results (except S5, S7, and S8, which had 15, 19, and 15 screened, respectively). The second was a full-text

search from Wiley using two full-sentence questions. We also included studies from backward citation chasing and initial reading. Table 1 contains strings that are combinations of keywords but not explicitly Boolean statements. These are listed as run. Searches were performed on [DATE]. We used Consensus and Elicit to identify related papers and compare citation records. The authors made all eligibility and coding decisions. Reference checks were performed against OpenAlex records, which use Crossref metadata.

Table 1. Search protocol

Source	Search string or query (as run)	Years	Estimated matches	Records screened
OpenAlex S1	interactional competence chatbot spoken dialogue system second language	2018-	2,267	20
OpenAlex S2	conversation analysis conversational agent robot language learners repair	2018-	760	20
OpenAlex S3	negotiation of meaning chatbot EFL learners interaction	2018-	886	20
OpenAlex S4	speaking assessment spoken dialogue system virtual interlocutor oral proficiency	2018-	988	20
OpenAlex S5	spoken dialogue system pragmatics requests second language learners Timpe-Laughlin Sydorenko	2016-	38	15
OpenAlex S6	interactional competence construct testing speaking conceptualizations	2014-	26,511	20
OpenAlex S7	interactional competence Galaczi Taylor Roever Kasper Pekarek Doehler Hall	2013-	23	19
OpenAlex S8	large language models conversation analysis turn-taking repair human-AI talk	2022-	3,656	15
Wiley full text Q1	How do second language learners manage interactional practices such as repair, turn-taking, openings and closings when talking with spoken dialogue systems, chatbots, or generative AI interlocutors, and what does this imply for interactional competence?	2018-	11 articles	11
Wiley full text Q2	Do large language model chatbots over-accommodate second language learners, taking over repair, topic management and sequence organization so that learners display less interactional competence than with human partners?	2022-	8 articles	8

Note. Estimated matches are the totals OpenAlex reported. Wiley full text = Wiley's full-text journal collection. Records screened: 20 unless stated in the text.

Publications had to be published in 2016 or later. The search start year varied by search string (Table 1). During screening, we dropped records returned by a search that were from before 2016. The main eligibility criterion was information on the interaction. If a study reported only proficiency or affective outcomes, it was ineligible because these do not reveal who did the interactional work.

Table 2. Eligibility criteria

Criterion	Inclusion	Exclusion
Design	Empirical study (quantitative, qualitative, or mixed)	Reviews, meta-analyses, editorials, and conceptual papers (kept as background)
Participants	L2 learners or L2 test takers of any age	L1 users only; speakers' L2 status

		cannot be verified
Interlocutor	At least one automated interlocutor: scripted simulation, rule- or intent-based SDS, LLM-based system, embodied agent, or robot	Human-human interaction, including video-mediated or VR talk between people
Data	Data on the interaction itself (transcripts, CA or discourse analysis, coded interactional features, IC ratings, or learner reports of specific events)	Outcome-only designs (proficiency, vocabulary, or affect without interaction data); written tasks
Publication	Peer-reviewed article or proceedings paper with a DOI, published 2016 or later, in English	Dissertations, books, chapters, and preprints

Note. SDS = spoken dialogue system; LLM = large language model; CA = conversation analysis; IC = interactional competence.

We used DOIs and titles to import records and remove duplicates. Two authors read titles and abstracts and screened full texts for inclusion, as shown in Table 2. Disagreements were settled by discussion. Figure 1 describes the process. Of 168 records, we removed 13 duplicates, leaving 155. We excluded 120 at the title and abstract level, primarily because they were medical, business, or general AI research. Of the 35 studies examined at full text, we rejected 25. Eleven were reviews and secondary sources (retained as background). Six had only an outcome measure (including Ayedoun et al., 2019, reporting expected willingness to communicate and no coded learner interaction). Four focused on human-human interaction; two involved only L1 users; one was a written task; and one was a study in which the L2 status of the speakers could not be confirmed (Voss & Waring, 2025, cited as background). Ten studies passed all criteria, and six further studies were found through citation chasing and initial reading, for a total of 16 studies.

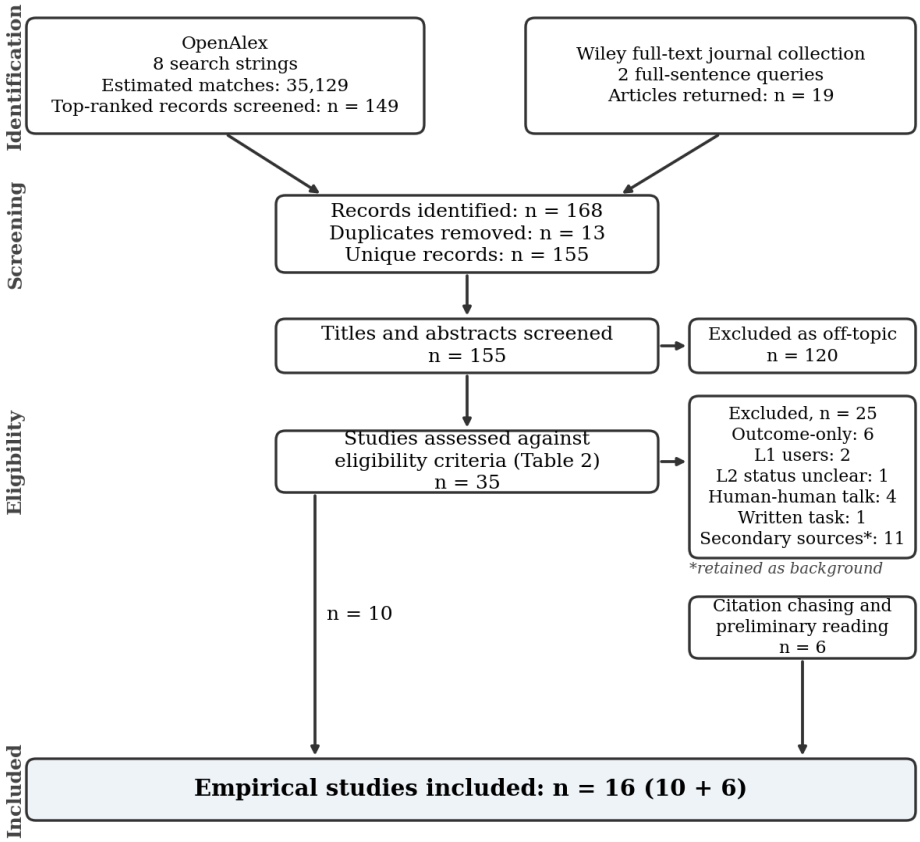


Figure 1. Flow of records through identification, screening, and eligibility

We coded each study on a coding sheet using dimensions A to E (Table 3). Dimensions A to D include information on the architecture, purpose and task, IC practices analyzed, and human comparison. Dimension E is the interactional labor code. One code is given for each reported finding about one practice in one study. If a study reported findings for each system, we coded each system separately, as in Timpe-Laughlin et al. (2025). We used a code other than R if the study indicated who initiated or completed the practice or described a system feature that determined the practice. The first author coded all 16 studies, and the second author independently recoded them. Coders agreed [X]% of the time (Cohen's κ = [X]) across [N] codes, and they discussed disagreements. Agreement for title and abstract screening was [X]% (κ = [X]), and full-text agreement was [X]% (κ = [X]). The authors did not appraise study quality. The unit of analysis is a reported finding about the direction of interactional work, not a pooled effect. One appraisal instrument does not cover different study designs (conversation analysis, quantitative, mixed, and learner reports); the numbers indicate whether evidence exists for each.

Table 3. Coding scheme

Dimension	Codes and decision rules
A. Interlocutor architecture	Scripted simulation; intent- or rule-based SDS; LLM-based system (text or voice); embodied agent or robot
B. Purpose and task	Learning, assessment, or both; role-play, paired discussion, examiner interview, or open conversation
C. IC practices analyzed	TT turn-taking; RP repair; OC openings and closings; TM topic and sequence management, including small talk; RD recipient design; IL interactive listening and backchanneling; PA pragmatic action sequences (e.g., requests, refusals)
D. Human comparison	Human comparison: present or absent (peer, teacher, or examiner partner). Report the direction of any difference in Section 4.
E. Interactional labor	T Transfer: learner-carried practice in a form found in human talk. A Adaptation: learner-carried practice in a form tied to machine limits. D Displacement: system-carried practice the learner would carry with a human partner. U System upkeep: machine-specific work that keeps the system running, recorded once per study. R Reduced, unattributed: reported as less frequent or less developed than with humans, with no evidence of who carried it; outside the framework.

Note. A reported finding is about one practice in one study. T, A, and D codes require evidence of who started or finished the practice or a system property determining it; otherwise R is assigned. U appears only once per study and is not included in the 36 findings. IC: R is a comprehensive IC rating reported lower with the machine. It is not counted in Figure 3.

Section 1 describes these four framework categories. Of the 34 findings, 25 have a code, which can be attributed to (T, A, or D). The nine R findings indicate cases where no one is identified as doing the work. The fourth category (U) covers system maintenance, which is machine-specific work not carried out by learners, nor transferred to talk with people, and is therefore coded at the study level.

3. Results

Corpus profile

Table 4 presents 16 studies (11 published in 2024 or later), including one scripted simulation, one social robot, seven intent- or rule-based SDSs, six LLM systems, and one study comparing an SDS to an LLM. Five studies focused on assessment, two included both learning and assessment, and nine focused on learning. Ten studies involved a human comparison condition. Six studies come from one research group (Sydorenko, Dombi, and Timpe-Laughlin), and two come from

Ockey and colleagues. Dombi, Sydorenko, and Timpe-Laughlin co-authored several of these studies. These studies are not fully independent.

Table 4. Coded corpus (N = 16)

Study	Interlocutor	Purpose and task	Human comp.	Main interactional finding	Labor codes
Sydorenko et al. (2018).	Scripted simulation	Learning; academic requests	No	Advanced learners took forms and content from expert models	PA: T; TM:D
Ockey & Chukharev-Hudilainen (2021).	SDS	Assessment; paired discussion	Yes	Higher IC scores with a human partner; computer partner more reliable	IC: R
Dombi et al. (2022).	SDS	Learning; role-play	No	Egocentric and cooperative behavior toward the system; common ground built	RD: A; TM: T
Ockey et al. (2023).	SDS	Assessment; paired discussion	Yes	More IC features are ratable with the SDS, but most features score lower.	IC: R
Ericsson et al. (2023).	SDS (virtual humans)	Learning; everyday scenarios	No	Practice felt safe but frustrating when not understood	RP: A; U
Timpe-Laughlin et al. (2024).	SDS	Learning and low-stakes assessment; role-play	Yes	Fewer turns with the SDS, fewer rapport moves, clarification requests, backchannels	TT:A; RP:R; IL:R; TM:R
Dombi et al. (2024).	SDS	Learning; role-play	Yes	Openings and closings transactional with SDS, relational with a human	OC: R; U
Karatay & Xu (2025)	LLM-based SDS	Assessment; examiner interview	No	IC features separated proficiency levels; system managed the talk	PA:T; TT:A; IL: R; TM:D
Timpe-Laughlin et al. (2025).	SDS vs LLM	Learning and assessment; service encounter	No	Repair 86% vs 24%; recipient design 78% vs 14% (SDS vs LLM)	RP:A/D; RD:A/D; OC:T/D; TM:T; U
Eguchi et al. (2025).	LLM (multimodal)	Assessment; role-play	Yes	Contingent, co-constructed responses; more overlaps and gaps than with humans	PA:T; TT:A
Sim et al. (2025).	LLM (ChatGPT)	Learning; role-play (requests, refusals)	Yes (peer)	Short-term gains with ChatGPT; better long-term retention with peers	PA:T; PA:D; PA:R
Jakonen et al. (2025).	Social robot	Learning; small-group activities	No	Recognition failures extended repair; teacher used it for pronunciation	RP: A
Meng et al. (2026).	LLM chatbot	Learning; pre-post practice	Yes (peer)	Both groups improved; peer group better at interactive listening	IL: R
Su & Chen (2026).	LLM (ChatGPT-4o)	Assessment; role-play (refusals)	Yes	Refusal sequences scorable; richer engagement with a human interlocutor	PA: T; PA: R

Timpe-Laughlin, Dombi, et al. (2026)	SDS	Learning; workplace requests	Yes	Requests elicited in both modes; SDS is more transactional	PA: T; PA: R
Qu & Chen (2026).	LLM (voice)	Learning; intercultural role-play	Yes (peer)	Politeness, repair, and topic extension co-occurred; lower anxiety	RP: T; TM: T

Note. Codes follow Table 3 (practice: category). IC: R is a holistic IC rating reported as lower with the machine. U marks system upkeep reported for the study as a whole. Timpe-Laughlin et al. (2025) is coded separately for each system. Yes (peer) indicates a human peer partner.

Coded findings

The 16 studies reported 34 practice-level findings: 11 transfer (T), 8 adaptation (A), 6 displacement (D), and 9 reduced, unattributed (R) (Figure 3). Figure 3 contributed seven findings, as its two systems were coded separately (Timpe-Laughlin et al., 2025). Three studies reported system upkeep (U), which was coded at the study level. Thus, it does not contribute to the 36 findings.

The most findings (10) resulted from pragmatic action sequences. Across all four codes, repair accounted for six findings (one T, three A, one D, and one R). In topic and sequence management (6), three findings were coded T, two D, and one R. Two findings for recipient design (3) were coded A and one D. Three findings for openings and closings (3) were coded T, D, and R. All three studies coded turn-taking (3) as A. All three studies coded interactive listening (3) as R. Nine of the 34 findings (26%) were coded R; that is, they reported a practice being less frequent or less developed with a machine than with a human partner but did not report who performed it.

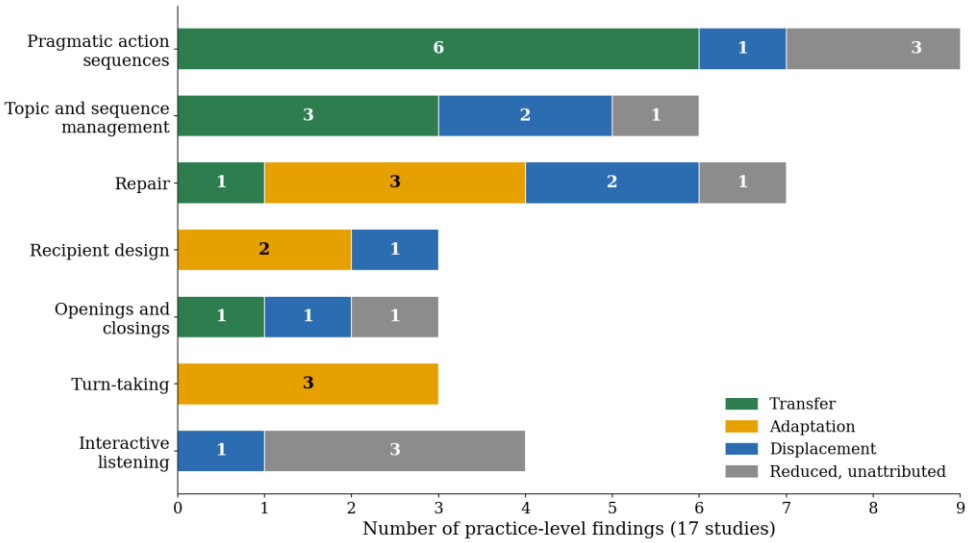


Figure 2. Practice-level findings by IC practice and labor category

4. Discussion

Action formats carry over

Sequence-level actions. The strongest evidence of transfer was for sequence-level actions. An SDS and a human partner both prompted relevant requests in a workplace task (Timpe-Laughlin, Dombi, et al., 2026). Raters could score refusal sequences with ChatGPT-4o, but ratings were more diverse when using a human partner (Su & Chen, 2026). Multimodal role-play with an LLM helped learners see the task as co-constructed and respond contingently, with some preference organization (Eguchi et al., 2025). A service encounter showed closing sequences in 98% of SDS conversations and all LLM conversations. Small talk was similar in both (Timpe-Laughlin et al., 2025). This is where practicing talking with people is most likely to help. There is less confidence

that such transfer will lead to better performance in other contexts. ChatGPT role-play helped with short-term requests, and peer interaction helped with retention over time (Sim et al., 2025).

Relational work and listening are thinner

Machines' weaknesses in affiliation and reciprocity were consistent. Learners took more turns with a human partner. Human interactions included more rapport-building moves, clarification requests, and backchannels (Timpe-Laughlin et al., 2024). Openings and closings were relational with a human interlocutor and transactional with an SDS (Dombi et al., 2024). Chatbot and peer practice both improved IC, yet peers did better on interactive listening (Meng et al., 2026). IC was much higher with a human partner than when interacting with an SDS rated by the rater (Ockey & Chukharev-Hudilainen, 2021). The LLM examiner provided no backchannels for test takers to use (Karatay & Xu, 2025). If the partner does not listen, the learner has nothing to give back. Timpe-Laughlin et al. (2025) found that learners' greetings matched each system's opening line.

Repair may run in opposite directions

Repair is where system design matters most. In constrained systems, learners handle repair due to recognition failures. Lower secondary students thought SDS was easy and safe, but also frustrating if the system did not understand them (Ericsson et al., 2023). Robot-based interactions had long repair cycles due to recognition failures, which the teacher transformed into pronunciation practice (Jakonen et al., 2025). The only study comparing both architectures in one task revealed repair in 86% of SDS conversations and 24% of LLM conversations, and recipient design in 78% and 14% (Timpe-Laughlin et al., 2025). The authors explained the high numbers by noting that the SDS has limited capacity, leaving recognition trouble for the learner to repair, and the low numbers by noting that the LLM inferred participants' intentions, leaving little room for learners to show IC.

The learner did not have to correct the same error in one LLM example. The system took care of the trouble (Timpe-Laughlin et al., 2025). We code the SDS behavior as adaptation because learners repair for the machine's benefit. We code the LLM behavior as displacement because the system takes over work that a human partner would pass back to the learner. Other studies are consistent with the SDS pattern, but the LLM pattern is not general. Qu and Chen (2026) found that learners made polite repairs and extended topics in an LLM intercultural role-play, suggesting that repair with the learner may depend on task type. Repair is foundational to human interaction (Dingemanse & Enfield, 2024), so displacement of repair merits attention.

Timing and turn-taking are adapted

The three turn-taking codes all point to adaptation. The degree of adaptation varies. For example, Timpe-Laughlin et al. (2024) report fewer learner turns with the SDS. Karatay and Xu (2025) show how the system manages the talk, with unexpected interruptions and topic switches it initiated, which might be more like displacement. Eguchi et al. (2025) note more overlaps and gaps than with human partners, which describes the timing of the exchange more than either party's moves. Thus, we consider the turn-taking adaptation code as provisional. Turn-taking in machine talk shows how learners manage system latency, but less how they transition turns with people.

Designed displacement

Some systems carried interactional work. An embodied agent with communication strategies and affective backchannels raised expected willingness to communicate over a version with one feature in a design study outside of the corpus (Ayedoun et al., 2019). It measured affect only, so it is not included in the counts. Advanced learners can use scripted expert request sequences to take the form and content from expert models (Sydorenko et al., 2018). Some show evidence for affective benefits. Ericsson et al. (2023) say practice is safe, and Qu and Chen (2026) report lower speaking anxiety with learner-carried repair and topic extension, so the system does not need to carry the work.

If scaffolding through designed displacement is to be effective, the scaffold must fade away. The only way to know if this hypothesis is true is to see if the learner can take on the displaced work, which no included study did. This means that if a partner does all the listening, closing, and

repair, it might appear that the practice is working even though the learner's repertoire has not changed.

An interactional labor framework

Figure 2. The results are grouped by who carries the practice and whether the practice, as carried, is transferable to talk with people. Transfer is learner-carried practice in forms found in human talk. Adaptation is learner-carried practice whose form is tied to machine limits. Displacement is system-carried practice that a learner would carry with a human partner. System upkeep is machine-specific work that keeps the system running, such as slot-filling questions, which appeared in three studies (Dombi et al., 2024; Ericsson et al., 2023; Timpe-Laughlin et al., 2025).

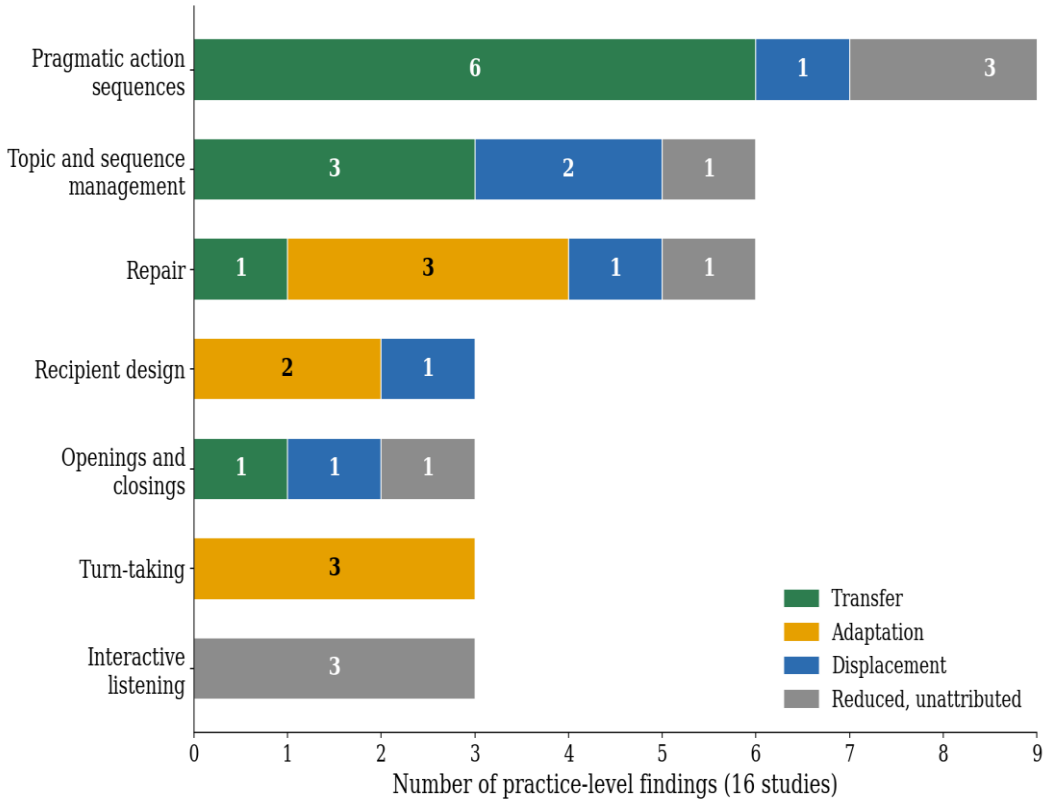


Figure 3. The interactional labor framework

Only repair appears in all four codes, as expected for something dependent on system structure. This work builds on Timpe-Laughlin, Davis, et al. (2026), who suggest human-machine communication as an evolving component of IC. We term this capacity adaptation, and how the learner handles system maintenance. Adaptation shows that learners can perform practices with machines. It is not a form of IC with people but instead represents human-machine competence. Of interest are findings about transfer since these are most relevant to IC with people. No studies reported a delayed post-test with human partners in the coding results.

Implications for speaking assessment

Some studies indicate higher rating consistency with automated partners. Karatay and Xu (2025) find an intraclass correlation of .95 for SDS-elicited speech. This shows consistency but no improvement over human partners, since the study includes no human-partner condition. Ockey and Chukharev-Hudilainen (2021) include a human comparison: ratings with a computer partner are more consistent, although IC scores are much higher with a human partner. Standardized behavior can reduce relational and listening features. Adding architecture can add non-test-taker-related variance. A study on the same task finds that repair occurs in 86% of conversations with one system and 24% with another (Timpe-Laughlin et al., 2025). This means that repair rubric categories mean different things for each system. Dai and Roever (2026) do not qualify as one of the 16 studies because they studied a smartphone role-play test of L1 and L2 Chinese speakers and

reported no automated interlocutor condition. They did note that IC was not consistent across social contexts and digital modalities.

The proposed recommendations, none of which the included studies have tested, are as follows: Report the architecture and model version of any automated interlocutor and its prompt. Claim IC with people from transfer practices only and report adaptation as a distinct variable called human-machine competence. Configure interlocutors to request clarification if they do not understand input, refrain from initiating closings, and pause for silence after uttering a response to allow the test taker to take over repair and sequence management. Audit transcripts for disrupted turns in validation studies before using scores as evidence. Map rating criteria onto observable aspects of interaction based on Lam et al. (2023).

Implications for teaching in EFL classrooms

This framework guides the design of AI-based speaking practice. In large classes with limited class time, like Indonesian EFL courses, learners may need an automated partner for out-of-class preparation activities, perhaps during the pre-class part of a flipped course (Baskara, 2023). The corpus suggests this type of activity could be effective for transfer activities such as practicing requests, refusals, and service interactions, with work sent back to the learner. Peer-to-peer is the most effective for interactive listening and relationship-building activities, as peers performed better at both listening and long-term retention (Meng et al., 2026; Sim et al., 2025). Non-English major students were more confident and fluent after real-audience project work (Putra & Prasetyo, 2025). However, this was not coded for AI or interactional aspects, so it remains an area to explore. Learners could review their AI conversation transcripts and see who repaired, who closed, and who listened. This might foster the crucial AI literacy that is called for by Kohnke et al. (2023) and Puri and Baskara (2023).

Limitations and research agenda

This review has several limitations. First, our search focused on two sources: OpenAlex, where we screened only the top 20 records for most strings, and one publisher's journal collection. We did not search Scopus, Web of Science, ERIC, or LLBA and may have missed relevant work. Second, the strings were based on studies known to the authors; two included author names, and Wiley question Q2 tests a specific expectation. We located six of the 16 studies through citation chasing, which suggests the searches were not fully sensitive. The main limitation of the search is that we did not search any additional databases in this version. Our corpus is small and skewed: six studies come from one research group, two from another, and only one compares an SDS and an LLM within one task. Studies had to be peer-reviewed or published in proceedings, be in English, and be published since 2016. One primary coder did all the coding, and we report the second coder's agreement in Section 2. Third, we do not assess study quality and use the same studies to develop our framework without testing it on other data. Fourth, technology is moving fast, and new low-latency voice systems may alter turn-taking results. Fifth, our counts rely on coding; if a study does not examine a practice, we cannot know who did what, and this will show up as R. Finally, the results here are hypotheses to test; they are not proven facts. We look forward to future work coding turn-by-turn interactional labor, testing whether practice with machines alters how learners interact with people, and testing whether design can return work to learners.

5. Conclusion

We asked who does the interactional work when an L2 learner talks with a machine (RQ1), how that work compares with talk between people (RQ2), and which framework captures the difference (RQ3). (RQ1): Learners bring to their interactions with machines action types at the sequence level, like requests and refusals, in forms that are typical of human talk (RQ1, RQ2). The most frequent coding is adaptation for repairs (3 of 6 findings), but turn-taking is more tentative. Only one study directly compares an SDS and an LLM, where the LLM took on repairs and closings that a human partner would share with the learner (Timpe-Laughlin et al., 2025). Relational practices and interactive listening tend to be less deep with machines (RQ2), but Meng et al. (2026) found that both groups improved, so it is primarily about listening.

This framework (RQ3) separates work with machines from practices ported to talking with people and from IC with people and human-machine competence. We suggest AI partners drill portable practice formats for returning repairs and closings to the learner, and that peers focus on relationships and listening, and that test developers communicate architecture and model version and constrain score claims accordingly. This has not been tested, and the above limitations apply. Next is coding interactional labor turn by turn and testing if working with machines affects how learners talk with people.

6. References

Ayedoun, E., Hayashi, Y., & Seta, K. (2019). Adding communicative and affective strategies to an embodied conversational agent to enhance second language learners' willingness to communicate. *International Journal of Artificial Intelligence in Education*, *29*(1), 29–57.

Baskara, F. R. (2023). Chatbots and flipped learning: Enhancing student engagement and learning outcomes through personalized support and collaboration. *IJORER: International Journal of Recent Educational Research*, *4*(2), 223–238.

Bibauw, S., Van den Noortgate, W., François, T., & Desmet, P. (2022). Dialogue systems for language learning: A meta-analysis. *Language Learning & Technology*, *26*(1).

Chen, A., Jia, J., Li, Y., & Fu, L. (2025). Investigating the effect of role-play activity with GenAI agent on EFL students' speaking performance. *Journal of Educational Computing Research*, *63*(1), 99–125.

Dai, D. W., & Roever, C. (2026). Assessing Digital Interactional Competence for Second-Language and First-Language Chinese Speakers: Effects of Proficiency, Mode, and Setting. *Language Learning*.

Dingemanse, M., & Enfield, N. J. (2024). Interactive repair and the foundations of language. *Trends in Cognitive Sciences*, *28*(1), 30–42.

Dingemanse, M., Roberts, S. G., Baranova, J., Blythe, J., Drew, P., Floyd, S., Gisladottir, R. S., Kendrick, K. H., Levinson, S. C., & Manrique, E. (2015). Universal principles in the repair of communication problems. *PloS One*, *10*(9), e0136100.

Dombi, J., Sydorenko, T., & Timpe-Laughlin, V. (2022). Common ground, cooperation, and recipient design in human-computer interactions. *Journal of Pragmatics*, *193*, 4–20.

Dombi, J., Sydorenko, T., & Timpe-Laughlin, V. (2024). Openings and closings in human-human versus human-spoken dialogue system conversations. *Language Learning & Technology*, *28*(2), 32.

Eguchi, M., Takizawa, K., Saeki, M., Kurata, F., Suzuki, S., Matsuyama, Y., & Sawaki, Y. (2025). Human-versus artificial intelligence-delivered roleplay tasks for assessing interactional competence: An applied conversation analytic study. *TESOL Quarterly*, *59*, S183–S219.

Ericsson, E., Sofkova Hashemi, S., & Lundin, J. (2023). Fun and frustrating: Students' perspectives on practicing speaking English with virtual humans. *Cogent Education*, *10*(1), 2170088.

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualizations, operationalizations, and outstanding questions. *Language Assessment Quarterly*, *15*(3), 219–236.

Godwin-Jones, R. (2024). Distributed agency in second language learning and teaching through generative AI. *arXiv Preprint arXiv:2403.20216*.

Hall, J. K. (2018). From L2 interactional competence to L2 interactional repertoires: Reconceptualizing the objects of L2 learning. *Classroom Discourse*, *9*(1), 25–39.

Hou, Z., & Min, S. (2026). Dialogue-based computer-assisted language learning systems for second language speaking development: A three-level meta-analysis. *ReCALL*, *38*(1), 40–56.

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, *38*(1), 237–257.

Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review*, *10*(1), 18–26.

Jakonen, T., Duran, D., & Peltonen, P. (2025). Situated L2 pronunciation instruction during small-group robot-assisted language learning activities. *Language Teaching Research*, 13621688251367852.

Karatay, Y., & Xu, J. (2025). Exploring the potential of conversational AI for assessing second language oral proficiency. *TESOL Quarterly*, *59*, S220–S250.

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal*, *54*(2), 537–550.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, *70*(4), 366–372.

Lam, D. M., Galaczi, E., Nakatsuhara, F., & May, L. (2023). Assessing interactional competence: Exploring ratability challenges. *Applied Pragmatics*, *5*(2), 208–233.

Lyu, B., Lai, C., & Guo, J. (2025). Effectiveness of chatbots in improving language learning: A meta-analysis of comparative studies. *International Journal of Applied Linguistics*, *35*(2), 834–851.

Meng, Y., Shen, C., Qian, X., & Wang, C. (2026). Chatbot Versus Human Peer: Effects on College English as a Foreign Language Learners' Interactional Competence Development. *Language Learning*.

Ockey, G. J., & Chukharev-Hudilainen, E. (2021). Human versus computer partner in the paired oral discussion test. *Applied Linguistics*, *42*(5), 924–944.

Ockey, G. J., Chukharev-Hudilainen, E., & Hirch, R. R. (2023). Assessing interactional competence: ICE versus a human partner. *Language Assessment Quarterly*, *20*(4–5), 377–398.

Pekarek Doehler, S. (2018). Elaborations on L2 interactional competence: The development of L2 grammar-for-interaction. *Classroom Discourse*, *9*(1), 3–24.

Pelikan, H. R., & Broth, M. (2016). Why that nao? How humans adapt to a conventional humanoid robot in taking turns at talk. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4921–4932.

Plow, I., Banerjee, J., & Iwashita, N. (2018). Interactional competence: Genie out of the bottle. *Language Testing*, *35*(3), 427–445.

Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice interfaces in everyday life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.

Puri, A. D., & Baskara, F. R. (2023). Enhancing pragmatic knowledge with ChatGPT: Benefits and considerations. *Prosiding Konferensi Linguistik Tahunan Atma Jaya (KOLITA)*, *21*(21), 18–23.

Putra, A. M. A., & Prasetyo, A. H. (2025). The Influence of Project-Based Learning on Speaking Skills: A Case Study of Political Science Students in Presenting SDGs Campaign. *Journal of Education and Teaching (JET)*, *6*(2), 396–417.

Qu, Z., & Chen, L. (2026). Simulating the "cultural other": The impact of GenAI interlocutors on EFL learners' intercultural communicative competence, speaking anxiety, and learning engagement. *Frontiers in Psychology*, *17*, 1799695.

Roever, C., & Kasper, G. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing*, *35*(3), 331–355.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simple system for the organization of turn-taking for conversation. *Language*, *50*(4), 696–735.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(2), 361–382.

Sim, S., Kim, Y., & Ku, K. (2025). Exploring Generative AI as a Roleplay Interlocutor in L2 Task-Based Pragmatics Learning: Comparing ChatGPT-Learner and Technology-Mediated Peer Interactions. *TESOL Quarterly*, *59*, S86–S116.

Su, Y., & Chen, X. (2026). Second Language Interactional Competence in Roleplay: Comparing Generative Artificial Intelligence and Human Interlocutors. *Language Learning*.

Su, Y., Dai, D. W., & Meng, Y. (2026). Digitally Empowered Assessment of Interactional Competence in a Second Language. Language Learning. https://doi.org/10.1111/lang.70050

Sydorenko, T., Daurio, P., & Thorne, S. L. (2018). Refining pragmatically appropriate oral communication via computer-simulated conversations. *Computer Assisted Language Learning*, *31*(1–2), 157–180.

Timpe-Laughlin, V., Davis, L., Oh, S., & Schmidgall, J. (2026). Conversational Agents in Practicing and Assessing Interactional Competence: Current Research and Future Directions. *Language Learning*.

Timpe-Laughlin, V., Divekar, R., Sydorenko, T., Dombi, J., & Oh, S. (2025). Intent-based versus GPT-based conversational agents: Benefits and challenges for practicing and assessing oral interaction. *TESOL Quarterly*, *59*, S117–S149.

Timpe-Laughlin, V., Dombi, J., Sydorenko, T., & Sasayama, S. (2026). L2 learners' pragmatic output in a face-to-face vs. A computer-guided role-play task: Implications for TBLT. *Language Teaching Research*, *30*(3), 1593–1618.

Timpe-Laughlin, V., Sydorenko, T., & Dombi, J. (2024). Human versus machine: Investigating L2 learner output in face-to-face versus fully automated role-plays. *Computer Assisted Language Learning*, *37*(1–2), 149–178.

Torraco, R. J. (2016). Writing integrative literature reviews: Using the past and present to explore the future. *Human Resource Development Review*, *15*(4), 404–428.

Voss, E., & Waring, H. Z. (2025). When ChatGPT cannot chat: The quest for naturalness. *TESOL Quarterly*, *59*(2), 1064–1075.

Yan, X., & Huang, B. H. (2025). Generative AI for the Teaching, Learning, and Assessment of Productive Skills: An Evidence-Based Approach to Understanding Its Real Impact. In *TESOL Quarterly* (Vol. 59, pp. S5–S18).