The 2nd NUS Symposium on T&I Evaluation

Advancing translation and interpreting quality assessment: Past lessons, transferrable insights, and future directions

21-22 May 2026

AS8 Level 5 05-49
National University of Singapore
10 Kent Ridge Crescent
Singapore 119260

Organized by Department of Chinese Studies
Faculty of Arts and Social Sciences
National University of Singapore

NUS National University of Singapore | Faculty of Arts & Social Sciences

Day 1: 21 May 2026
Morning session

8:30-8:40 Opening remarks

- Han Chao, National University of Singapore

8:40-9:40 Keynote speech 1

- Measurement, reception, and the space between: Interpreting assessment reconsidered
- Liu Min-hua, National Taiwan University / Hong Kong Baptist University

9:40-10:30 Thematic presentation 1

- The construct validity of analytic rubrics in interpreting: Linking ratings to performance indicators
- Chen Sijia, Guangdong University of Foreign Studies

Break 10:30-10:50

10:50-11:30 Individual presentation 1

- Rethinking rating quality in interpreting assessment: The role of rater experience and reference texts
- Chen Shirong, University of International Business and Economics

11:30-12:10 Individual presentation 2

- Measuring rater cognition in interpreting assessment: Development and validation of the Raters' Scoring Process Scale
- Jiang Mengting, Xiamen University

Lunch 12:10-14:00

Afternoon session

14:00-14:50 Thematic presentation 2

- Dependency distance, task complexity, and interpreting performance: A quantitative linguistic approach to interpreting assessment
- Jiang Xinlei, Xi'an Jiaotong University

14:50-15:30 Individual presentation 3

- Modeling fluency as process: A temporal-sequential analysis of interlingual interpreting performance
- Jiang Zhaokun, National University of Singapore

Tea break 15:30-16:00

NUS National University of Singapore | Faculty of Arts & Social Sciences

16:00-16:40 Individual presentation 4

- **Exploring the predictive validity of a Rasch-calibrated integrative test for interpreting aptitude: Rediscovering the SynCloze**
- **Liu Yubo, Beijing Foreign Studies University**

16:40-17:20 Individual presentation 5

- **Measuring washback on learning in interpreting: Development and validation of the CATTI Self-directed Preparation Strategy Scale**
- **Kang Yuhan, Xiamen University**

Group dinner 17:20-19:30

Day 2: 22 May 2026
Morning session

8:30-9:20 Thematic presentation 3

- **Leveraging pretrained language models for the automated assessment of English-Chinese interpreting: Explorations, advances, and reflections**
- **Han Chao, National University of Singapore**

9:20-10:00 Individual presentation 6

- **An interdisciplinary approach to notetaking research: What can consecutive interpretation and higher education researchers learn from each other?**
- **Rebecca Yeager, University of Illinois Urbana-Champaign** (via Zoom)

Break 10:00-10:20

10:20-11:00 Individual presentation 7

- **Beyond one-turn prompting: Leveraging self-reflection and multi-agent interaction in assessing students' consecutive interpreting**
- **Lu Xiaolei, Xiamen University**

NUS National University of Singapore | Faculty of Arts & Social Sciences

11:00-11:40 Individual presentation 8

- **Machine vs. human: An exploratory study of Chinese-Portuguese classical poetry translation quality assessment**
- **Jiang Lili & Han Lili, Macao Polytechnic University**

11:40-12:20 Individual presentation 9

- **Investigating student interpreters' feedback-seeking processes from GenAI: An exploratory study using behavioral sequence analysis**
- **Chen Qionglu, Xiamen University**

Lunch 12:20-14:30

Afternoon session

14:30-15:10 Individual presentation 10

- **Diagnosing language-pair bias in quality assessment of speech translation using a multi-agent AI framework**
- **Wang Xiaoman & Claudia Angelelli, Heriot-Watt University** (via Zoom)

15:10-16:10 Keynote speech 2

- **Automated speaking assessment: Development, implementation, and implications**
- **Yan Xun, University of Illinois Urbana-Champaign**

Tea break 16:10-16:30

16:30-17:30 Fireside chat

- **Fireside chat with Prof Liu Min-hua and Prof Yan Xun**
- **Moderator: Han Chao**

Group dinner 17:30-20:00

NUS National University of Singapore | Faculty of Arts & Social Sciences

Keynote speech 1

Measurement, reception, and the space between: Interpreting assessment reconsidered

Liu Min-hua
National Taiwan University / Hong Kong Baptist University

Abstract

Assessment of interpreting quality has made considerable technical progress in recent years. However, its foundational question is left largely unexamined: quality for whom, and to what end? In this talk, I will trace the field's persistent tensions — between reliability and validity, between output measurement and audience reception — and ask what 50 years of research has established. Against this backdrop, automatic assessment is positioned not as a solution to the measurement problem but as a redistribution of it: advancing what can be assessed objectively while making more visible what cannot. The emerging prospect of automatic interpreting further intensifies the challenge: when machines interpret, norms formed around human performance offer only partial guidance. The talk concludes with a proposal: for conference interpreting, output-based criteria remain the practical reality, but the epistemological stance toward them must change. Beyond conference interpreting, where communicative purpose is more often institutionally defined, quality criteria should be grounded in evidence of what users actually receive.

Bio

Liu Min-hua (Ph.D., The University of Texas at Austin) is an Adjunct Professor in the Graduate Program of Translation and Interpretation at National Taiwan University and an Honorary Research Fellow of the Centre for Translation, Hong Kong Baptist University. Her research covers interpreting cognition, bilingualism, and interpreting assessment, supported by several government grants. She currently serves as co-editor of the journal *Interpreting* and sits on the advisory boards of several journals and book series in translation and interpreting studies.

Keynote speech 2

Automated speaking assessment: Development, implementation, and implications

Yan Xun
University of Illinois Urbana-Champaign

Abstract

This presentation surveys recent developments in automated speaking assessment and considers how artificial intelligence (AI) is reshaping multiple stages of the assessment process. From the generation of speaking tasks, through the scoring of spoken responses, to the provision of feedback, I argue that the key to automated speaking assessment is not the development and improvement of individual AI systems, but how they relate to one another in the quality control system. In addition, the integration of AI technology needs to be evaluated against the broader questions of validity, fairness, and sustainability. As such, while technical reliability is the starting point of automated speaking assessment, ongoing efforts must be spent on construct representation, task authenticity, score interpretation, and the role of human oversight in automated speaking assessments. The talk concludes by discussing the challenges and opportunities of current approaches and outlining priorities for future research and practice in automated speaking assessment.

Bio

Yan Xun is a professor of Linguistics, SLATE, Educational Psychology at the University of Illinois Urbana-Champaign. He is a faculty member at the Beckman Institute for Advanced Science and Technology. His research interest include writing and speaking assessment, psycholinguistic and computational approaches to language testing, and language assessment literacy. His work has been published in *Language Testing*, *Language Assessment Quarterly*, *Language Learning*, *Studies in Second Language Acquisition*, *Applied Linguistics*, *TESOL Quarterly*, and *System*. He is the co-editor for *Language Testing*.

Thematic presentation 1

The construct validity of analytic rubrics in interpreting: Linking ratings to performance indicators

Chen Sijia
Guangdong University of Foreign Studies

Abstract

Interpreter performance is increasingly evaluated via multi-dimensional scales; however, the extent to which individual facets capture distinct underlying constructs warrants closer empirical scrutiny. This study examines how subscales of analytic rubrics relate to independently derived performance indicators. Using a three-dimensional rubric as a focal case, we investigated how subscale scores — Information Completeness (Information), Fluency of Delivery (Fluency), and Target-Language Quality (Language) — correspond to objective measures of accuracy and fluency. The dataset comprised 504 performances from 102 participants across 36 stimuli. Subscale ratings were modelled as a function of propositional scores and a suite of fluency indicators, including manually coded disfluencies and automatically extracted speech measures. Linear mixed-effects models estimated Indicator × Subscale interaction effects, with random intercepts for participants and stimuli. Results revealed distinct patterns across dimensions. Propositional scores strongly predicted Information ratings, showing a significantly tighter association than with Fluency or Language. Regarding fluency, the patterns were more selective: among automatic indicators, speech rate and mean length of run exhibited significantly stronger associations with the Fluency subscale than with Information. Conversely, pause-based measures tended to relate more strongly to Information (and occasionally Language) than to Fluency. Manual disfluency counts showed less consistent differentiation. These findings suggest that the relationships between objective performance features and rated dimensions are not uniformly distributed, underscoring the need to empirically validate how quality constructs are operationalized in interpreting assessment frameworks.

Bio

Chen Sijia is Associate Professor of Interpreting and Translation Studies at Guangdong University of Foreign Studies and Honorary Associate Professor of Linguistics at Macquarie University. Her primary research interests include cognitive approaches to translation and interpreting, translation and interpreting technology, and audiovisual translation. She investigates the cognitive processes underlying translation and interpreting using pen recording, eye tracking, think-aloud protocols, and psychometric methods.

Thematic presentation 2

Dependency distance, task complexity, and interpreting performance: A quantitative linguistic approach to interpreting assessment

Jiang Xinlei
Xi'an Jiaotong University

Abstract

This talk presents a systematic overview of my ongoing research program on interpreting assessment from a quantitative linguistic perspective. Drawing on dependency grammar as the primary theoretical and analytical framework, this program seeks to unravel the intricate relationships among multi-dimensional task complexity indices, interpreting performance measures, and holistic quality scores. The first line of inquiry examines the effect of source-text complexity, operationalized through dependency-based and traditional linguistic indices, on learners' interpreting performance. In particular, dependency distance is introduced as a core syntactic indicator to capture the structural processing demands of interpreting task. Based on a syntactically annotated bidirectional corpus of learners' English-Chinese consecutive interpreting, the findings demonstrate how source-text complexity significantly shapes learners' production under different interpreting directions, revealing differential sensitivity to structural demands. The second line of research focuses on quantitatively profiling interpreting performance and modelling its relationship with expert judgments of quality. By examining performance measures across the dimensions of complexity, accuracy, and fluency, the study applies statistical and machine learning methods to identify effective predictors of quality scores, with a particular focus on the explanatory power of quantitative linguistic indices. The findings contribute to advancing interpreting quality assessment by integrating quantitative linguistic indices. Future research will extend this framework to LLM- or AI-powered interpreting and explore the applicability of quantitative linguistic approaches to the quality assessment of AI-mediated interpreting.

Bio

Jiang Xinlei is an associate professor in the School of Foreign Studies and a Postdoctoral Fellow in the Institute of Artificial Intelligence and Robotics at Xi'an Jiaotong University. She holds a PhD in interdisciplinary linguistics and computer science. Her research integrates cognitive interpreting studies, quantitative linguistics, and corpus-based translation studies. She has published several interdisciplinary, peer-reviewed SSCI-indexed articles in journals including *IRAL*, *JoSTrans*, *Journal of Quantitative Linguistics*, and *Lingua*. She currently leads a national social science project on quantitative approaches to interpreting process research.

Thematic presentation 3

Leveraging pretrained language models for the automated assessment of English-Chinese interpreting: Explorations, advances, and reflections

Han Chao
National University of Singapore

Abstract

Translation and interpreting (T&I) are important components of many large-scale, high-stakes certification examinations and language proficiency tests, particularly across Asia. The large volume of T&I performance data generated in these contexts presents considerable challenges for testing agencies. Recently, artificial intelligence (AI) has been increasingly explored as a means of enabling automated scoring to improve scalability and efficiency. This presentation begins by outlining key findings from a recent systematic review of AI-based approaches to evaluating human-generated T&I, highlighting major trends and persistent limitations in assessment design and validation practices. Building on this background, I will then present a program of empirical research conducted with my colleagues on the use of pretrained language models (PLMs) to assess English-Chinese interpreting, based on the *Interpreting Quality Evaluation Corpus* (IQEC). In particular, I will focus on two approaches to automated interpreting assessment: prompting large language models (LLMs) and finetuning PLMs. Drawing on our critical reflections of the research findings, I will propose a tentative framework for the design and analysis of LLM-based assessment. I will conclude by calling for the development of large annotated benchmark datasets, the adoption of multi-pronged validation frameworks, and stronger commitments to interpretability and explainability.

Bio

Han Chao is Associate Professor (Dean's Chair) in the Faculty of Arts and Social Sciences at the National University of Singapore. He focuses on research and practice in the testing and assessment of translation and interpreting. His work has been published in *Interpreting*, *Target*, *Language Testing*, *Language Assessment Quarterly*, and *Computer Assisted Language Learning*. He is currently Associate Editor of *Language Testing in Asia* (Springer Nature) and serves on the editorial/advisory boards of *Interpreting* (John Benjamins) and *Language Testing* (Sage).

Individual presentation 1

Rethinking rating quality in interpreting assessment: The role of rater experience and reference texts

Chen Shirong
University of International Business and Economics

Abstract

This study explores how rater experience and reference text use shape rating quality in interpreting assessment. Sixty-two experienced and 62 novice raters evaluated 24 English-to-Chinese consecutive interpretations using either a source text or an exemplar. For fidelity, experienced raters and the use of the source text were associated with higher reliability and greater severity, whereas patterns for fluency and expression were more variable. In contrast, neither rater experience nor reference text had a significant effect on rating accuracy. These findings highlight the nuanced roles of rater expertise and reference use in interpreting assessment.

Bio

Chen Shirong received her PhD from Xiamen University. She is currently a Lecturer at the School of International Studies, University of International Business and Economics, China. Her research focuses on quality assessment in translation and interpreting (T&I) and cognitive approaches to T&I. Her doctoral work integrated eye-tracking, pen-recording, and score analysis to investigate interpreting assessment processes and outcomes. She has published in journals including *Asia-Pacific Education Researcher*, *Target*, *PLOS ONE*, and the *Interpreter and Translator Trainer*.

Individual presentation 2

Measuring rater cognition in interpreting assessment: Development and validation of the Raters' Scoring Process Scale

Jiang Mengting
Xiamen University

Abstract

In rater-mediated performance assessment, the process through which raters acquire, retrieve, and synthesize information before assigning scores is central to establishing scoring validity. While various models of scoring process have been proposed for monolingual writing and speaking assessment, research concerning interlingual interpreting assessment remains scant. Furthermore, much of the extant literature relies on qualitative methods, such as think-aloud protocols, which may limit the reliability, comparability, and generalizability of findings. Such methodological constraints underscore the need to quantify the scoring process with a psychometrically validated measurement tool. Against this background, this study aims to develop and validate the Raters' Scoring Process Scale (RSPS) in the assessment of bidirectional English-Chinese consecutive interpreting. Adopting an exploratory sequential mixed-methods approach, the research was conducted in four interrelated phases: 1) item generation based on stimulated recalls (n = 40) and focus groups (n = 16); 2) content validation via cognitive interviews (n = 5) and expert reviews (n = 5); 3) construct validation through exploratory factor analysis (n = 300) and confirmatory factor analysis (n = 300); and 4) an examination of criterion-related validity by analyzing the relationship between the scoring process and scoring outcomes. The findings may contribute theoretical insights into the scoring process of interlingual language assessment and offer a validated instrument to standardize future empirical investigations.

Bio

Jiang Mengting is a doctoral student majoring in Interpreting Studies at the College of Foreign Languages and Cultures, Xiamen University, China. She is co-supervised by Professor Han Chao from the National University of Singapore and Professor Chen Jing from Xiamen University. Her research focuses on interpreting assessment, scale development, and meta-research.

Individual presentation 3

Modeling fluency as process: A temporal-sequential analysis of interlingual interpreting performance

Jiang Zhaokun
National University of Singapore

Abstract

Fluency assessment in interlingual interpreting has conventionally relied on aggregated indicators such as mean speech rate and overall pause frequency. While useful for proficiency judgments, these static measures conceal important temporal characteristics of cognitive processing. For example, frequent short hesitations associated with lexical retrieval may yield the same total pause duration as rare but extended silences reflecting higher-level planning, even though they point to distinct cognitive bottlenecks and instructional needs. To capture these differences, this study adopts a process-oriented perspective in which student interpreting output is modeled as a continuous stream of time-aligned speech and disfluency events. Using latent state modeling on annotated audio data, this study aims to identify recurrent temporal-sequential patterns and estimate the duration of different behavioral states, thereby inferring underlying phases of cognitive activity during interpreting. These dynamic profiles are then used to characterize individual production styles and to examine whether process-sensitive features provide stronger predictions of interpreting quality than conventional aggregate measures. By modeling fluency as a temporally unfolding process, this study advances cognitively informed interpreting assessment and supports the development of diagnostic feedback that can guide targeted pedagogical intervention based on learners' specific processing constraints.

Bio

Jiang Zhaokun is a PhD student in the Department of Chinese Studies at the National University of Singapore. Her recent research focuses on cognitively informed approaches to interpreting assessment and the enhancement of assessment through human-AI teaming. Her work has appeared in peer-reviewed journals including *International Journal of Applied Linguistics*, *Research Methods in Applied Linguistics*, *Translation Studies*, and *Digital Scholarship in the Humanities*.

Individual presentation 4

Exploring the predictive validity of a Rasch-calibrated integrative test for interpreting aptitude: Rediscovering the SynCloze

Liu Yubo
Beijing Foreign Studies University

Abstract

This study investigates the psychometric properties and predictive validity of the SynCloze as an integrative aptitude test for conference interpreting. The SynCloze requires test takers to orally generate contextually appropriate and synonymic sentence completions under time pressure, tapping context-driven comprehension, anticipation, and associational fluency. In the first phase of this study, a cohort of 58 postgraduate interpreting trainees in China completed an English and a Chinese SynCloze at program entry. Responses were evaluated using an enhanced, fine-grained scoring framework and subsequently calibrated with Rasch models to monitor unidimensionality, item fit and scale functioning. After multiple rounds of iterative Rasch calibration and resulting item removal and scale restructuring, both SynClozes demonstrated satisfactory psychometric properties, including acceptable item fit, reliable separation of candidates, and stable scale functioning. In the second phase, the trainees participated in Chinese-English bidirectional consecutive interpreting and simultaneous interpreting tests after one year of instruction. Interpreting performances were assessed by four interpreting trainers and practitioners using an analytic, multi-criteria rating scale. Rater training and trial rating were conducted before operational rating. Rater scores were analyzed using the Many Facet Rasch Model to monitor rater self-consistency and scale functioning, which entailed rater calibration and scale collapsing. Results demonstrated acceptable rater fit and stable rating scale functioning. In the third phase, scores of the SynClozes and interpreting tests were subjected to correlation, regression, and mediation analyses. Results show that the English SynCloze performance is significantly associated with, and predictive of, interpreting performance across directions and modes, with stronger relationships observed for simultaneous interpreting. Moreover, the predictive mechanism between the first response for each item of the English SynCloze and final interpreting performance is fully mediated by synonymic responses, demonstrating the value of expressional and associational fluency as strong interpreting aptitude predictors.

Bio

Liu Yubo is an associate professor at the Graduate School of Translation and Interpretation, Beijing Foreign Studies University. His research focuses on interpreting aptitude testing and interpreting quality assessment. He leads a National Social Science Fund project on AI-driven interpreting aptitude testing and assessment. His publications have appeared in peer-reviewed journals such as *Interpreting*, *Interpreter and Translator Trainer*, and *Across Languages and Cultures*.

Individual presentation 5

Measuring washback on learning in interpreting: Development and validation of the CATTI Self-directed Preparation Strategy Scale

Kang Yuhan
Xiamen University

Abstract

Research on washback in test preparation has primarily focused on second language (L2) tests, with limited attention given to the distinct context of translation and interpreting (T&I) tests. Although the China Accreditation Test for Translators and Interpreters (CATTI) is a high-stakes examination influencing T&I training and selection, empirically validated instruments to measure its impact on learning remain scarce. To address this methodological gap, this study conceptualizes interpreting test preparation strategy and develops the Self-Directed Test Preparation (SDTP) Strategy Scale specifically for the T&I context in China. Adopting an exploratory sequential mixed-methods design, the study first identified constructs and developed an initial item pool through qualitative analysis of focus group interviews, learner journals, and online Q&A forum data. Subsequently, a survey was administered to a large sample of CATTI candidates (N=2,262). Psychometric properties were established using a split-sample approach. Exploratory Factor Analysis (EFA) (n_1=554) identified a latent structure, which was confirmed by Confirmatory Factor Analysis (CFA) (n_2=555). Rigorous tests for composite reliability (CR), average variance extracted (AVE), and discriminant validity (HTMT) were conducted. The final SDTP strategy scale includes 25 items representing five dimensions: Goal Setting and Planning, Post-practice Self-evaluation, Test-focused Resource Building, Affect and Motivation Regulation, and Social Support and Collaboration. This study provides a psychometrically sound tool to quantify learner-level preparation practices in interpreting and supports future research on the consequences of high-stakes T&I tests by enabling systematic investigations of candidate behaviors and related learning processes.

Bio

Kang Yuhan is currently a doctoral candidate at the College of Foreign Language and Culture, Xiamen University. Her research centers on interpreting testing and assessment, with a particular focus on the washback effect.

Individual presentation 6

An interdisciplinary approach to notetaking research: What can consecutive interpretation and higher education researchers learn from each other?

Rebecca Yeager
University of Illinois Urbana-Champaign

Abstract

This talk opens the door to more productive interaction between researchers who study notetaking for consecutive interpretation and higher education contexts. The talk begins with an outline of the similarities and differences between notetaking for consecutive interpretation and academic purposes, and continues by defining areas of overlap between the disciplines in terms of shared research questions and methodologies. This potential is illustrated by an innovative methodology to examine notetaking for second language listening assessment, with potential for application in studies of notetaking for consecutive interpretation. A concluding Q and A session invites more dialogue and collaboration across disciplines.

Bio

Rebecca Yeager has an MA in TESOL and Applied Linguistics from Indiana University and taught ESL for thirteen years at the University of Iowa, where she also assumed responsibility for English placement testing. She is now a PhD student at the University of Illinois Urbana-Champaign, where her research focuses on the intersection between listening, writing, and integrated assessment. She has published in *Language Teaching Research*, the *International Journal of Listening*, the *Journal of English for Academic Purposes*, and *Language Testing*.

Individual presentation 7

Beyond one-turn prompting: Leveraging self-reflection and multi-agent interaction in assessing students' consecutive interpreting

Lu Xiaolei
Xiamen University

Abstract

Previous research revealed that interim scoring is a common strategy employed by human raters. Under such strategy, human raters assign an initial score, deliberate over it, and revise their judgments when needed. However, current LLM-based interpreting assessment has largely relied on one-turn prompting, in which the LLM produces a single-pass score that cannot be further reviewed or revised. To examine whether process-oriented scoring can improve LLM-based assessment, the present study moves beyond one-turn prompting and operationalizes LLM-based assessment as a multi turn and iterative process that consists of initial scoring, deliberation, and revision. Study 1 examines whether self-reflection improves scoring performance by reviewing score justification, rubric compliance, and possible rater effects. Study 2 investigates whether multi-agent debate enhances assessment performance through collaborative scoring. A follow-up bias analysis evaluates the robustness of multi-agent debate by probing position bias, verbosity bias, and self-enhancement bias. Overall, this study explores how moving from one-turn scoring to a process-oriented workflow can support a more reliable, interpretable, and fairer LLM-based assessment of consecutive interpreting.

Bio

Lu Xiaolei is an associate professor at the College of Foreign Languages and Cultures of Xiamen University, China. Her research interests include translation technology and automated translation and interpreting assessment. Her articles have appeared in peer-reviewed journals such as *Computer Assisted Language Learning*, *Language Testing*, *Interpreting*, *Target*, the *Interpreter and Translator Trainer*, and *Natural Language Engineering*. She is the co-author of *Applied Corpus Processing with Python*, a volume dedicated to corpus data processing and analysis.

Individual presentation 8

Machine vs. human: An exploratory study of Chinese-Portuguese classical poetry translation quality assessment

Jiang Lili & Han Lili
Macao Polytechnic University

Abstract

AI-driven machine translation (MT) has expanded possibilities for low-resource language pairs; yet poetry translation remains difficult to assess with practicable and reusable translation quality assessment (TQA) designs, particularly for Chinese-Portuguese. This descriptive-exploratory study benchmarks large language model (LLM) translations against published human translations and examines cross-method evidence for classical poetry TQA. A bidirectional corpus of 20 canonical texts (ten Tang Dynasty Chinese poems and ten Renaissance Portuguese sonnets) was translated by three LLM-based systems (ChatGPT, DeepSeek, and Qwen/Tongyi Qianwen) and compared with published human translations as baselines. Five bilingual expert raters conducted blind evaluations using Comparative Judgement (CJ; full ranking and pairwise comparisons) and holistic scoring, complemented by a post-hoc questionnaire. Our findings reveal distinct directional effects and system-specific strengths. In Chinese-Portuguese direction, human translations show an overall advantage over MT outputs. Conversely, in Portuguese-Chinese, the highest-performing LLM achieves a comparatively higher share of top ranks on some items than the human baseline. Moreover, cross-method analyses examine the extent to which CJ and holistic scoring provide convergent and complementary evidence for poetry TQA, and how pairwise-based and ranking-based CJ outcomes relate. Raters' feedback further clarifies perceived affordances and preferences across methods.

Bio

Jiang Lili is an experienced conference and diplomatic interpreter-translator for multilateral international organizations since 2018 in the languages of Chinese, Portuguese, and English. She provided hundreds of simultaneous and consecutive interpretation services in areas of multilateral political, economic, and commercial cooperation and negotiations. She is a fourth-year PhD student at Macao Polytechnic University. Her research interests are interpreting studies, interpreting testing & assessment, and meta-research in translation and interpreting.

Han Lili is Associate Professor of the Faculty of Applied Sciences of Macao Polytechnic University, Macau. She has lectured and conducted research in interpreting studies, acting as trainer for the Conference Interpreting (Chinese-Portuguese-English) course in partnership with the DG (SCIC) of the European Commission. Her research interests include interpreting studies, intercultural studies, language and translation policy studies, interpreting testing & assessment, and computer-aided interpreting.

Individual presentation 9

Investigating student interpreters' feedback-seeking processes from GenAI: An exploratory study using behavioral sequence analysis

Chen Qionglu
Xiamen University

Abstract

This study employed cluster analysis and behavioral sequence analysis to investigate how language interpreting learners seek feedback from GenAI after completing language interpreting tasks. Data were collected from language interpreting recordings, GenAI logs and participants' think-aloud protocols. Thirty three participants were categorised into three clusters based on their frequencies of feedback seeking behaviors, and the clusters were labelled as analysis-oriented seekers, assessment-oriented seekers, and understanding-oriented seekers. According to the results of behavioral sequence analysis, analyzing, assessing, understanding, and remembering are the most frequent feedback seeking behaviors of learners. Although analysis-oriented seekers and assessment-oriented seekers demonstrated higher language interpreting performance than understanding-oriented seekers, understanding-oriented seekers exhibited significantly greater improvement in language interpreting quality compared to the other two groups after seeking GenAI feedback. Based on the results, we presented a feedback seeking model for describing the process of diverse feedback seekers in GenAI environments. These findings underscore the importance of developing GenAI feedback literacy in language interpreting training, as it enables learners to more effectively leverage GenAI for self-regulated learning.

Bio

Chen Qionglu is a Ph.D. student in the Department of English Language and Literature at Xiamen University, China. Her research interests are in the GenAI-supported language interpreting assessment and feedback. She has published in SSCI-indexed journals such as *Language Testing* and *Innovations in Education and Teaching International*.

Individual presentation 10

Diagnosing language-pair bias in quality assessment of speech translation using a multi-agent AI framework

Wang Xiaoman & Claudia Angelelli
Heriot-Watt University

Abstract

LLM agents are increasingly used to assess human and machine translation and interpreting quality. It remains unclear whether they treat different language pairs with equal strictness, and whether multi-agent deliberation reduces or amplifies any bias they carry. If agents systematically assign lower scores to one language pair than another, the resulting rankings across systems or languages become unfair. We present MATIA (Multi-Agent Translation & Interpreting Assessment), a framework with four LLM agents, each covering one of Angelelli's quality criteria: comprehension, grammar, style, and situational appropriateness. The agents work through a three-phase pipeline: 1) independent per-criterion scoring, 2) two rounds of structured deliberation among agents, and 3) leader synthesis into a single 0-100 quality score. Applied to a stratified sample of segments covering the full score range, drawn from the IWSLT 2026 Metrics shared-task speech-translation dev set in which each segment pairs an English source audio clip (with transcription) to German and Chinese target translations with human quality scores (0-100), the pipeline issues 2,520 agent calls across EN→DE and EN→ZH. We compare agent scores against human gold ratings using three diagnostics: the Language Pair Sensitivity Index (LPSI), Assessment Convergence Coefficient (ACC), and Dimensional Bias Index (DBI). The results show that MATIA agents systematically underrate speech translations relative to human judges in both directions, with mean agent scores of 33.7 (DE) and 36.7 (ZH) versus human means of 64.1 and 61.0, yielding LPSIs of −30.3 and −24.4. The LPSI gap between language pairs (−5.98; t = −1.61, p = 0.11; Mann−Whitney, p = 0.11) is not statistically significant, indicating that agents are uniformly harsh across both language pairs rather than biased against one. However, deliberation amplifies rather than corrects bias, with ACC values of −9.08 (DE) and −9.29 (ZH) showing that Phase-3 consensus drifts further from human ratings than Phase-1 independent scores. Criterion-level DBI reveals that comprehension dimension is the most divergent dimension between pairs (DBI gap = 15.7), with German comprehension scored substantially more leniently than Chinese, while grammar and style behave more symmetrically. Multi-agent LLM evaluators can inherit and amplify individual-agent severity through deliberation, and cross-lingual bias concentrates in specific quality dimensions. Agent-based QE systems therefore need per-criterion calibration and deliberation-aware aggregation before they can reliably complement human assessment.

NUS National University of Singapore | Faculty of Arts & Social Sciences

Bio

Wang Xiaoman is an assistant professor working at Heriot-Watt University, UK. Her research lies at the intersection of Interpreting Studies and Natural Language Processing, with a particular focus on the automatic assessment of interpreting quality. She also works more broadly in the areas of language and technology, and digital humanities. Her works have been published in refereed SSCI/A&HCI/CSSCI journals such as *Perspectives*, *Across Languages and Cultures*, *LANS-TTS*, the *Interpreter and Translator Trainer*, and *Chinese Translators Journal* (中国翻译).

Prof. Claudia V. Angelelli is Chair in Multilingualism and Communication at Heriot-Watt University, Emeritus Professor of Spanish Linguistics at San Diego State University, and Visiting Professor at Beijing Foreign Studies University. Holding a Ph.D. from Stanford, her work spans Sociolinguistics, Applied Linguistics, and Translation and Interpreting Studies. She has led research projects across Europe, the Americas, and Australia, served as World Leader for *ISO 13611 on Community Interpreting*, and is Past President of ATISA. She is sole author of three monographs and has published widely in leading international journals.

NUS National University of Singapore
Faculty of Arts & Social Sciences