AIRiAL 2026 Schedule

AIRiAL 2026 Schedule

AIRiAL 2026 Pre-Conference Workshop Thursday, September 24, 2026	
9:00 – 4:00	***Alfred Essa*** **Title: From Study to Scale:** **Moving Multimodal and Multimodel AI Research in Applied Linguistics into Real-World Use**

AIRiAL 2026 Presentations Day 1: Friday, September 25, 2026	
8:00 – 9:00	Registration & Coffee
9:00 – 9:15	Welcome and Opening Remarks **Erik Voss**, Teachers College, Columbia University
9:20 – 10:20	**Session F1**
Paper 1: 9:20 – 9:40	**Evaluating Humanlikeness in AI-Generated Language: A Register-Based Framework for Benchmarking (Mis)Alignment** *Yağmur Demir, Northern Arizona University*
Paper 2: 9:40 – 10:00	**Grammatical Features of Successful Exam Writing: A Multi-Dimensional Comparison of Human and LLM texts** *Ednalvo Apostolo Campos, State University of Pará, Brazil*
Paper 3: 10:00 – 10:20	**A Corpus-Based Comparison Between AI and Human Teacher Classroom Talk** *Marilisa Shimazumi, Sao Paulo State Technology College, Brazil* *Tony Berber Sardinha, Pontifical Catholic University of Sao Paulo, Brazil* *Randi Reppen, Northern Arizona University*

10:20 – 10:35	Coffee Break
10:35 – 11:35	**Session F2**
Paper 1: 10:35 – 10:55 *Paper 2:* 10:55 – 11:15 *Paper 3:* 11:15 – 11:35	**Practicing Pronunciation with ASR: Insights into Error Awareness among French EFL Learners** *Ana Bumber, Toulouse University, France* **PRAGMO: Evaluating Multimodal AI-Mediated Feedback for L2 Pragmatic Competence in Speaking Practice** *Jiwon Choi, Iowa State University* *Inyoung Na, Iowa State University* **Multimodal and Multimodel AI Applied to Assess Language Learners' Depth of Processing of Written Feedback** *Ángeles Ortega-Luque, Georgetown University*
11:35 – 12:30	Lunch Break *Grace Dodge Hall 177 & 179*
12:30 – 1:25	**Poster and Technology Demonstration Session** (See presenters below)

1:30 – 2:30	**Plenary Speaker**
	Veronika Timpe-Laughlin, *Educational Testing Service* **Title: Generating Empathetic Dialogue for English Language Learning** (Followed by Group Picture)
2:30 – 2:45	Coffee Break
2:45 – 4:05	**Session F3**
Paper 1: 2:45 – 3:05 *Paper 2:* 3:05 – 3:25	**How Avatar-Based Interfaces Shape Speech and Engagement in Spoken Interaction with Generative AI?** *Seiji Takahashi, Northern Arizona University* **Same Language, Different Partner: Fluency and Anxiety in Real-Time Spoken AI Interaction** *Chris Litten, Iowa State University* *Young-joo Lee, Iowa State University* *Ubed Amrullah, Iowa State University* *Ben Godard, Iowa State University* *Mahdi Duris, Iowa State University*

Paper 3: 3:25 – 3:45	**AI-Powered Multimodal Preparation for Court Interpreter Certification: Tool Design and Evidence of Learner Self-Efficacy** *Elsa Perez, Utah State University*
Paper 4: 3:45 – 4:05	**Multimodal AI Interaction in ESL Job Interviews: A Comparison of Mixed Reality, ChatGPT Audio, and Traditional Instruction** *Amany Alkhayat, Media University of Applied Design, Germany*
4:05 – 4:15	Break
4:15 – 5:15	**Session F4**
Paper 1: 4:15 – 4:35	**Developing and Validating an AI-Enhanced Multimodal Listening Assessment** *Shanshan He, University of Western Ontario, Canada*
Paper 2: 4:35 – 4:55	**Can AI Speech Replace Human Speech in L2 Listening Assessment for Young Learners? Insights from the TOEFL Junior® Standard Test** *Sanshiroh Ogawa, University of Maryland* *Ekaterina Sudina, University of Maryland* *Bronson Hui, University of Maryland*
Paper 3: 4:55 – 5:15	**Dual-Layer AI Assessment in EFL Speaking Practice: Integrating Automated Scoring and LLM Formative Feedback in a Curriculum-Aligned Web Application** *John McKee, Military Technological College, Oman*

5:15 – 5:30	Announcements
5:30 – 6:45	Opening Reception (Wine & Cheese in Grace Dodge Hall 177 &179)

AIRiAL 2026 Poster and Technology Demonstration Session #1

Friday, September 25, 2026
12:30 – 1:25 pm

POSTERS

A Veces, Comes Agua: Measuring Hallucination of Idiomatic Expressions in a Language Learning Application
Timothy Farnsworth, CUNY Hunter College

Are Human Teachers Becoming Optional in EFL Speaking Classrooms, or Are We Simply Placing a Native-Speaking Tutor in Every Learner's Hand?
Heejin Park, Gyeongang National University, Korea

Automated Topic Detection in Spoken Language Assessment: Comparing an Instruction-Tuned Generative Model Against a Cross-Encoder Approach
Sabrina Sun, Language Testing International
Scott Gravina, Language Testing International
Kim Sallee, Language Testing International
Meg Malone, ACTFL

Designing a Language for Specific Purposes (LSP) Course in the Age of AI: Practical Strategies from a Course in Medical Spanish
David Ortega, Yale University

Evaluating the Validity Arguments for a ChatGPT-Powered Dynamic Pragmatics Assessment
Gi Jung Kim, Iowa State University

GenAI in a Language Teacher Education Practicum: A Case Study in ESP
Amanda Brown, Syracuse University

"I Wish Anika Would Contribute More": Designing Multimodal AI Characters for Learning-Oriented SBLA
Soo Hyoung Joo, Teachers College, Columbia University

Know a Word by Its Neighbours: Teaching Vocabulary Through LLM-Assisted Lexical-Semantic Grounding
Büşra Marşan, Stanford University
Ecem Fırat Kopuz, CUNY, City University of New York

Preserving Voices through Cross-Cultural Game Design: An LLM-Assisted Multimodal Pedagogical Approach
C. Joseph Sorell, Purdue University
Lixia Cheng, Purdue University
Srilani Wickramasinghe, Purdue University
Yutika V. Sawant, Purdue University

Using Generative AI for Speaking Practice in Language Education
Fatemeh Soleimani, University of Texas at Austin

TECHNOLOGY DEMONSTRATIONS

A Mediation Environment for Concept-Based Vocabulary Learning in Second Language Learners
Yuwei Xia, Pennsylvania State University

Evalora: A Voiced-Avatar Examiner for L2 French — What It Scores, What It Cannot
Imane Harbi, Sorbonne University Abu Dhabi, United Arab Emirates

Talio Language Tech Demo
Matthew Wise, Talio Language

WriteMed: An AI Mediator for Dynamic Assessment of L2 Academic Writing
Yiwei Qin, University of Pennsylvania
Yuwei Xia, Pennsylvania State University
Junjing Zhang, Stanford University

AIRiAL 2026 Presentations Day 2: Saturday, September 26, 2026	
8:00 – 8:50	Registration
8:50 – 9:00	Opening Remarks **Erik Voss**, Teachers College, Columbia University
9:00 – 10:00	**Session S1**
Paper 1: 9:00 – 9:20	**Stylistic (A)synchrony in Human-Machine Dyadic Interaction: Investigating Co-Adaptation Through the Lens of Communicative Naturalness** *Abby Massaro, Teachers College, Columbia University* *Ioana Wicker, Teachers College, Columbia University*
Paper 2: 9:20 – 9:40	**Co-Adaptation in Learner–ChatGPT Dyadic Interaction: A Multi-Leveled Linguistic Analysis** *Ashley Beccia, Teachers College, Columbia University* *Jill Williams, Teachers College, Columbia University*
Paper 3: 9:40 – 10:00	**It Takes Two to Tango: A Longitudinal Mixed-Methods Investigation of Human-AI Co-Adaptation Across Iterative Dialogues** *Zhizi Chen, Teachers College, Columbia University* *Liza Ostolaza, Teachers College, Columbia University*

10:00 – 10:20	Coffee Break
10:20 – 11:40	**Session S2**
Paper 1: 10:20 – 10:40	**What LLMs 'Know' about Language Use: A Case Study of Lexical Patterns in Song Lyrics** *Maria Claudia Delfino, Centro Paula Souza / Fatec Praia Grande, Sao Paulo Catholic University, Brazil*
Paper 2: 10:40 – 11:00	**Enhancing LLM Register Awareness Through Grammar: A Key Feature Analysis of Research Articles** *Tony Berber Sardinha, Pontifical Catholic University of Sao Paulo, Brazil* *Anderson Avila, Institut National de la Recherche Scientifique, Canada* *Maria Claudia Nunes Delfino, Sao Paulo Technology College (FATEC), Brazil* *Marilisa Shimazumi, Sao Paulo State University at S. José do Rio Preto (UNESP), Brazil* *Deise Prina Dutra, Federal University of Minas Gerais (UFMG), Brazil* *Paula Tavares Pinto, Sao Paulo State University at S. José do Rio Preto (UNESP), Brazil* *Ana Bocorny, Federal University of Rio Grande do Sul (UFRGS), Brazil* *Carlos Kauffmann, Pontifical Catholic University of Sao Paulo, Brazil* *Patricia Bértoli, Rio de Janeiro State University (UERJ), Brazil* *Cicero Soares da Silva, Pontifical Catholic University of Sao Paulo, Brazil* *Mirella Whiteman, Pontifical Catholic University of Sao Paulo, Brazil* *Marcos Roberto de Oliveira, Sao Paulo Technology College (FATEC), Brazil* *Rogerio Yamada, Pontifical Catholic University of Sao Paulo, Brazil* *Leandro Tessler, State University at Campinas (Unicamp), Brazil*
Paper 3: 11:00 – 11:20	**Can LLMs Capture Fashion Discourse? A Corpus-Based Evaluation of AI Using the ELFC** *Katherine Oliva Ortolani, Unimontes, Brazil*

Paper 4: 11:20 – 11:40	**Evaluating LLM-as-Judge for Interpretive Analysis of Interview Discourse: Implications for Qualitative Research** *Songhee Han, Florida State University* *Jueun Shin, Florida State University* *Jiyoon Han, Florida State University* *Bung-Woo Jun, Florida State University* *Hilal Ayan Karabatman, Florida State University*
11:40 – 12:40	Lunch Break
12:40 – 1:40	**AI Research in X Meetings**
	• AI Research and Language Assessment • AI Research and Corpus Linguistics • AI Research and Language Teaching/Learning • AI Research and Conversation Analysis • AI Research and Second Language Acquisition
1:45 – 2:45	**Session S3**
Paper 1: 1:45 – 2:05	**Beyond Detection: Rethinking Writing Assessment in the Age of Generative AI** *Yiwei Qin, University of Pennsylvania* *Yuwei Xia, Pennsylvania State University*

Paper 2: 2:05 – 2:25	**LLM-based Mediation: Exploring the Potential of AI in DA of L2 Academic Writing** *Daniel Murcia, Penn State University* *Matthew Poehner, Penn State University*
Paper 3: 2:25 – 2:45	**Supervised Finetuning for Speech Act Assessment** *Feng Xiao, Pomona College* *Kun Nie, Pomona College*
2:45 – 2:55	Break
2:55 – 4:15	**Session S4**
Paper 1: 2:55 – 3:15	**The Impact of Structured and Ethical ChatGPT Training on Academic Writing Performance in EFL Contexts** *Mohamed Almahdi, Arizona State University*
Paper 2: 3:15 – 3:35	**Ethics of AI-Mediated Language Education: Critical, Ecological, and Technomoral Perspectives** *Chaoran Wang, Colby College* *Ben Baker, Colby College*
Paper 3: 3:35 – 3:55	**Plurilingual Graduate Students' Agency in AI-Assisted Academic Writing: A Multiple-Case Study** *Anne-Marie Sénécal, Université de Sherbrooke, Canada*
Paper 4: 3:55 – 4:15	**Demographic Variation in Hedge Use in English Language Learners' Writing: A LLM-Based Metadiscourse Analysis** *Tianyu Xu, Teachers College, Columbia University* *Xinyu Xu, Teachers College, Columbia University*

4:15 – 5:10	**Poster and Technology Demonstration Session** (See presenters below)
5:10 – 5:30	Closing Remarks & Best Student Paper Award Announcement
6:30 – 9:30	Networking Dinner *The Ellington on Broadway - The Strauss Room* *2745 Broadway, New York, NY 10025*

AIRiAL 2026 Poster and Technology Demonstration Session #2

Saturday, September 26, 2026
4:15 – 5:10 pm

POSTERS

A Testable Multimodal AI Framework for Replication of Listening-to-Write Research in English for Academic Purposes (EAP)
John Bandman, Lancaster University, UK

Comparing the Effects of AI-Generated and Teacher-Delivered Formative Assessment on EFL Learners' Writing Performance and Writing Self-Efficacy
Golnoush Haddadian, Georgia State University
Nooshin Haddadian, Georgia State University
Saba Soleimani, Interdisciplinary Transformation University IT:U, Austria

Designing a Digital Diagnostic Writing Assessment for Post-IELTS EAP Learners
Yoohee Kim, Centre for Applied Language Studies (CALS), University of Jyväkylä, Finland

Developing an AI-Powered Writing Assessment Platform for Portuguese as an Additional Language: Leveraging the CorCel Corpus
Juliana Schoffen, Federal University of Rio Grande do Sul, Brazil
Elisa Stumpf, Federal University of Rio Grande do Sul, Brazil
Deise Amaral, Federal University of Rio Grande do Sul, Brazil
Tanara Kuhn, University of Coimbra, Portugal
Larissa Goulart, MontClair State University
Luiza Divino, Federal University of Rio Grande do Sul, Brazil
Isadora Hanauer, Federal University of Rio Grande do Sul, Brazil
Amanda Raupp, Federal University of Rio Grande do Sul, Brazil
Brenda Xavier, Federal University of Rio Grande do Sul, Brazil

João Victor Pacheco, Federal University of Rio Grande do Sul, Brazil
Gabriel Pazinatto, Federal University of Rio Grande do Sul, Brazil
Rafael Nunes, Federal University of Rio Grande do Sul, Brazil
Dennis Balreira, Federal University of Rio Grande do Sul, Brazil

Divergent Policies, Shared Goals: Understanding ELL Students' AI-Mediated Academic Writing Across Policy Contexts
Nathan Schrader, Fordham University
Xiaole Shan, Fordham University

Exploring EFL Doctoral Students' Use of ChatGPT in Higher-Order Academic Writing Through a Self-Regulated Learning Lens: A Pilot Qualitative Study
Erpin Said, University of Rochester

Operationalizing AI Literacy in Multilingual Writing: A Curriculum-Integrated Pedagogical Design
Gözde Durgut, University of Arizona

Position, Design, Enact: A Framework for Critical GenAI Pedagogy in Language Education
Xinyue Lu, Howard University
Zhongfeng Tian, Rutgers University–Newark

Trait-Based Automated Essay Scoring in Rater Training for the Evaluation of Young English Language Learner's Writing
Aubrey Sahouria, Center for Applied Linguistics
Yu Lan Su, Center for Applied Linguistics
Frank Wuckinski, Center for Applied Linguistics
Joshua Smith, Center for Applied Linguistics
Hyeonseong Lee, Center for Applied Linguistics, University of Maryland

TECHNOLOGY DEMONSTRATIONS

Bruhh.ai: An AI-Based System for College Advising
Prince Bashangezi, Pomona College
Feng Xiao, Pomona College

AIRiAL 2026 Program

Conference Abstracts
Paper Abstracts: Friday

Session F1
Friday, September 25, 2026 09:20 - 10:20 am

Evaluating Humanlikeness in AI-Generated Language: A Register-Based Framework for Benchmarking (Mis)Alignment

Yağmur Demir, Northern Arizona University

Achieving human-like language is often described as a central goal of generative AI systems (Clark et al., 2021). A key dimension of human language is register and register variation, yet emerging research suggests that AI-generated texts do not consistently mimic human register variation (e.g., Berber Sardinha, 2024; Demir & Egbert, 2026; Goulart et al., 2024). This raises important questions for the evaluation and benchmarking of AI language: how can we determine whether AI output (mis)aligns with human register norms, and what counts as meaningful similarity and dissimilarity?

This study introduces a corpus- and register-based framework for evaluating the humanlikeness of AI-generated language. The framework integrates three complementary analytical layers—multidimensional analysis (MDA), lexico-grammatical and semantic key feature analyses (KFA)—to assess (mis)alignment across different aspects of register variation.

A central methodological contribution is the adaptation of equivalence testing (ET) to operationalize register (mis)alignment. Effect size (Cohen's d) thresholds are established using the distribution of effect sizes reported in prior register research (Biber & Egbert, 2018), following a percentile-based approach (Plonsky & Oswald, 2014).

The framework is demonstrated through a case study comparing 100 human-authored and 100 ChatGPT-generated texts across three informational registers (research articles, university textbooks, Wikipedia) and two disciplines (biology and political science). Results from the lexico-grammatical KFA illustrate how ET can help distinguish between practical equivalence and meaningful divergence.

This study proposes a framework that enables context-sensitive benchmarks for interpreting (mis)alignment and contributes to ongoing efforts to develop more reliable, transparent, and linguistically informed evaluation practices for AI-generated language.

Grammatical Features of Successful Exam Writing: A Multi-Dimensional Comparison of Human and LLM texts
Ednalvo Apostolo Campos, State University of Pará, Brazil

This study looks at the grammatical correlates of writing success in university entrance composition exams and considers how large language models can be used as writing partners in exam preparation. Previous research has shown that linguistic features such as syntactic complexity and lexical sophistication are associated with writing quality (e.g. Crossley et al., 2010). However, such studies tend to rely on overall pre-defined measures rather than identifying the specific grammatical features that characterize exam writing. To address this issue and determine the extent to which different LLMs can provide human-like texts that meet entrance exam requirements, we conduct a corpus-based study on the grammatical features associated with composition writing. The human-authored corpus consists of 500 compositions written for the entrance exam of a major state university in Brazil, where failure in the composition component results in overall exam failure. The corpus was divided into passing (P) and failing (F) texts, with the latter being rewritten submitted by two LLMs (GPT 5.2 and Gemini 3.1) according to the official six-criterion rating scheme. All texts were tagged with the PALAVRAS Parser and counts the linguistic features were taken by a specialized script, which were then submitted to a Multi-Dimensional Analysis (Berber Sardinha & Veirano Pinto, 2019; Biber, 1988; Veirano Pinto et al., 2026). The resulting dimensions were interpreted and the LLM-generated texts compared to those of human-authored passing texts. Results indicate differences in grammatical resources between successful human writing and successful LLM texts, with implications for exam preparation.

A Corpus-Based Comparison Between AI and Human Teacher Classroom Talk
Marilisa Shimazumi, Sao Paulo State Technology College, Brazil
Tony Berber Sardinha, Pontifical Catholic University of Sao Paulo, Brazil
Randi Reppen, Northern Arizona University

This paper looks at how different large language models (GPT-5 and Gemini 3) simulate classroom teaching in higher education, in comparison with real human-led classroom interaction. LLMs have been reported to take on tasks in education (Kasneci et al., 2023), but whether they could take on the role of educators in a classroom setting remains unclear. To explore that scenario from a multimodel corpus-based applied linguistics perspective, the study uses two human-taught classroom corpora: one of face-to-face university classes and another of online classes. These corpora were paired with AI-generated teacher talk simulations. The resulting corpus contains 30+ million words of human and AI teaching discourse. The analysis uses Lexical Multi-Dimensional Analysis (Berber Sardinha & Fitzsimmons-Doolan, 2025) to identify how human and AI teachers conduct the classroom interaction with the students. The dimensions for the online classes include contrasts such as macro-structuring versus performative stagecraft and informal facilitation versus formal deductive exposition. The results show differences between human and AI teacher talk. The LLM exaggerates or introduces traits not present in the human interaction, which include, for example, extreme evaluative language, sarcastic remarks, or exaggerated expressions of appreciation. Additionally, the LLM-simulated classes often did not differentiate among the different teaching styles. In conclusion, the study shows that LLMs can generally reproduce aspects of teaching discourse

but use different language traits from human interaction. These findings contribute to the understanding of how AI performs institutional roles and point to the need for careful assessment of AI-generated language in educational contexts.

Session F2
Friday, September 25, 2026 10:35 - 11:35 am

Practicing Pronunciation with ASR: Insights into Error Awareness among French EFL Learners

Ana Bumber, Toulouse University, France

Automatic Speech Recognition (ASR) technology has been used for pronunciation practice for more than a decade, as attested by Liu et al. 2025 systematic literature review (SLR). While most cited articles involving ASR for pronunciation practice use dictational models, fewer are those such as Spring and Tabuchi, 2022 that use Automatically Generated Closed Captions (AGCCs) as the basis for ASR assisted pronunciation training. Even though AGCCs are continuously improving thanks to ongoing training of the models behind the ASR, they are still not one hundred percent accurate. AGCCs were not initially invented for pronunciation practice; their purpose is to ensure the accessibility of video content on the Internet. The very hallucinations that AGCCs produce can be valuable indicators that elements in the audio track are non-standard. These could be used by learners to point to non-standard pronunciation in recordings of their own speech. In addition, none of the 24 studies selected by Liu et al. 2025 targeted French learners of English as a Foreign Language (EFL) and most of them were conducted with Asian students of EFL.

In the study presented here, two groups of French students recorded videos of themselves speaking in English and generated AGCCs using the ASR in Microsoft Stream. Was the method user friendly? According to students, how effective are AGCCs for identifying and correcting pronunciation errors? While most concluded on the usefulness of AGCCs for training pronunciation, many students pointed out its limitations in their qualitative answers.

PRAGMO: Evaluating Multimodal AI-Mediated Feedback for L2 Pragmatic Competence in Speaking Practice

Jiwon Choi, Iowa State University
Inyoung Na, Iowa State University

Pragmatic competence is a key component of second language (L2) oral proficiency, yet it remains challenging for learners to develop (Taguchi, 2015). It requires exposure to social contexts beyond classroom instruction, and pragmatic errors are rarely addressed in natural interaction, limiting opportunities to notice and repair inappropriate language use (Bardovi-Harlig & Dörnyei, 1998). This study develops and evaluates PRAGMO (Pragmatics in Motion), a multimodal AI-mediated speaking practice platform. PRAGMO integrates GPT-4o realtime preview as the AI interlocutor, GPT-4.1 for pragmatic analysis and evaluation, and HeyGen API for video feedback generation. Learners engage in a role-play scenario (SP1), receive targeted multimodal feedback on their weakest

pragmatic areas, and produce a revised response (SP2), allowing analysis of modified output as evidence of feedback uptake. At the system level, assessment experts (n = 2) and L2 learners (n = 3) evaluated feedback effectiveness, the AI's appropriateness in conversation, and the human-likeness of the AI interlocutor. Semi-structured interviews explored perceptions of the system's conversational quality and assessment potential. At the learner level, performance analysis focused on (1) change in pragmatic performance from SP1 to SP2 and (2) video feedback uptake. The system's pragmatic scores were compared with those of two human raters. Results showed overall improvement in learners' pragmatic performance in SP2, with modified output from video-based feedback. Participants perceived feedback as effective; however, the AI interlocutor showed limitations in conversational variation. This study offers insights into AI-mediated speaking systems for L2 pragmatic development, highlighting the role of multimodal feedback in supporting pragmatic competence.

Multimodal and Multimodel AI Applied to Assess Language Learners' Depth of Processing of Written Feedback
Ángeles Ortega-Luque, Georgetown University

This study investigates how multimodal and multimodel AI may support the assessment of second language (L2) learners' depth of processing (DoP) of written corrective feedback. Prior research shows that deeper processing facilitates L2 learning (Leow, 2020), yet analyzing L2 learners' cognitive engagement with feedback in instructional contexts has been methodologically challenging and time-consuming. To address this gap, a corpus was constructed integrating audio and textual data from bilingual (Spanish and English) think-aloud protocols conducted during rewrites by 25 advanced L2 Spanish learners. Audio recordings were transcribed automatically using Whisper. The resulting transcripts were analyzed with large language models (LLMs) (ChatGPT and Claude) and Python-based natural language processing for linguistic feature extraction. Results show the (1) validity and time-efficiency improvement of LLMs' automated scoring of Leow's (2015) three levels of DoP, (2) reliability between human and LLM coders, and (3) top linguistic features associated with Leow's (2015) three levels of DoP, and their consequent learning outcomes measured by accuracy in the rewrites. These findings shed light on the potential of multimodal and multimodel AI to assist teachers and scholars in assessing well-established and empirically-proven theoretical underpinnings and instructed language learning. This presentation aims to contribute to ongoing discussions on the validity, scalability, and reliability of AI-assisted L2 learning assessment, methodological innovation, and the ethical integration of AI in Applied Linguistics.

Session F3
Friday, September 25, 2026 02:45 - 04:05

How Avatar-Based Interfaces Shape Speech and Engagement in Spoken Interaction with Generative AI?
Seiji Takahashi, Northern Arizona University

Avatar-based generative AI (GenAI) chatbots with visual interfaces have been suggested to enable more human-like communication, with growing interest in their applications in second language conversational practice and assessment. However, empirical evidence on how such interfaces shape users' speech in interaction with those systems remains limited. To address this gap, the present study investigated whether differences in GenAI visual interface (a human-like, lip-syncing avatar vs. a conventional display, using the same system) influence speech in ways that are perceptible to listeners, as reflected in judgments of machine-directedness (i.e., how much speech sounds "addressed to a machine") and perceived engagement (i.e., how engaged the speaker sounds).
To generate speech samples, four native-English speakers interacted with the GenAI under two conditions (avatar vs. conventional) and with a human interlocutor (baseline condition). A total of 72 utterance clips were extracted and evaluated by 44 native English-speaking listeners for machine-directedness and engagement. Linear mixed-effects modeling revealed no significant differences in perceived machine-directedness across the three conditions. For engagement, speech directed to both AI systems was rated similarly, while speech directed to a human received higher ratings. Consistent with these findings, acoustic analyses showed no differences between the two AI conditions across all measures. In contrast, human-directed speech exhibited higher mean F0 and greater intensity—features commonly associated with increased engagement—than AI-directed speech. These findings suggest that while spoken GenAI systems may not elicit overtly machine-directed speech, they may still fall short of eliciting the level of engagement typically observed in human–human interaction.

Same Language, Different Partner: Fluency and Anxiety in Real-Time Spoken AI Interaction
Chris Litten, Iowa State University
Young-joo Lee, Iowa State University
Ubed Amrullah, Iowa State University
Ben Godard, Iowa State University
Mahdi Duris, Iowa State University

AI systems capable of processing and producing real-time speech with low latency represent a fundamentally different type of interlocutor from the text-based and asynchronous tools that dominate research on AI-mediated language learning. Yet no empirical work has examined how real-time spoken AI affects fluency and anxiety in interaction, nor has L1 speaker performance been used to contextualize what AI-mediated speech actually looks like. This study investigates how interaction with Sesame.ai, a real-time,

multimodal, multi-model spoken AI agent, shapes L1 spoken fluency and whether speaker anxiety acts as a confounding variable, raising concerns about whether AI-mediated interaction may influence performance.

Eleven L1 English speakers completed three structured speaking tasks using Sesame.ai: descriptive, argumentative, and cooperative. By focusing on L1 speakers, the study isolates the interactional effects of the AI interlocutor, providing a reference point for interpreting L2 fluency and affect in future research. To capture the multidimensional nature of spoken performance, the study triangulates acoustic measures of non-linguistic resources (Lennon, 1990), perceptual fluency ratings — in which listeners draw on both linguistic and non-linguistic resources (Kormos & Dénes, 2004; Freed, 1995) — Linear mixed model is employed to determine which acoustic features predict human perception.

Early results suggest that human speakers adjust their fluency to accommodate the AI's computer-driven delivery, and that the non-human fluency characteristics produced by a clash of interactional "architectures" trigger anxiety unrelated to a speaker's actual speaking ability, echoing prior findings that system-driven breakdowns in AI interaction can lead to user frustration (Chhabra et al., 2024).

AI-Powered Multimodal Preparation for Court Interpreter Certification: Tool Design and Evidence of Learner Self-Efficacy

Elsa Perez, Utah State University

Court interpreting is an inherently multimodal task: interpreters must process written legal documents, spoken testimony, prosodic cues, and courtroom visual context — often simultaneously, and always under high stakes. Yet conventional preparation for U.S. certification exams (Utah Courts Approved Interpreter; ACTFL Oral Proficiency Interview, Superior; Oral Performance Exam, OPE) remains largely unimodal, relying on role plays that scale poorly and offer limited feedback. This study reports the design and evaluation of a multimodal AI system for an advanced court interpreting course at Utah State University. The system targets each exam component with a distinct AI architecture. For sight translation, generative LLMs produce authentic legal passages (motions, jury instructions, plea forms) calibrated to courtroom register. For consecutive practice, conversational agents simulate witness, attorney, and defendant turns with realistic pacing, accent variation, and embedded ethical dilemmas. For simultaneous practice, AI-generated multi-speaker audio is paired with sight-text overlays. For OPE preparation, learners complete mock exams scored against official criteria: completeness, accuracy, register fidelity, fluency, neutrality, and terminology. For OPI Superior, an interactive examiner simulates the four-task ACTFL protocol. Across modes, AI feedback combines linguistic, prosodic, and time-coded analysis of learners' recorded renditions. A 20-item self-efficacy instrument was administered (N = 29). Bartlett's test was significant (χ^2 = 546.81, $p < .001$), KMO = .738, and a six-factor solution explained 85.34% of variance. Learners reported highest confidence in ethical practice (M = 6.33) and lowest in legal-terminology accuracy (M = 5.39); α = .69–.91. Findings contribute preliminary evidence on multimodal AI architectures in high-stakes interpreter education.

Multimodal AI Interaction in ESL Job Interviews: A Comparison of Mixed Reality, ChatGPT Audio, and Traditional Instruction

Amany Alkhayat, Media University of Applied Design, Germany

Speaking a second language is often considered one of the most anxiety-inducing skills to acquire. For many ESL learners, limited vocabulary, fear of judgement, and lack of sufficient practice in crowded classrooms can impede speaking performance. Although the literature has predominantly focused on interventions in general English classrooms, English for Specific Purposes courses remain underdeveloped. This study employed a quasi-experimental, mixed-method design, where experimental groups practiced job interview questions with a multimodal embodied conversational AI (MECAI) in mixed reality (MR) simulation or the ChatGPT voice application, in comparison with a control group. Students (N = 75), ages 18–64, from a community college participated in the study. All students completed three pre- and posttests (mock job interview, Foreign Language Classroom Anxiety Scale, and heart rate measures) and were assigned to a six-week intervention. Finally, participants were interviewed to explore their perceptions of the study intervention.

Key findings showed that although there was no significant difference between groups in the pre-intervention mock job interview, FLCAS, or heart rate, both AI groups (MECAI in MR and ChatGPT voice app) showed significant post-intervention improvement, with the MR group demonstrating the greatest improvement in fluency, vocabulary, and total language performance. However, despite students reporting anxiety reduction in reflection interviews, posttests of FLCAS and heart rate showed no statistically significant reduction in physiological or self-reported anxiety. Both AI conditions helped address gaps in teacher and curriculum knowledge in ESP-oriented pedagogy, where speaking remains underexplored.

Session F4
Friday, September 25, 2026 04:15 - 05:15

Developing and Validating an AI-Enhanced Multimodal Listening Assessment

Shanshan He, University of Western Ontario, Canada

Academic L2 listening is increasingly conceptualised as a multimodal construct (Campoy-Cubillo & Querol-Julián, 2021), requiring the integration of auditory and visual information (e.g., facial expressions) that characterises real-world academic communication. However, most high-stakes listening assessments remain predominantly audio-only, creating a misalignment between construct definition and operationalisation. Recent advances in generative AI and multimodal systems offer new opportunities to address this gap, yet research has focused largely on automated scoring, with limited attention to AI-driven stimulus and item generation for listening assessment. This study develops and validates an AI-enhanced multimodal academic listening assessment, addressing the question: To what extent does the AI-enhanced multimodal academic listening assessment demonstrate reliability and validity? The

study comprises two iterative stages—test development and validation—guided by the Successive Approximation Model (Allen, 2024) and a human-in-the-loop approach to AI integration. Twenty-four 1-minute excerpts from authentic lectures were selected and analysed using NLP tools (e.g., Coh-Metrix and TAALES) to ensure comparability across 29 indices of lexical, syntactic, and discourse complexity, followed by hierarchical clustering. Multimodal stimuli were created using AI-generated talking-head avatars (HeyGen) synchronised with the original audio. Forty-eight multiple-choice items were generated using ChatGPT and iteratively refined through prompt engineering and expert review. Validation followed an argument-based approach (Chapelle, 2020), incorporating expert judgements, questionnaire data from 50 test takers, item analysis, and analyses of score reliability and performance differences between multimodal and audio-only versions. The study concludes by discussing the implications of AI-driven multimodal assessment for innovation in language learning and testing in the age of GenAI.

Can AI Speech Replace Human Speech in L2 Listening Assessment for Young Learners? Insights from the TOEFL Junior® Standard Test

Sanshiroh Ogawa, University of Maryland
Ekaterina Sudina, University of Maryland
Bronson Hui, University of Maryland

Artificial intelligence (AI) has enabled synthetic speech to sound nearly indistinguishable from natural speech. Indeed, some high-stakes tests (e.g., Duolingo English Test) have adopted text-to-speech (TTS) technology for listening stimuli. However, research suggests potential challenges associated with synthetic speech, including reduced comprehensibility, lack of naturalness, and higher cognitive load compared to human speech. These issues have implications for the construct validity of L2 listening assessments, particularly cognitive validity (Field, 2013). Young learners may be especially susceptible to such effects, putting a vulnerable population at risk. This study explored the use of synthetic speech in L2 listening assessment using items from the TOEFL Junior® Standard Test. Listening stimuli were presented in human speech, traditional concatenative TTS, and AI-based TTS. A total of 438 Japanese middle and high school learners of English completed different parts of the listening section in different voices in a within-subject, counterbalanced design. Two research questions (RQs) were addressed: (1) Does synthetic speech affect item difficulty? (2) Are test-taker impressions of the voices associated with item-level accuracy? These RQs were answered by differential item functioning (DIF) analysis and logistic mixed-effects modeling. Results suggest that synthetic speech may result in minimal differences in item difficulty estimates, and that comprehensibility is significantly associated with item response accuracy, but not naturalness, prosody, or social impression. Qualitative data suggest that lower-quality TTS may affect test-taking experience, despite its minimal impact on test performance. Findings are discussed in terms of construct validity, ethical implementation of AI-generated stimuli in standardized assessment, and pedagogical applications.

Dual-Layer AI Assessment in EFL Speaking Practice: Integrating Automated Scoring and LLM Formative Feedback in a Curriculum-Aligned Web Application

John McKee, Military Technological College, Oman

Foundation-level English as a Foreign Language (EFL) learners in pre-intermediate programmes often demonstrate low speaking confidence in spite of regular classroom instruction. Despite growing interest in artificial intelligence (AI)-supported language assessment, few studies examine the integration of automated pronunciation scoring with large language model (LLM) formative feedback within a single, curriculum-aligned intervention designed to reduce speaking anxiety. This paper presents SpeakCheck, a purpose-built web application supporting autonomous speaking practice among foundation-level EFL learners at Common European Framework of Reference (CEFR) B1–B2 level at a tertiary institution in Oman. SpeakCheck combines Speechace API pronunciation and fluency scoring with LLM-generated formative feedback across four criteria aligned with widely recognised EFL speaking assessments: fluency, grammar, vocabulary, and clarity delivered through a scaffolded, curriculum-aligned Learning Check model covering seven units of the programme's pre-intermediate speaking curriculum. A quasi-experimental mixed-methods design compares two experimental and two control intact classes, comprising approximately 80 learners, over a seven-week pilot. Quantitative instruments include Speechace scores across eighteen learning checks, pre/post departmental speaking rubric scores, and Foreign Language Classroom Anxiety Scale (FLCAS) ratings. Qualitative data include AI feedback transcripts, student perception surveys, and semi-structured interviews analysed thematically following Braun and Clarke. This study contributes an original evaluation of dual-layer AI assessment: combining quantitative automated scoring with qualitative LLM feedback within a low-stakes, self-access pedagogical design. Findings address reliability, student perceptions of AI feedback, and implications for responsible AI integration in language assessment contexts.

Session S1
Saturday, September 26, 2026 09:00 - 10:00 am

Stylistic (A)synchrony in Human-Machine Dyadic Interaction: Investigating Co-Adaptation Through the Lens of Communicative Naturalness

Abby Massaro, Teachers College, Columbia University
Ioana Wicker, Teachers College, Columbia University

Massaro and Wicker investigate the development of stylistic synchrony between a human learner and ChatGPT over a seven-week period, using the communicative naturalness framework proposed by Voss and Waring (2025). Naturalness in interaction is generally perceived as a necessary foundation for ensuring spontaneous, fluent and comfortable exchange of information. Though naturalness is a given, more or less, in interpersonal (i.e., human–human) interaction, this is not always the case for human–AI interaction. While large language models like ChatGPT undergo continual improvement, they still tend to fall short of natural "humanlike" communication (Voss & Waring, 2025).

This analysis addresses how ChatGPT's perceived naturalness influences co-adaptation in the learner–AI dyad. The data were iteratively coded across Voss and Waring's (2025) three naturalness principles: emergence, indexicality, and recipient design. Analyses revealed that early initial stylistic synchrony between the dyad became increasingly divergent, with data suggesting that violations of conversational naturalness shown by ChatGPT may have prompted the learner's shift toward increasingly informal and unpolished language—and inhibited stylistic co-adaptation overall.

The analysis demonstrates that stylistic synchrony in the human–AI dyad is a bidirectional phenomenon. Because ChatGPT relies on a dynamic, pattern-based, and context-aware neural network to adjust its style, minimal or incomplete input from the human makes accurate stylistic adaptation difficult. These findings illustrate that stylistic co-adaptation in human–AI interaction emerges from the interaction itself and is primarily driven by the human user's input. The analysis thereby offers useful insights for developing more adaptive educational AI tools for language acquisition.

Co-Adaptation in Learner–ChatGPT Dyadic Interaction: A Multi-Leveled Linguistic Analysis

Ashley Beccia, Teachers College, Columbia University
Jill Williams, Teachers College, Columbia University

Beccia and Williams examine how co-adaptation unfolds in a human–ChatGPT dyad by conducting a multi-level linguistic analysis of the seven-week interactional dataset via process tracing. Existing research on AI–learner interactions has primarily focused on the products of these exchanges, such as task performance gains (Sok & Shin, 2025) or the degree of overlap between AI-generated text and learners' final drafts (Kusumaningrum et al., 2024), while the interactional process itself is rarely a central concern. As a result, little is known about how learners and ChatGPT mutually adapt at a granular level over time. This analysis addresses this gap by tracing co-adaptation across language (form) and content (meaning) through iterative coding of the longitudinal data.

The analysis revealed that co-adaptation occurred unevenly across levels. At the level of language, lexical synchrony was modest and mostly local within individual exchanges, and syntactic synchrony was limited to two interrogative–declarative pairings that emerged sporadically across the data. By contrast, content synchrony was stable and cumulative: the learner consistently initiated topics through task-oriented prompts, while ChatGPT responded with substantially longer elaborations, yielding high topical alignment alongside persistent non-convergence in turn length. This asymmetry suggests that co-adaptation in learner–ChatGPT interaction can be strong in what is addressed but weak in how it is addressed. The findings contribute to emerging work on co-adaptation as a mechanism for L2 development by showing that co-adaptation in human–AI dyadic interaction is level-specific, emerging robustly in content yet appearing fleetingly and selectively at the level of language.

It Takes Two to Tango: A Longitudinal Mixed-Methods Investigation of Human-AI Co-Adaptation Across Iterative Dialogues

Zhizi Chen, Teachers College, Columbia University
Liza Ostolaza, Teachers College, Columbia University

Chen and Ostolaza explore psychological, cognitive, and interactional dimensions of co-adaptation across the seven weeks of learner–ChatGPT dyadic interaction. While existing research has examined what learners gain from AI interaction, the cognitive and psychological processes through which learners and AI systems mutually adapt during sustained dialogue remain underexplored. Analyses combined quantitative indices via Linguistic Inquiry Word Count (LIWC), an automated text-based computational program for natural language processing (Pennebaker et al., 2015), with qualitative coding of turn-by-turn interactional practices, linking quantitative trends to the interactional moves that produced them.

LIWC indices, including analytic thinking and authenticity, fluctuated across weeks, with the clearest convergence occurring in the middle of the dataset, while divergence characterized later weeks. Co-adaptation emerged nonlinearly through clarification, repair, reformulation, and semantic alignment. Qualitative analysis focused on Weeks 5-7, when repair sequences and collaborative alignment became more pronounced, revealing that the learner increasingly assumed interactional control through correction and topic-setting, while ChatGPT adapted through uptake, reformulation, and shifts in tone and stance. The findings suggest that human–AI co-adaptation is a nonlinear, co-constructed process in which sustained dialogue can evolve from question-answer routines into more collaborative exchange.

Session S2
Saturday, September 26, 2026 10:20 - 11:40 am

What LLMs 'Know' about Language Use: A Case Study of Lexical Patterns in Song Lyrics

Maria Claudia Delfino, Centro Paula Souza / Fatec Praia Grande, Sao Paulo Catholic University, Brazil

Fluent language use is primed by repeated encounters with language patterns, which are stored in memory (Hoey, 2005). Yet, corpus linguistics has shown that speakers' intuitions about language use do not match patterns observed in corpora (Sinclair, 2004). Human linguistic intuition involves conscious retrieval restricted by memory that tends to yield simplified, idealized accounts of language. Large language models (LLMs), on the other hand, should have no such restrictions, as they generate fluent text without relying on conscious retrieval and are not subject to the same constraints as human intuition. On this basis, one might expect them to provide more accurate judgments about linguistic patterns, and thus to approximate the patterns found in a corpus more closely than human intuition. This study tests that expectation by comparing judgments about a specific register (pop song lyrics) with lexical patterns identified through Lexical Multidimensional Analysis (Berber Sardinha & Fitzsimmons-Doolan, 2025), through the analysis of a 1.2 million-word, 4,000-text corpus, balanced both between human-authored and AI-generated lyrics, and across musical genres

(country, pop, rap, rock, and soul). Expectations about the lexical patterns were elicited from human participants and three LLMs (GPT, Gemini, and Llama), all asked to characterize the typical patterns in pop lyrics. Their judgments were compared with the corpus findings. The results show that neither human participants nor LLMs could reliably anticipate the lexical dimensions identified. These findings reflect the problem of a lack of register awareness (Berber Sardinha, 2024) in LLMs, suggesting they are register-underspecified language-wide models.

Enhancing LLM Register Awareness Through Grammar: A Key Feature Analysis of Research Articles

Tony Berber Sardinha, Pontifical Catholic University of Sao Paulo, Brazil
Anderson Avila, Institut National de la Recherche Scientifique, Canada
Maria Claudia Nunes Delfino, Sao Paulo Technology College (FATEC), Brazil
Marilisa Shimazumi, Sao Paulo State University at S. José do Rio Preto (UNESP), Brazil
Deise Prina Dutra, Federal University of Minas Gerais (UFMG), Brazil
Paula Tavares Pinto, Sao Paulo State University at S. José do Rio Preto (UNESP), Brazil
Ana Bocorny, Federal University of Rio Grande do Sul (UFRGS), Brazil
Carlos Kauffmann, Pontifical Catholic University of Sao Paulo, Brazil
Patricia Bértoli, Rio de Janeiro State University (UERJ), Brazil
Cicero Soares da Silva, Pontifical Catholic University of Sao Paulo, Brazil
Mirella Whiteman, Pontifical Catholic University of Sao Paulo, Brazil
Marcos Roberto de Oliveira, Sao Paulo Technology College (FATEC), Brazil
Rogerio Yamada, Pontifical Catholic University of Sao Paulo, Brazil
Leandro Tessler, State University at Campinas (Unicamp), Brazil

In this paper, we propose steering LLMs toward greater register awareness in research article writing by explicitly instructing models to adopt key grammatical features (Egbert & Biber, 2023) from human texts. Corpus-based research has shown that LLMs struggle to generate register-specific texts that match human writing (e.g. Berber Sardinha, 2024). A central difficulty in approximating LLM output to human texts from a grammatical perspective is that grammar is learned indirectly by LLMs, as a by-product of the word sequences they produce, since models do not represent grammar explicitly as an inventory of features. As a result, grammatical features emerge from lexical choices rather than being directly controlled. When instructed to increase the use of a given feature, the model must realize it through wording, which creates uncertainty: the feature may appear infrequently due to lexical constraints or a lack of grammatical knowledge. To address this issue, we compiled a 2,700-text (2.8 million words) corpus of research articles in Applied Linguistics, Biology, and Chemistry. Articles were segmented into abstract, introduction, method, results, discussion, and conclusion, and key features were computed for each field-section combination. We then generated a base corpus of 50 paragraphs per combination using GPT 5.4 and Gemini 3.1. Two steering strategies were tested: prompting and fine-tuning. Prompting relied on feature-based instructions and examples, and fine-tuning used instruction-response pairs derived from the base corpus. The results

showed that prompting induced the targeted grammatical features more efficiently, with GPT outperforming Gemini, whereas fine-tuning produced lower rates of feature realization.

Can LLMs Capture Fashion Discourse? A Corpus-Based Evaluation of AI Using the ELFC

Katherine Oliva Ortolani, Unimontes, Brazil

This study investigates whether large language models (LLMs) can reproduce discourse patterns identified through corpus research methods in applied linguistics, using the English Language Fashion Corpus (ELFC; Ortolani, 2024), a large multi-register corpus composed of diverse textual and media sources, namely newspaper and magazine articles, television programs, social media posts, blogs, documentaries, academic texts, fiction books, films and tutorials. This research builds on prior findings obtained through Lexical Multidimensional Analysis (LMD; Berber Sardinha, 2019, 2021; Berber Sardinha & Fitzsimmons-Dooley, in prep; Fitzsimmons-Dooley, 2019, forthcoming; Clarke et al. 2022), which identified five discourse dimensions in the domain of fashion. The present study prompts LLMs to analyze comparable datasets and generate discourse categorizations, which are then systematically compared to LMD derived dimensions. The comparison focuses on alignment, divergence, and interpretability of results, examining whether LLM outputs capture the same discourse patterns identified by humans. The findings contribute to ongoing discussions on the reliability, validity, and explainability of AI systems in applied linguistics, highlighting both the potential and limitations of LLMs as tools for discourse analysis. By bridging corpus linguistics and artificial intelligence, this study proposes a novel framework for evaluating AI-generated linguistic interpretations. The results have implications for the use of AI in linguistic research, and the analysis of complex discourse domains.

Evaluating LLM-as-Judge for Interpretive Analysis of Interview Discourse: Implications for Qualitative Research

Songhee Han, Florida State University
Jueun Shin, Florida State University
Jiyoon Han, Florida State University
Bung-Woo Jun, Florida State University
Hilal Ayan Karabatman, Florida State University

As qualitative researchers increasingly use large language models (LLMs) to support interpretive analysis, model outputs are often adopted without systematic evaluation of interpretive quality or comparison across models. This is especially problematic for interview-based research, where meaning depends on discourse context, speaker intent, and nuance beyond literal wording. This study examines whether LLM-as-judge evaluation can meaningfully align with human judgments of interpretive quality and support model selection for language-centered qualitative analysis. Using 712 conversational excerpts from semi-structured interviews with K-12 mathematics teachers, we generated one-sentence interpretive responses with five inference models: Command R+, Gemini 2.5 Pro, GPT-5.1, Llama 4 Scout-17B Instruct, and Qwen 3-32B Dense. We then evaluated these responses using AWS Bedrock's LLM-as- judge framework across five metrics and compared automated scores with human ratings on interpretive accuracy, nuance

preservation, and interpretive coherence. Results show that automated scores capture broad directional trends in human evaluation at the model level but diverge substantially in score magnitude. Among the automated metrics, Coherence showed the strongest alignment with aggregated human ratings, while Faithfulness and Correctness were less reliable for non-literal, context-dependent interpretations. Safety-oriented metrics were largely irrelevant to interpretive quality. These findings suggest that LLM-as-judge methods may be useful for screening or eliminating underperforming models, but they should not replace human judgment in interpretive analysis of interview discourse. The study offers practical guidance for more systematic and transparent use of LLMs in qualitative research workflows.

Session S3
Saturday, September 26, 2026 01:45 - 02:45

Beyond Detection: Rethinking Writing Assessment in the Age of Generative AI

Yiwei Qin, University of Pennsylvania
Yuwei Xia, Pennsylvania State University

Generative AI tools have become routinely embedded in students' writing processes, raising urgent questions about how instructors should assess work whose authorship is increasingly distributed between students and machines. The dominant institutional response—deploying AI detection software—rests on two faulty assumptions. Empirically, detectors produce high false-positive rates and disproportionately misclassify the writing of multilingual students and non-native English writers (Liang et al., 2023; Weber-Wulff et al., 2023), making them unreliable instruments for high-stakes judgment. Conceptually, detection reduces writing to a textual artifact and asks the wrong question: it tells instructors whether a text appears human-written but reveals nothing about how a student planned, struggled, or developed across drafts.

This paper proposes that the field move beyond detection toward process-oriented assessment. Drawing on Vygotsky's sociocultural account of mediated learning and Freire's critique of banking education, it argues that generative AI is best understood not as a categorically new threat but as a mediational tool whose pedagogical value depends on how learners engage with it. The paper develops a portfolio-based assessment framework with four components—planning artifacts, multiple drafts, a structured process memo on AI use, and a final draft—evaluated along four dimensions that distribute weight across product, revision depth, reflective quality, and transparency about tool use. The framework reframes assessment from surveillance to dialogue, repositioning multilingual student writers as agents who account for their own learning rather than suspects subjected to biased automated judgment. Implications for L2 writing pedagogy and institutional policy are discussed.

LLM-based Mediation: Exploring the Potential of AI in DA of L2 Academic Writing

Daniel Murcia, Penn State University
Matthew Poehner, Penn State University

Integrating Large Language Models (LLMs) to support learner L2 writing development has generated interest among researchers and led to several approaches reflecting different operationalizations of the writing process (e.g., Dai et al., 2023; Escalante et al., 2023, Kurt & Kurt, 2024; Meyer et al., 2024). Runge et al. (2025), for example, conceptualize process as elaboration through a contingent second writing stage while Meyer et al. (2024) treat it as movement from draft to feedback to revision. Our project aims for an expanded integration of LLMs into the composition and revision process guided by dynamic assessment, or DA (Poehner, 2008; Poehner & Lantolf, 2024). In DA, support, or mediation, that is dialogic and responsive to learner needs is used to diagnose and promote learner abilities, including writing (Wang-Doyle, 2025; Yu, 2022). However, a challenge to implementing DA has been its time- and labor-intensive nature. Working with examples of human mediation in published L2 DA studies and using cascade logic-tree prompting, our project explores the potential of two OpenAI models, GPT-4.1 and GPT-5.2, to provide graduated (i.e., tiered) mediation to learners as they require it. Findings from a small-scale pilot with students in a U.S. university academic writing program reveal the value of modelling instructions to reduce information flooding and that guide each GenAI model to more contingent and responsive forms of support. The study reconsiders LLMs as mediational systems by identifying future possibilities for multimodal and multilingual DA, including L1 mediation and the interpretation of learner nonverbal behavior.

Supervised Finetuning for Speech Act Assessment

Feng Xiao, Pomona College
Kun Nie, Pomona College

This study investigates supervised fine-tuning (SFT) as a methodology for automated assessment of L2 pragmatic competence. Conventional approaches rely on human raters, which limits scalability and introduces variability in scoring. To address this, we fine-tune a compact open-weight language model (Qwen-3B) to learn a mapping from learner responses to human-assigned scores, focusing on L2 Chinese request-making. The dataset comprises 1,240 responses produced by 62 American learners through a computerized oral discourse completion task with 20 situational prompts. To maximize label reliability, we use a high-agreement subset of 489 responses with unanimous human ratings for supervised fine-tuning and reserve the remaining 751 responses for out-of-sample evaluation. The model is trained to predict scalar scores, treating assessment as a regression problem aligned with human judgment. Results show that SFT yields strong alignment with human ratings, achieving a Pearson correlation of $r = .65$ ($p < .001$). In addition, the model demonstrates robust internal consistency (MnSq = 0.87), suggesting stable score calibration. These findings indicate that SFT effectively transfers human rating behavior to a small language model under limited but high-quality supervision. Overall, this study positions supervised fine-tuning as a viable and data-efficient paradigm for automated pragmatic assessment. We highlight its potential for improving scalability and consistency in L2 evaluation, and discuss implications for annotation design, model interpretability, and the development of reliable NLP-based assessment systems.

Session S4
Saturday, September 26, 2026 02:55 - 04:15

The Impact of Structured and Ethical ChatGPT Training on Academic Writing Performance in EFL Contexts
Mohamed Almahdi, Arizona State University

Grounded in cognitive interactionist theory, this study examined the long-term effects of ChatGPT as a writing support tool on students' academic writing performance across four domains: task response, coherence and cohesion, lexical resource, and grammatical accuracy and range. To investigate these effects, a 28-week study was conducted. Participants were assigned to either a control group or an experimental group, with both groups receiving equal instructional time. The experimental group received conventional teacher-led writing instruction supplemented by training in the ethical use of ChatGPT, whereas the control group received standard writing instruction without ChatGPT.

To measure the effects of the intervention, both groups completed a pretest, a posttest, and a delayed posttest over the course of the study. All tests were blind-rated by two professional reviewers. The scores were then analyzed using generalized estimating equations (GEE) to examine changes over time. Results indicate that the intervention was effective, with the experimental group outperforming the control group on both the posttest and the delayed posttest. The experimental group also demonstrated significant gains across all four IELTS writing criteria. Although some decline was observed in the delayed posttest, performance remained significantly above pretest levels. In contrast, the control group showed some improvement, but the differences in scores were not statistically significant. Overall, the findings suggest that structured and ethical AI training was associated with higher performance across all analytic writing criteria.

Ethics of AI-Mediated Language Education: Critical, Ecological, and Technomoral Perspectives
Chaoran Wang, Colby College
Ben Baker, Colby College

Ethical concerns surrounding generative AI in applied linguistics are often framed as responses to technological disruption, such as algorithmic bias, unfair assessment, or challenges to academic integrity (Dovchin, 2025; Voss et al., 2023). However, these seemingly unprecedented challenges are better understood as emerging at the intersection of longstanding tensions around multilingual practices and monolingualism, now reshaped through new forms of AI mediation.

In this presentation, we will first discuss how current ethical dilemmas around AI in language learning reflect a deeper systemic misalignment between rapidly evolving language practices and the institutional conventions through which learning, competence, and

authorship have been understood and evaluated. We specifically examine how AI unsettles inherited pedagogical and evaluative norms while exposing longstanding inequities, especially for multilingual learners.

We then propose a tripartite conceptual framework for ethical AI praxis in language education, drawing upon scholarship across applied linguistics and philosophy: (1) a critical and algorithmic justice perspective (e.g., Pennycook, 2021; Birhane, 2021), which provides a ground for questioning linguistic norms and judgements automated through AI, positioning AI as a critical site for negotiating normativity; (2) an ecological perspective (Godwin-Jones, 2025; Rietveld & Kiverstein, 2014), which situates AI as part of the semiotic and sociotechnical environment in which language learning emerges through dynamic relations among learners, linguistic and multimodal resources, technologies, tasks, and institutional contexts; and (3) technomoral virtue ethics (Vallor, 2024). We will conclude by discussing the framework's pedagogical implications and possibilities.

Plurilingual Graduate Students' Agency in AI-Assisted Academic Writing: A Multiple-Case Study

Anne-Marie Sénécal, Université de Sherbrooke, Canada

Technological innovations, particularly generative AI (GenAI), are reshaping language education and redefining academic writing (Barrot, 2023; Godwin-Jones, 2024). While GenAI tools can support plurilingual students by facilitating writing-related tasks (Shi et al., 2025), their increasing sophistication has intensified concerns about over-reliance on technology and its implications for learners' agency (Jacob et al., 2025). Scholars caution that when students offload academic tasks to AI, they may achieve desired outcomes without engaging deeply in the underlying writing processes (Fan et al., 2024), effectively using AI as a non-human agent in the writing process (Bandura, 2001). These issues are especially salient for plurilingual graduate students, whose academic writing experiences are shaped by emerging disciplinary expertise and complex linguistic and cultural backgrounds (Booth Olson et al., 2023).

This presentation establishes the conceptual and methodological coherence of a project that explores AI-assisted academic writing by plurilingual graduate students. I first situate the problem within current scholarship, highlighting the gap in research on how human-AI interaction shapes learner agency in academic writing. The second part elaborates the project's theoretical framework, integrating concepts of agency (Bandura, 2001), plurilingualism (Piccardo, 2020), and AI-assisted writing (Feuerriegel et al., 2024; OECD, 2025). The third part details the qualitative, descriptive, longitudinal, instrumental multiple-case study design used to investigate this issue.

The presentation concludes by outlining the broader implications for applied linguistics and writing pedagogy in higher education, recognizing GenAI as a potential step toward "linguistic democratization" in contexts where language remains entangled with power structures (Chapelle, 2025, p. 2).

Demographic Variation in Hedge Use in English Language Learners' Writing: A LLM-Based Metadiscourse Analysis

Tianyu Xu, Teachers College, Columbia University
Xinyu Xu, Teachers College, Columbia University

Metadiscourse, a pragmatic tool used to signal writers' stance and engagement, has been widely studied in writing performance, while little is known about how sociodemographic factors influence its development in English Language Learners' writing. Using the open-source ELLIPSE corpus, this study uses an LLM-assisted pipeline to investigate how sociodemographic factors (SES, grade, gender, and race) predict the use of hedges, an epistemic modality marker to express uncertainty about a proposition, among ELLs. Following a pilot validation comparing LLM-based annotation against human coding, the study analyzes two primary metrics: hedge density and categories (epistemic verbs, adverbs, adjectives, modal verbs, and approximations) and applies this automated discourse analysis to the corpus. Findings indicate that while demographics alone do not explain hedging variance, grade level serves as a significant main effect, suggesting a developmental shift in rhetorical strategies as students progress. Modal verbs and approximations emerged as the most frequent categories, with significant distributional differences implying higher student familiarity with these forms over epistemic verbs, adverbs, and adjectives. Consequently, learners require targeted practice to master more nuanced stance-marking. Additionally, hedge variety effectively captures differences in rhetorical range, necessitating instructional approaches that prioritize writing proficiency and individual style within students' specific cultural contexts. This study stresses that though the developmental nature of hedging justifies a scaffolded curriculum, the significant impact of grade level indicates that more proactive pragmatic instruction could empower ELLs to master a broader lexical range earlier.

Poster and Technology Demonstration Abstracts

Friday, September 25, 2026 12:30 - 01:25 pm

Posters

A Veces, Comes Agua: Measuring Hallucination of Idiomatic Expressions in a Language Learning Application

Timothy Farnsworth, CUNY Hunter College

Hallucination is a well-documented inherent property of large language models. In many domains, hallucinated output can be verified against external sources with relative ease. In L2 instruction, however, the learner is limited in their ability to evaluate the target language input they receive, making fabricated vocabulary, idioms, and usage patterns uniquely dangerous, as these may be internalized as authentic L2 knowledge. This risk is amplified in role-playing tutoring applications, where system prompts instruct models to act as tutors with specific regional identities and consequently generate idiomatic language that may extend beyond their

reliable knowledge. Preliminary review of output from Charlapren, an LLM-based language learning application, identified multiple suspected fabrications in Spanish instruction, including a nonexistent slang expression (comer agua - to eat water - as an idiom purportedly meaning to "wipe out" in surfing) and an invented English calque plausible enough to pass unnoticed by any learner.

This study estimates the rate of figurative hallucination across 300 simulated L2 Spanish lessons generated by three role-playing AI tutors representing Castilian, Argentine, and Puerto Rican varieties — a within-language comparison serving as a proxy for differential training data representation. Tutor output is screened by four flagship LLMs using structured criteria targeting fabricated collocations, nonexistent English calques, sense-boundary errors, and variety mismatches. Flagged items are verified against Spanish reference corpora and coded by native-speaker expert reviewers. The methodology adapts corpus-linguistic collocation analysis to LLM hallucination detection, producing both a hallucination rate estimate and an emergent typology of fabricated expressions in AI-mediated L2 instruction.

Are Human Teachers Becoming Optional in EFL Speaking Classrooms, or Are We Simply Placing a Native-Speaking Tutor in Every Learner's Hand?

Heejin Park, Gyeongang National University, Korea

As generative AI rapidly enters language education, its pedagogical role demands closer scrutiny. While previous studies report generally positive effects of AI on speaking proficiency, they have largely overlooked how its effectiveness may vary depending on task type and learner characteristics.

This study addresses this gap through a controlled experimental design involving 80 EFL university students. A 2 × 2 factorial design was employed, crossing two speaking task types—shadow reading (behaviorist, repetition-based) and conversation learning (constructivist, interaction-based)—with two instructional conditions: AI-assisted learning using ChatGPT and traditional instruction. Speaking proficiency was measured through pre- and post-tests focusing on accuracy, fluency, and vocabulary, alongside affective variables including confidence, motivation, and interest.

The findings reveal that AI effectiveness is not uniform but conditional. AI significantly enhances performance in structured tasks such as shadow reading, while its advantage is limited in conversational tasks. Moreover, lower-proficiency learners benefit most from AI-supported learning. In particular, AI also improves learners' confidence, a key affective factor.

These results suggest that AI does not replace teachers, but rather amplifies learning when pedagogically aligned with task structure and learner needs.

Automated Topic Detection in Spoken Language Assessment: Comparing an Instruction-Tuned Generative Model Against a Cross-Encoder Approach

Sabrina Sun, Language Testing International
Scott Gravina, Language Testing International
Kim Sallee, Language Testing International
Meg Malone, ACTFL

Automated scoring systems of spoken language assessments must consider both response quality and topicality. This is especially important in large-scale assessments such as the ACTFL Oral Proficiency Interview–computer (OPIc), which is used across diverse settings including business contexts in South Korea. However, the use of AI for scoring speaking assessments creates new vulnerabilities. Test takers may rely on memorized or generic answers (Knill & Gales, 2024; Malinin et al., 2017), raising their scores without reflecting true ability (Wang et al., 2019). The automatic detection of off-topic responses is thus key to maintaining score validity.

This study builds upon prior work on off-topic detection for the English OPIc in Korea by comparing an instruction-tuned generative model from the previous study against a new cross-encoder approach, both trained on an expanded dataset. Earlier work utilized balanced training and evaluation datasets consisting of 1,000 and 330 human-labeled OPIc responses (2020-2021) respectively. The present study expands these datasets to 2,866 for training and 1,444 for evaluation. The generative model, LLaMA 3.2 3B, is instruction-tuned and optimized using GRPO to classify topical alignment. The cross-encoder, DeBERTa-v3-base, is fine-tuned for pairwise prompt-response comparison. Results show that the cross-encoder outperforms the generative model, and we discuss possible explanations for this gap.

Designing a Language for Specific Purposes (LSP) Course in the Age of AI: Practical Strategies from a Course in Medical Spanish

David Ortega, Yale University

The increasing demand for and desirability of Medical Spanish skills in the United States among medical professionals underscores the importance of better integrating language training experiences within medical education (Lopez Vera et al, 2025; Ortega et al, 2020). This presentation introduces a Medical Spanish course tailored for students in professional schools, including medical, nursing, and physician assistant programs that considers student needs, interests, scheduling issues as well as disparate language abilities. The session offers general practical tips and actionable strategies to create efficient, relevant, and scalable language training for healthcare contexts that could be applied to various global scenarios.

The presentation will outline the course structure, key learning outcomes, and assessment methods, showcasing a model of potential interest to instructors interested in designing LSP courses.

Designing a Language for Specific Purposes (LSP) course for medical professionals requires aligning linguistic instruction with authentic clinical communication needs. This presentation examines challenges such as balancing medical accuracy with varied proficiency levels and integrating cultural competence despite time-constraining circumstances (Basturkmen, 2010; Long, 2005). It highlights how technology—such as simulation platforms, role-plays, and AI-driven conversational tools—can support situated learning and interactional competence (Chapelle, 2001; Godwin-Jones, 2018). Assessment is addressed through performance-based approaches, including scenario-based evaluations and tasks (Miller De Rutté et al., 2024).

Evaluating the Validity Arguments for a ChatGPT-Powered Dynamic Pragmatics Assessment

Gi Jung Kim, Iowa State University

Generative AI's interactive affordances are pushing the boundaries of the construct of language assessment, moving the meaning of a test score from a static proficiency to a dynamically growing proficiency. However, there has been little validation of scores obtained from generative AI-based dynamic language assessment. This study employs an argument-based approach to validate Talk Smart Trainer, a ChatGPT-powered dynamic pragmatics assessment that I developed for Korean young EFL learners. Four Korean middle school students completed two request role plays with Talk Smart Trainer and interviews. Experts, including Korean secondary school English teachers, professors of applied linguistics/TESL, and Ph.D. students in applied linguistics, were asked to evaluate the appropriateness of the rating scale, rate students' performance and provide feedback based on the rating scale, and judge the relevance of Talk Smart Trainer's feedback to the pragmatic competence. The findings show that the evaluation inference was supported by the expert review of the rating scale. The generalization inference was partially supported by the inter-rater dependability between Talk Smart Trainer and human raters. The explanation inference was supported by the expert review of the relevance of Talk Smart Trainer's feedback. The utilization inference was partially supported by the percentage of feedback partially addressed by students. The consequence implication inference was supported by the discursive analysis of students' role-play performance based on Vygotsky's sociocultural theory and the interview data. The present study represents meaningful progress in using generative AI in a grounded manner that measures and promotes students' learning of the target construct.

GenAI in a Language Teacher Education Practicum: A Case Study in ESP

Amanda Brown, Syracuse University

English for Specific Purposes (ESP) instruction requires attention to learners' professional contexts and communicative needs (Basturkmen, 2019), and ESP in airline cabin services is argued to present unique pedagogical challenges (Cho and Choe, 2024). Language teacher education programs rarely provide preparation for ESP teaching (e.g. absence in Tajeddin & Farrell, 2025), leaving pre-service teachers underequipped for the field. Finally, scholarship also notes limited research on GenAI integration in language teacher education (Kerr & Kim, 2025; Kohnke et al., 2023; Moorhouse et al., 2024), despite its professional potential (Weng & Fu, 2025).

This case study investigates how pre-service language teachers leveraged GenAI during an ESP teaching practicum focused on

airline cabin services. Drawing on reflective journals, teaching artifacts, and observations, the study identifies how pre-service teachers employed GenAI to support needs-based language analysis, materials development, and activity design in this challenging domain. Benefits included broad uptake of the technological tools, accelerated materials creation, increased teacher confidence, and the ability to produce diverse and context-specific resources. Challenges included inaccuracies from insufficient proofreading and lack of personal knowledge, as well as limited task variety. Findings highlight both the promise and pitfalls of integrating GenAI into language teacher education for ESP contexts.

"I Wish Anika Would Contribute More": Designing Multimodal AI Characters for Learning-Oriented SBLA

Soo Hyoung Joo, Teachers College, Columbia University

Recent advances in genAI have prompted growing interest in their applications for second language learning. Research on AI in language learning has focused largely on performance outcomes in text- and speech-based systems (Law, 2024; Pipia & Gurgenishvili, 2025), leaving the design and interactional effects of multimodal AI characters enabled by text-to-video technologies underexplored. Unlike text-only or voice-based systems, these characters integrate visual embodiment (e.g., gender and race) and auditory features (e.g., accent, speech rate), presenting human-like cues, while their interactional design (e.g., roles, collaboration style) governs social context.

Drawing on the Learning-Oriented Language Assessment (LOLA) framework (Purpura, 2024), this exploratory mixed-methods study examines how multimodal AI character design shapes the social-interactional dimension of a scenario-based language assessment (SBLA). High-school learners were presented with AI characters created using text-to-video platforms (e.g., Synthesia; HeyGen), which varied in visual embodiment, accent, and collaboration style. Survey data on perceived collaboration, trust, and engagement, along with learners' self-reported preferences, were integrated with interviews and open-ended responses.

Findings indicate that collaboration style and accent were the most salient factors shaping engagement. Visual identity cues contributed to how learners perceived relatability and social presence, influencing their willingness to engage with the characters. Analyses of learner preferences further revealed that alignment between expected and enacted collaboration styles supported smoother interaction, whereas mismatches contributed to perceptions of imbalance, limited contribution, and interactional friction.

These findings inform a preliminary framework for multimodal AI character design and prompt engineering, highlighting how visual, auditory, and interactional features can shape the social-interactional dimension.

Know a Word by Its Neighbours: Teaching Vocabulary Through LLM-Assisted Lexical-Semantic Grounding

Büşra Marşan, Stanford University
Ecem Fırat Kopuz, CUNY, City University of New York

This study tests whether vocabulary learning improves when words are taught through structured semantic relations rather than in isolation or through thematic grouping, and whether large language models (LLMs) can support this instructional design. We define lexical-semantic grounding as teaching a word within a local semantic network including synonyms, antonyms, and multiple senses. This builds on work showing that semantic connectivity and deeper processing support more durable lexical representations (Collins & Loftus, 1975; Craik & Lockhart, 1972), and on research emphasizing meaningful lexical relationships in vocabulary learning (Nation, 2022; Schmitt, 2008; Webb, 2007).

Participants are Turkish learners of English. All groups are taught the same 10 target words alongside 10 filler words within a guided lesson. Instructional time, number of items, and examples are controlled across groups. In the control condition, target words are taught individually with definitions, and filler words are unrelated. In thematic condition, target and filler words share a broad topic (e.g., travel), without explicit semantic links. In lexical-semantic grounding condition, target words form a structured network through synonymy, antonymy, and teaching of multiple senses of each target word; filler words are selected to support these relations.

Instructional materials for this condition are generated with LLM assistance and curated prior to teaching; teaching procedures are otherwise identical across groups.

Learning is measured via pretest, immediate and delayed post-tests. Data will be analyzed using mixed-effects models. We expect lexical-semantic grounding to yield stronger retention, particularly over time. This study evaluates how LLM-generated semantic structure can support vocabulary learning.

Preserving Voices through Cross-Cultural Game Design: An LLM-Assisted Multimodal Pedagogical Approach

C. Joseph Sorell, Purdue University
Lixia Cheng, Purdue University
Srilani Wickramasinghe, Purdue University
Yutika V. Sawant, Purdue University

This work-in-progress study investigates the use of large language models (LLMs) in cross-cultural game design as a multimodal pedagogical approach to enhancing language and intercultural competence. Participants from diverse linguistic and cultural backgrounds composed personal narratives and integrated them into a shared metanarrative, using an LLM to support brainstorming, drafting, and iterative refinement (Godwin-Jones, 2023; Mittal et al., 2024). The resulting texts were embedded in interactive

puzzle-based games, stimulating discussion and analysis of behaviors and cultural values through multimodal content (Reinhardt, 2019).

The project explores how LLMs support game design by providing access to linguistic and cultural knowledge—ranging from specialized vocabulary (e.g., Quechua morphology) to culturally grounded practices (e.g., dietary systems)—that would otherwise require extensive study or expert informants. At the same time, it examines risks of AI-generated content obscuring individual authorial voices and introducing inaccuracies (Bender et al., 2021). Participants negotiated these tensions through iterative prompt-engineering and critical evaluation of AI outputs.

Preliminary observations suggest that LLM integration in game design fosters deep engagement, sophisticated language production, and meaningful intercultural dialogue. This study contributes to AI for Applied Linguistics by illustrating how LLM-assisted, multimodal, collaborative game design supports language instruction and the preservation of diverse linguistic and cultural voices.

Using Generative AI for Speaking Practice in Language Education

Fatemeh Soleimani, University of Texas at Austin

This poster presents a classroom experience using LLMs as conversational tools in an intermediate undergraduate French course (A2–B1) at a public university in the United States. As AI systems increasingly shape how language is produced and practiced, this project explores how learners interact with AI-generated language within real classroom communication. During speaking activities, students used their preferred chatbots (such as ChatGPT or Claude) to ask for alternative ways to express their ideas in French. They then tested these suggestions in paired and small-group conversations and discussed whether the expressions were natural, appropriate, or meaningful in context.

AI output was treated as tentative rather than correct. Students were encouraged to notice pragmatic problems, register mismatches, and unclear meanings, and to ask follow-up questions or revise their speech with peer and instructor support. This created a multimodel learning environment where AI-generated text, spoken interaction, and human judgment worked together.

Based on classroom observations, teaching materials, and student reflections, the project shows how AI limitations can support pragmatic awareness and communicative competence. The poster contributes to applied linguistics research by illustrating how LLMs can be integrated into speaking instruction while sustaining human interaction, discourse negotiation, and meaning-making in AI-mediated language classrooms.

Technology Demonstrations

A Mediation Environment for Concept-Based Vocabulary Learning in Second Language Learners

Yuwei Xia, Pennsylvania State University

This demonstration presents a multimodel AI vocabulary learning environment that operationalizes Concept-Based Language Instruction (Lantolf & Poehner, 2014) for second language learners. The system mediates not only individual words but the concept of vocabulary knowledge, Form, Meaning, and Use (Nation, 2013), through which learners build a richer understanding of what it means to know a word and reorganize their mental lexicon.

A learner query first elicits prior knowledge or prompts recall on re-encounter, then opens a 3×3 conceptual matrix: Form (word parts, morphologically related words, word family), Meaning (core meaning, dictionary senses, associations), and Use (example sentences, usage patterns, collocations), each addressed through three complementary modules unlocking progressively. Each step scaffolds a mediation cycle in which the learner exercises agency over pace, depth, and language of engagement. An opt-in AI assistant expands any card on demand, offering explicit clarification or eliciting productive use.

A multimodel pipeline combines two LLM streams, a POS tagger, and a Postgres-backed lexical RAG layer keyed on lemma and POS. Example sentences are retrieved from POS-tagged COCA fragments across five registers, not LLM-generated. Multimodal input and output—text, TTS audio, and register/POS labels—support diverse processing routes.

Attendees will query words and observe the mediation cycle.

Evalora: A Voiced-Avatar Examiner for L2 French — What It Scores, What It Cannot

Imane Harbi, Sorbonne University Abu Dhabi, United Arab Emirates

This technology demonstration presents Evalora, a voiced-avatar chatbot designed to simulate a French as a Foreign Language (FFL) institutional speaking exam at Sorbonne Université Abu Dhabi. Evalora operationalizes an 11-criterion rubric aligned with DELF standards, calibrated to intelligibility norms rather than native-speaker benchmarks, in a multilingual context where Arabic and English are the dominant first languages, alongside Russian, Ukrainian, and Armenian.The demonstration will showcase Evalora's three-phase architecture: a monitored monologue, an avatar-led debate with dynamic relaunches, and automated criterion-by-criterion feedback. Four avatar examiners with distinct interactional styles modulate evaluative pressure across sessions. The core argument of this demonstration is not that automation succeeds, but where and why it fails, and how engineered workarounds can reduce the gap. Factual criteria such as source identification transfer readily to automated verification. Others, including fluency, spontaneity, and gaze interpretation, resist despite sustained calibration. Proxy indicators, simplified heuristics, and score transparency were developed to partially bridge this gap, revealing what remains irreducibly human in speaking assessment.

Evalora functions as a live case study in instructor-led AI co-design for language assessment, raising empirical questions about validity, transparency, and the role of pedagogical judgment in automated scoring.

Talio Language Tech Demo

Matthew Wise, Talio Language

Talio Language is a multimodel L2 learning tool focused on conversation practice. Conversation is a primary bottleneck in rapid L2 advancement--without good, regular conversation practice, students fossilize bad habits or stop progressing altogether.

The demo focuses primarily on the four aspects that set Talio apart:

* Conversation naturalness. A speech-to-speech model drives live conversation--not the typical ASR→LLM→TTS chain. Beta users consistently call out how natural the interaction feels; even skeptics end up talking far longer than they expected.

* NLP-powered feedback. Multiple NLP techniques identify where students are struggling, and pedagogical principles drive when and how to deliver feedback.

* AI-enhanced exercise creation. Teachers can sketch a rough scenario and let our LLM-driven scenario design handle the prompt engineering--they get a working exercise in seconds. They can also write detailed prompts themselves. Either way, teachers keep final control.

* Complex language support. We started with Mandarin and solved complex issues like text alignment, sandhi, and tone contour analysis that aren't handled by other companies.

Together, these technologies address the conversation bottleneck by providing students with high-quality conversation practice and supporting teachers in facilitating and analyzing student conversations.

WriteMed: An AI Mediator for Dynamic Assessment of L2 Academic Writing

Yiwei Qin, University of Pennsylvania
Yuwei Xia, Pennsylvania State University
Junjing Zhang, Stanford University

This demonstration presents WriteMed, an AI-mediated writing instruction system that operationalizes Dynamic Assessment (Poehner, 2008) for second language learners working on academic discussion essays. Unlike systems that primarily score, correct, or rewrite student work, WriteMed positions AI as a mediator: teaching and assessment occur within the same revision loop, and learner ability is diagnosed through responsiveness to graduated mediation rather than final-draft quality alone.

After a learner submits a full draft, the system scores and diagnoses the essay across three writing focus areas: content, organization, and language use. It then selects one focus per revision round to avoid feedback overload. Within each round, mediation proceeds through graduated hint levels, beginning with open questions, escalating to options or analogies, and finally offering explicit revision direction. The system compares the previous and current drafts to judge whether the focus has been addressed; if so, it auto-advances to the next round.

An orchestrated multi-agent architecture coordinates global scoring, focus selection, dialogic mediation, and revision-uptake judgment. All learner submissions, hints, AI judgments, and observed edits are persisted, supporting research on revision behavior, feedback uptake, and AI explainability.

Attendees will observe a complete mediation cycle from initial draft to final diagnosis.

Saturday, September 26, 2026 04:15 - 05:10 pm

Posters

A Testable Multimodal AI Framework for Replication of Listening-to-Write Research in English for Academic Purposes (EAP)

John Bandman, Lancaster University, UK

Integrated listening-to-write tasks are central to English for Academic Purposes (EAP), requiring coordination of multiple linguistic and cognitive processes. This presentation revisits a completed PhD study that examined the effects of task repetition and feedback on EAP students' writing performance and demonstrates how such a study could be systematically replicated using AI. The original study, conducted without AI, involved 64 university students completing repeated listening-to-write tasks, with one group receiving teacher feedback between performances. Results showed improvements in complexity, accuracy, and fluency (CAF), as well as in listening-input summarization and knowledge transfer, with reduced trade-off effects over time and limited impact of teacher feedback.

This presentation does not report an AI-based replication; rather, it proposes an empirically testable and operationalizable framework for AI-mediated replication using currently available AI tools. Specifically, it outlines how multimodal AI systems could be used to (1) generate and adapt listening-to-write tasks from audio and textual inputs and (2) provide scalable feedback through large language models, using measures aligned with CAF constructs and discourse-level features.

By explicitly mapping a traditional SLA study design onto contemporary AI capabilities, this work contributes a replicable and empirically testable model for research, pedagogy, and assessment, while raising considerations regarding validity, reliability, and responsible AI use in applied linguistics.

Comparing the Effects of AI-Generated and Teacher-Delivered Formative Assessment on EFL Learners' Writing Performance and Writing Self-Efficacy

Golnoush Haddadian, Georgia State University
Nooshin Haddadian, Georgia State University
Saba Soleimani, Interdisciplinary Transformation University IT:U, Austria

Although the influence of AI-generated and teacher-delivered feedback on writing has been the subject of much research, less is known about the comparative role of AI-generated and teacher-delivered formative assessment (FA) on EFL writing and self-efficacy. This quasi-experimental pretest-posttest study examines this issue. The participants, comprising 68 Iranian upper-intermediate learners, were selected from an initial pool of 131 students based on Oxford Quick Placement Test (OQPT). They were selected through convenience sampling, divided into two groups. Before interventions, the two groups received a writing pretest and a writing self-efficacy questionnaire (WSEQ). Next, the first experimental group (N = 33) received teacher-delivered FA based on three strategies of FA including a) determining clear learning goals and success criteria, b) specifying evidence of current writing performance; and c) providing timely, specific, and actionable feedback. The second experimental group (N = 35) was exposed to FA implementing the same strategies via ChatGPT. Both groups practiced 15 randomly selected topics from a pool of 60 with the same teacher. After a 15-session treatment, both groups completed a writing posttest and the WSEQ. The results of independent samples t-test indicated that AI-generated FA significantly improved writing performance compared with teacher-delivered FA. Conversely, one-way ANCOVA results demonstrated teacher-delivered FA had a more significant effect on writing self-efficacy compared with AI-generated FA. It is suggested that EFL teachers consider integrating AI-generated and teacher-delivered FA to potentially capitalize on the merits of both teacher and AI technology in enhancing EFL learners' writing performance and self-efficacy beliefs.

Designing a Digital Diagnostic Writing Assessment for Post-IELTS EAP Learners

Yoohee Kim, Centre for Applied Language Studies (CALS), University of Jyväkylä, Finland

A digitally delivered diagnostic writing assessment is designed for advanced English learners enrolled in an English for Academic Purposes (EAP) course at one small institution. Grounded in constructivist theory and communicative competence models, the assessment aims to bridge students' prior learning in grammar, vocabulary and writing with the academic demands of university-level writing. A dual evaluation system integrates AI-powered and human judgement: surface-level features (e.g., grammar, vocabulary, punctuation) are assessed by AI, while deeper discourse-level elements (e.g., coherence, specificity, tone) are scored by human raters. A Boolean checklist and a 0-4 analytic rubric are employed to enhance inter-rater reliability and to identify leaner's strengths and weaknesses in writing. The assessment task reflects real-world writing demands by incorporating a reading-into-writing prompt

requiring students to produce structured essays. Feedback is aligned with instructional resources used in the institution to support a sustainable assessment approach that informs pedagogical instruction and promotes long-term academic writing development. Assessment quality is evaluated through the VRIP-Q framework.

Developing an AI-Powered Writing Assessment Platform for Portuguese as an Additional Language: Leveraging the CorCel Corpus

Juliana Schoffen, Federal University of Rio Grande do Sul, Brazil
Elisa Stumpf, Federal University of Rio Grande do Sul, Brazil
Deise Amaral, Federal University of Rio Grande do Sul, Brazil
Tanara Kuhn, University of Coimbra, Portugal
Larissa Goulart, MontClair State University
Luiza Divino, Federal University of Rio Grande do Sul, Brazil
Isadora Hanauer, Federal University of Rio Grande do Sul, Brazil
Amanda Raupp, Federal University of Rio Grande do Sul, Brazil
Brenda Xavier, Federal University of Rio Grande do Sul, Brazil
João Victor Pacheco, Federal University of Rio Grande do Sul, Brazil
Gabriel Pazinatto, Federal University of Rio Grande do Sul, Brazil
Rafael Nunes, Federal University of Rio Grande do Sul, Brazil
Dennis Balreira, Federal University of Rio Grande do Sul, Brazil

This paper presents an ongoing project developing an interactive web platform for the automated assessment of written proficiency in Portuguese as an Additional Language (PAL), grounded in the theoretical construct of Celpe-Bras — Brazil's official PAL proficiency exam. The platform draws on the CorCel corpus, a pioneering corpus comprising over 15,000 texts produced under exam conditions across four Celpe-Bras editions, rated on a 0–5 holistic scale and organized by proficiency level and task type. To date, two main tools have been developed: 1) CorSpell, a customizable spelling normalization tool that supports corpus-based analysis preventing the artificial inflation of type-based indices, and 2) an automated essay scoring (AES) system trained on CorCel data, resulting from multi-architectural experiments involving machine learning algorithms, neural models, and LLMs, which classifies user-submitted texts according to Celpe-Bras proficiency levels. As a next step, the project will develop a feedback module translating AES classifications into pedagogically meaningful guidance, highlighting strengths and areas for improvement. By anchoring automated scoring in a task-based, genre-oriented construct, the platform addresses a key gap in AI-supported language assessment for under-resourced languages, informing exam preparation, rating scale refinement, construct validation, and the development of adaptive pedagogical tools for PAL learners and teachers.

Divergent Policies, Shared Goals: Understanding ELL Students' AI-Mediated Academic Writing Across Policy Contexts

Nathan Schrader, Fordham University
Xiaole Shan, Fordham University

As AI tools become increasingly prevalent in higher education, debate has emerged over their role in academic settings. Sullivan et al. (2023) characterize this as a tension between "banning vs. embracing," with initial reactions often framing AI as a threat to academic integrity while others argue that prohibition fails to prepare students for an AI-integrated workplace. This unresolved tension has resulted in widely divergent AI policies among instructors, leaving English Language Learner (ELL) students to navigate inconsistent expectations when using AI to support academic writing. This mixed-methods study investigates how ELL students use AI for writing development under these varying policy conditions. Participants were asked to complete a multidimensional survey assessing affective, behavioral, cognitive, and ethical dimensions of AI literacy. Semi-structured interviews captured students' experiences with AI-mediated writing processes, and pre- and post-AI drafts along with corresponding chat histories were collected to trace student-AI interaction during revision. Quantitative survey data revealed patterns in how students conceptualize AI across policy contexts, while qualitative analysis of interviews, chat logs, and draft revisions illuminated the strategies students employed and how interactions shaped their writing. Findings contribute to responsible AI integration in language education by grounding policy recommendations in empirical evidence of student behavior.

Exploring EFL Doctoral Students' Use of ChatGPT in Higher-Order Academic Writing Through a Self-Regulated Learning Lens: A Pilot Qualitative Study

Erpin Said, University of Rochester

This poster presents a pilot qualitative study exploring how EFL doctoral students use ChatGPT in higher-order academic writing through a Self-Regulated Learning (SRL) lens. Prior research on ChatGPT in academic writing has often emphasized outcomes, short-term tasks, or general perceptions, with less attention to how doctoral writers describe planning, monitoring, and evaluating tool use across sustained writing tasks. The pilot primarily draws on semi-structured interview data, with participant-provided screenshots of ChatGPT interactions as supporting artifacts. Using episode-level analysis, the study examines how participants described using ChatGPT during brainstorming, outlining, drafting, revising, and source checking. Preliminary analysis suggests that participants used ChatGPT selectively to expand ideas, improve organization, and review alignment with task expectations, while maintaining authorial control by comparing, verifying, and rejecting unsuitable outputs. Framed through SRL, these patterns highlight planning, monitoring, and evaluative decision-making in AI-mediated academic writing rather than simple reliance on text generation. As a pilot study, this work offers process-oriented insight into ChatGPT-supported writing and helps refine the analytic direction of the broader dissertation project.

Operationalizing AI Literacy in Multilingual Writing: A Curriculum-Integrated Pedagogical Design

Gözde Durgut, University of Arizona

This paper examines how AI literacy can be operationalized through pedagogical design in multilingual first-year writing courses. While existing research has proposed conceptual frameworks of AI literacy, there remains limited work demonstrating how these frameworks can be translated into classroom practice, particularly in writing-focused, non-STEM contexts.

Grounded in Ng et al.'s (2021) AI literacy framework and informed by an ecological perspective, this paper presents a curriculum-integrated AI literacy module embedded within a genre-based writing curriculum. The design aligns AI literacy domains with core writing outcomes, including rhetorical awareness, critical thinking, revision, and reflection. The module incorporates a structured sequence of conceptual, applied, multimodal, and ethics-focused activities, integrated into existing course units and assignments.

The instructional design engages students with multiple AI tools, including text-based systems as well as AI-supported research and multimodal tools, and situates AI use within ongoing composing practices. Rather than treating AI literacy as a set of abstract competencies, this approach frames it as a set of observable practices that emerge through structured classroom engagement and discussion of AI-related discourse as a focus of critical inquiry in the classroom.

This paper contributes to applied linguistics by demonstrating how AI literacy can be operationalized through curriculum-integrated design in writing instruction. It positions AI literacy not as a set of skills, but as practices that develop through writing, revision, and reflection. The paper offers a pedagogical model for integrating AI into writing curricula that supports critical engagement, maintains student agency, and situates AI use within broader academic and social contexts.

Position, Design, Enact: A Framework for Critical GenAI Pedagogy in Language Education

Xinyue Lu, Howard University
Zhongfeng Tian, Rutgers University–Newark

Generative AI (GenAI) tools are increasingly entering language classrooms, yet educators still lack pedagogically grounded frameworks for deciding how these tools should be used in ways that support language development while addressing bias, ethics, and instructional responsibility. Existing discussions of AI in education often emphasize efficiency, productivity, or technical capability, but these orientations do not fully address the social, contextual, and meaning-centered goals of language teaching. This paper proposes a Position–Design–Enact (PDE) framework for critical GenAI pedagogy that centers human judgment and teacher decision-making. The framework highlights four areas of critical awareness that should guide classroom use of GenAI: language ideologies, representation and bias, authority and agency, and ownership and ethics. We illustrate the framework through two classroom-based examples: a secondary world language classroom in which GenAI is positioned as a writing coach, and a K–5

bilingual classroom in which GenAI supports the design of interpretive assessment tasks. Across these examples, we show how PDE helps educators make more deliberate decisions about the role of GenAI in classroom ecologies, the design of learning tasks, and the enactment of AI-supported instruction in practice. We argue that the PDE framework offers a useful model for applied linguistics scholars and educators seeking to engage multimodal GenAI critically while maintaining clear instructional goals and commitments to equitable language teaching.

Trait-Based Automated Essay Scoring in Rater Training for the Evaluation of Young English Language Learner's Writing

Aubrey Sahouria, Center for Applied Linguistics
Yu Lan Su, Center for Applied Linguistics
Frank Wuckinski, Center for Applied Linguistics
Joshua Smith, Center for Applied Linguistics
Hyeonseong Lee, Center for Applied Linguistics, University of Maryland

Trait-based automated essay scoring (AES) has been integrated into several high-stakes language proficiency assessments to expedite scoring and verify human ratings for writing. Both AES engines and human raters require extensive training to reliably assign scores across multiple traits, especially when assessing younger English language learners' (ELLs; grades 3-5) writing skills, which demand developmentally sensitive evaluation. Assessing language proficiency in young children presents unique challenges due to developmental variation in attention, cognition, emotional regulation, and verbal expression. Young children may have limited attention spans, fatigue quickly, or struggle with unfamiliar testing formats. Training raters for this context requires careful review and selection of benchmark writing samples across a range of criteria, which can be resource-intensive, especially when accounting for young ELL's developing skills. Recent advances in population-specific, trait-based AES methods present an opportunity to streamline the selection of representative writing samples for rater training, potentially reducing resource-intensity and diversifying training options. This literature review explores and summarizes the utility of tailored, trait-based AES for the selection of rater training writing samples in high-stakes English language proficiency assessments for younger learners.

Technology Demonstrations

Bruhh.ai: An AI-Based System for College Advising

Prince Bashangezi, Pomona College
Feng Xiao, Pomona College

Bruhh.ai is an AI-based system designed to address complex and time-consuming advising challenges and support college students and advisors. It offers three core services: academic planning, course scheduling, and personalized academic and career advising. Using large language models, retrieval-augmented generation, active search, and institutional academic data, Bruhh.ai delivers

context-aware recommendations tailored to each student's degree requirements, academic history, scheduling constraints, and long-term goals. It translates complex university systems into accessible, natural-language guidance, helping students navigate curriculum pathways, understand graduation requirements, and make informed, career-oriented academic decisions. The system features personalized course planning, prerequisite tracking, adaptive schedule generation, requirement completion analysis, and AI-assisted academic and career exploration. It also incorporates explainability and contextual retrieval mechanisms to enhance the reliability and transparency of its recommendations. This work contributes to ongoing discussions in AI-enhanced educational interaction by examining how AI can improve comprehension, accessibility, and decision-making in higher education. Bruhh.ai demonstrates the potential of AI architectures to strengthen student support systems while emphasizing the importance of responsible and explainable AI in educational contexts.

Luduan.ai: A System Solution to AI-Enhanced Pedagogy

Feng Xiao, Pomona College
Kun Nie, Pomona College

This technology showcase introduces Luduan.ai, an integrated platform for AI-enhanced language teaching and learning. The system streamlines instructional workflows while providing individualized, real-time support for learners across 17 languages. Luduan.ai integrates task design, automated feedback, assessment, and progress tracking within a unified interface.For instructors, the platform reduces time spent on grading, feedback, and materials development through AI-assisted evaluation aligned with pedagogical criteria. It also offers analytics dashboards that highlight learner performance patterns, enabling more targeted and data-informed instruction. For learners, Luduan.ai delivers adaptive support based on proficiency level, task type, and communicative context, supporting personalized practice and continuous improvement. Attendees will engage with interactive examples illustrating how the system supports classroom instruction, formative assessment, and self-directed learning.

OUTREACH: Using AI to Shorten the Distance Between School Leaders and Stakeholders

Shawn Filer, School Board and Youth Engagement Lab
Jonathan Collins, School Board and Youth Engagement Lab

Even in the largest school district in the country, parent engagement is limited and ad hoc. All major discussions happen in English and in the evening, making participation in the community comment section practically impossible for hundreds of thousands of working-class parents. In the NYCDOE, where 40% of students are multilingual and 150 languages are spoken, the "language tax" results in massive fiscal waste: districts spend taxpayer money on programs that stakeholders never asked for and don't care about because they were never truly in the room. Professor Collins's time on the NYCDOE Panel for Education Policy has made this gap glaring.

This demonstration presents OUTREACH, a community dialogue platform designed to bridge this gap by enabling real-time, bidirectional critical engagement. We move beyond static, one-way translated flyers to live, nuanced dialogue. By allowing parents to

converse with other parents about school policy and budget priorities in their native language in real time, OUTREACH transforms the community comment section from a performative English-only ritual into a scalable tool for accountability. We will demonstrate how this technology allows school leaders to stop guessing what their community needs and start listening, regardless of the language spoken or the time of day.