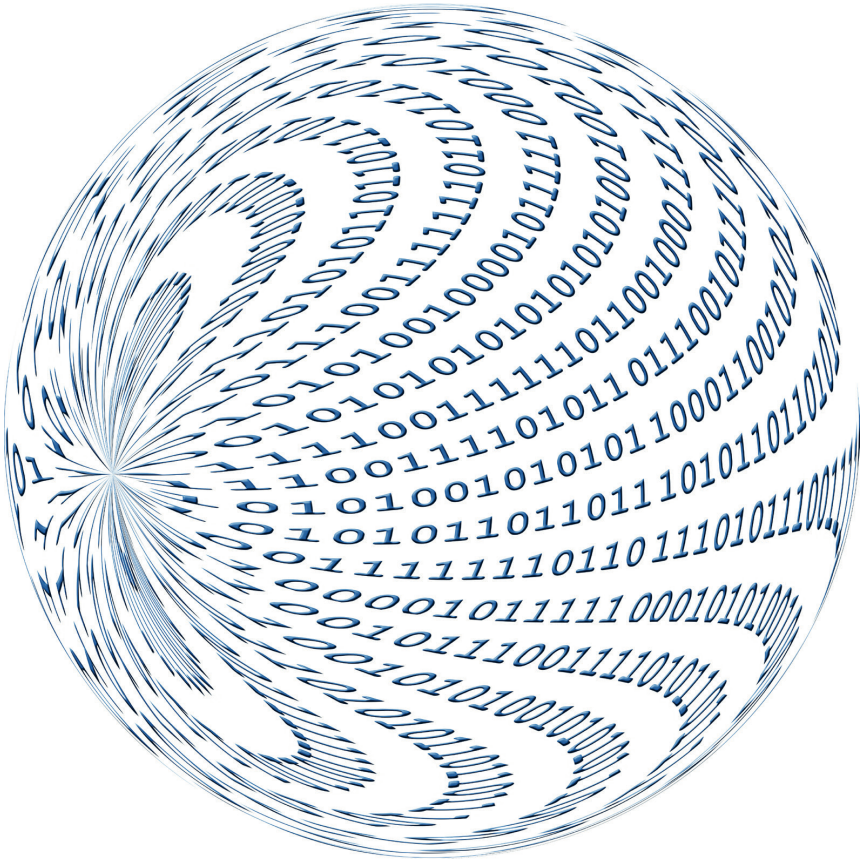
Rethinking English Language Teaching in an Algorithmic World

Critical Approaches to Human-Centered and Digitally-Enhanced Education

Edited by
Antonella Luporini and Antonella Giacosa



Bologna
University Press

Rethinking English Language Teaching in an Algorithmic World

Critical Approaches to Human-Centered and Digitally-Enhanced Education

Edited by
Antonella Luporini and Antonella Giacosa

Bologna
University Press

Il volume è stato pubblicato con il contributo dell'Alma Mater Studiorum - Università di Bologna.

Fondazione Bologna University Press
Via Saragozza 10,
40123 Bologna
tel. (+39) 051 232 882
www.buponline.com
e-mail: info@buponline.com

Quest'opera è pubblicata sotto licenza Creative Commons CC BY-4.0

ISBN on line 979-12-5477-965-1
DOI 10.30682/9791254779651

Questo volume è stato realizzato a partire da un impaginato camera-ready fornito dalle curatrici

In copertina: geralt/pixabay.com

Prima edizione: ottobre 2026

TABLE OF CONTENTS

Preface 5
Wissia Fiorucci

Questioning and rethinking technology-enhanced English Language Teaching in a fast-evolving educational landscape 9
Antonella Luporini, Antonella Giacosa

Lost in translation, found in post-editing: Machine translation and post-editing strategies in an English-Italian MA translation course 17
Antonella Luporini

A corpus-based approach to teaching English collocations in the AI era: A study of Persian EFL learners 37
Farideh Amirfarhangi

Automated assessment of L2 English skills with Talketet–A tool based on LLMs and speech technologies 63
Claudia Roberta Combei, Francesco Zappulla

Beyond templates: The DiFree repository and Moodle course for digital freelancing in ESP pedagogy 83
Anna Romagnuolo

Rebutting GenAI: Engaging with *IOW Dictionary* and podcast in classroom practices to develop critical language awareness and critical digital literacy 103
Paola Giorgis, Andrea C. Valente

Negotiating identity and agency in AI-mediated pedagogy: A critical mixed-methods analysis of EFL students' perspectives across educational levels 131
Antonella Giacosa

Contributors 153

AUTOMATED ASSESSMENT OF L2 ENGLISH SKILLS WITH TALKETET– A TOOL BASED ON LLMS AND SPEECH TECHNOLOGIES

Claudia Roberta Combei, University of Rome "Tor Vergata"
Francesco Zappulla, Etet S.r.l. Società Benefit

1. Introduction

Technology-enhanced language learning (TELL) has transformed the interaction between teachers and learners, as well as the ways in which learners behave within the language-learning process (Walker & White, 2013). These transformations regard lesson design, classroom management, and the assessment of acquired second language (L2) skills (Stockwell, 2018). The use of new technologies has brought about changes not only in methodology but also in terminology, giving rise to concepts such as intelligent computer-assisted language learning (ICALL, see Heift, 2021). The growing interest in ICALL is demonstrated by the creation of a Special Interest Group within the Computer Assisted Language Instruction Consortium, aimed at fostering the development of tools based on Artificial Intelligence (AI) and natural language processing (NLP) technologies, including Large Language Models (LLMs) and automatic speech recognition (ASR).

Even if recent studies in the field of ICALL have shown that AI can be successfully applied both in the generation of test items (Donati et al., 2024) and in the automated assessment of language skills (Yavuz et al., 2025), including oral skills (Nebhi & Szaszák, 2023), language assessment should not be viewed as a mere technical exercise in measurement. This is important given that language testing functions as a socially situated institutional practice through which individuals are channeled into educational pathways (e.g., admission to university degree programs), professional opportunities (e.g., employment, promotion, or access to internationally oriented roles), and, in some contexts, legal categories and rights (e.g., applications for visa, resident permits, citizenship, or asylum), all

of which often depend on demonstrated L2 proficiency. From an applied linguistics perspective, this foregrounds the consequential nature of assessment decisions and the need to evaluate not only reliability and efficiency, but also construct validity, fairness, and equity. Such considerations are especially relevant nowadays, in light of the growing demand for language certification and the increasing digitization and automation of testing infrastructures (some being AI-mediated).

Over recent decades, language testing/assessment (LT/LA) has developed along two closely intertwined trajectories. The first is theoretical and advances in the field of Second Language Acquisition (SLA) where models of communicative competence have motivated a gradual move away from exclusively structuralist and narrowly psychometric approaches (Lado, 1961) towards assessment designs that better reflect language use as social action in context (Hymes, 1972). This shift has supported greater attention to authenticity, interactional demands, and the integration of skills (Vignoli, 2026). The second trajectory is technological and depends on the fact that digital delivery, multimedia affordances, and networked platforms have transformed how language performance can be elicited, recorded, analyzed, and evaluated, giving rise to a broad research and practice ecosystem around computer-assisted language testing/assessment (CALT/CALA, see Suvorov & Hegelheimer, 2014; Winke & Isbell, 2017)[1]. More recently, LLMs and speech technologies have introduced an additional significant change, enabling systems that do not merely deliver tests but can participate in item generation, adaptivity, scoring, and feedback at multiple stages of the assessment lifecycle (Arronte Álvarez & Xie Fincham, 2025; Pack et al., 2024).

Following the recent developments in NLP and AI, our contribution presents *Talketet*, a web-app for the automated assessment of L2 English. The tool deploys ASR and LLMs to evaluate the learners' performance across four core skills for L2 testing: listening, reading, writing, and speaking. Talketet also integrates real-time feedback algorithms and a speech component capable of measuring various aspects of oral production. In the remainder of this chapter, we first outline the scope of our research on the development of an automated tool for L2 assessment (Section 2). We then describe Talketet (Section 3), present the results of a pilot study involving 43 participants whose L2 English skills were assessed using the tool, discuss the implications of the findings, and reflect on the current

[1] In the literature, the terms computer-assisted language testing (CALT) and computer-assisted language assessment (CALA) are widely used. Although some authors draw distinctions between testing–associated with standardized measurement and summative decisions (Suvorov & Hegelheimer, 2014) and assessment–associated with broader interpretive judgment, educational purposes, and classroom-based practices (Winke & Isbell, 2017), the boundaries are not consistently applied across research traditions or operational contexts. In this chapter, we use these terms to refer to computer-supported procedures for eliciting, interpreting, and scoring L2 skills.

limitations of the tool (Section 4). We also present some concluding remarks and outline possible directions for future development (Section 5).

2. The scope of AI-mediated L2 assessment

Traditionally, large-scale language testing relied heavily on paper-and-pencil instruments and had rather controlled, objective formats. Early mechanized scoring technologies often used multiple-choice tasks, operationally attractive at scale, despite critiques regarding their limited ability to represent complex communicative competence (Fulcher, 2000). The introduction of computer delivery, several decades ago, involved mostly computer-administered conventional tests, namely digital replicas of paper exams that preserved item types and test structure, while enabling faster scoring and data handling (Lynch, 2022). Although this paradigm is effective for certain standardized uses, it sits uneasily with assessment designs that aim to capture more dynamic, integrated, and contextually embedded language ability (Vignoli, 2026).

During the 1990s, the capacity to embed audio, images, and video expanded the range of task types available to test developers and supported more communicatively oriented assessment designs (Chapelle & Douglas, 2006). From the SLA standpoint, these developments matter insofar as they can increase both situational authenticity (alignment between test tasks and real-world communicative settings) and interactive authenticity (the extent to which tasks engage the language knowledge, strategies, and pragmatic resources that candidates would mobilize outside the testing context).

In the early 2000s, web-based delivery further extended reach and scalability through browser-mediated access, enabling remote participation and centralized, real-time data collection. Later on, the integration of AI has expanded the role of computation in LA from supporting delivery and scoring to intervening across multiple stages of the assessment cycle. The tool we present in this chapter exploits the advances in NLP, in particular ASR and large-scale neural architectures such as transformers. These techniques enable new forms of automated scoring and feedback, but they also introduce new risks related to robustness, bias, explainability, and the defensibility of score-based decisions.

AI is used in CALA across at least three interrelated domains: (i) test design and item development, (ii) test delivery and adaptivity, and (iii) automated scoring of productive performance. In the first domain, the central issue extends beyond the feasibility of generating test items to the more consequential question of validity: whether AI-produced items appropriately operationalize the intended construct, function as expected for the target population, and remain defensible

given the stakes attached to score use (Shin et al., 2025). In the second domain, AI-enabled delivery systems can move beyond conventional Computer Adaptive Testing (CAT) by supporting fine-grained, real-time monitoring of candidate performance (e.g., systematic error patterns and response behaviors). Such monitoring can inform dynamic adjustments to task selection and sequencing, potentially mitigating frustration when tasks are overly difficult and reducing disengagement when tasks are insufficiently challenging (Lavanya et al., 2024). In the third domain, Automated Essay Scoring (AES) and Automated Speaking Assessment (ASA) systems typically rely on deep learning architectures that learn scoring-relevant representations from data and, in some implementations, incorporate rubric-related information via prompting to produce both scores and accompanying textual rationales. ASA systems usually integrate ASR with acoustic-prosodic analysis and pragmatics evaluation, capturing dimensions of spoken performance that are not recoverable from transcripts alone. However, prompt-engineered[2] AES tools derive much of their effectiveness from the quality and design of the prompts that guide them, including how clearly scoring criteria and expectations are articulated. Their performance is, therefore, sensitive to prompt formulation, and less intrinsically stable than systems trained directly on annotated data, such as the fine-tuned[3] AES tools. Moreover, because these tools rely on a general-purpose underlying LLM, their behavior is inherently dependent on ongoing model updates (e.g., changes to the base model can alter scoring patterns and rationales, even when the prompt remains unchanged). As a result, reliability in such systems reflects not only prompt design, but also the evolving characteristics of the LLM that drives the assessment process. On the same note, the validity of ASA remains closely contingent on ASR robustness, which is frequently weaker for L2 speech than for native speech and can be affected by recording quality, environmental noise, and algorithmic biases.

In fact, a recurrent topic in recent work in SLA is the need to retain meaningful human oversight in AI-mediated L2 assessment. Human-in-the-Loop (HITL) approaches integrate human judgment across data curation and annotation, rubric design, model training and calibration, and the validation of automated outputs, allowing to monitor for error patterns and bias and to support more defensible decision-making (Fajardo-Ramos et al., 2025). This

[2] Here, prompt-engineered tools refer to tools that depend on how instructions are formulated for the LLM, such as asking it to score a text written by a learner according to specified criteria (e.g., lexical range, coherence, task fulfillment, etc.).

[3] Fine-tuned tools refer to tools that were further trained on annotated learner data (e.g., essays and other textual typologies scored by humans). In practice, however, fine-tuning and prompting may also be combined.

aligns with the view that assessment validity is inseparable from the interpretive arguments and consequences surrounding score use (Vignoli et al., 2025).

Taken together, the road from paper-based testing to AI-enabled CALA suggests that technological progress does not automatically entail better assessment. Instead, advances in delivery, adaptivity, and automation must be evaluated against core SLA criteria (e.g., construct representation, authenticity, fairness, impact, etc.), while NLP contributes tools for modeling language and speech, but also methods for auditing robustness, bias, and explainability. We believe that positioning Talketet at the intersection of these fields provides a basis for understanding both the opportunities, the limits, and the risks of AI-based L2 assessment, and for motivating principled design choices.

3. The Talketet architecture[4]

This Section describes the 2024/2025 version of Talketet used in the case study presented in Section 4. Since the completion of this study, the tool has undergone further development, partly driven by the linguistic experiments discussed in Section 4 and by other studies (Feccia, 2026; Vignoli, 2026; Vignoli et al., 2025).

Talketet is a commercial multilingual web-app designed to provide an automated L2 assessment of written and oral language skills by means of both closed-ended and open-ended questions. The assessed skills include oral production and interaction, listening, reading comprehension, written production, and grammar. Talketet returns scores on a 0-100 scale, weighted according to question difficulty and item type; scores are also mapped onto the Common European Framework of Reference for Languages (CEFR)[5] levels, both globally and for each assessed skill.

3.1. Technical features and types of tasks

The platform used in this study integrates real-time feedback algorithms, LLMs, and speech technologies, drawing on recent work in AES (Yavuz et al., 2025) and ASA (Zechner & Evanini, 2019). From an architectural point of view, the 2024/2025 version of Talketet runs in a Linux (Debian) environment. The back end is implemented in Ruby on Rails 8, while the management interface

[4] The architecture of Talketet and its Italian version are described in detail in Vignoli et al. (2025).
[5] For details on the Common European Framework of Reference for Languages (CEFR), see the official website: https://europass.europa.eu/en/common-european-framework-reference-language-skills (accessed 10 May 2026).

(back-office) is developed in Vue.js. The end-user front-end module is also based on Vue.js. Authentication is performed via email and password, and the session is verified via JWT tokens. The ASR component is based on AzureAI that uses OpenAI's Whisper model. Voice inputs are captured in the browser in .ogg or .wav format (mono-channel) and are not normalized. Decoding parameters are not configurable. Open-ended questions are evaluated via OpenAI's GPT-4o, an API-accessible model, using customized prompts for each test and question type. Scoring is asynchronous, runs in the background, and is returned only upon completion of the test. All examination-related data (e.g., texts and audio files produced by users as responses) are stored in a PostgreSQL database hosted in the cloud on Azure infrastructure, in compliance with the GDPR. Disks are encrypted with platform-managed keys, and a periodic snapshot policy is applied to ensure data security and preservation.

The Talketet web-app allows the creation of different test types depending on the needs of individual examiners. The starting point for test construction is the definition of the objective of the assessment, which entails specific requirements in terms of content and administration time. Tests may be designed to assess all four core language skills (reading, listening, writing, and speaking) or only the specific skills of interest to the examiner. Test creation follows an incremental process: for each core skill, a questionnaire is created with questions selected from a database prepared by internal experts and chosen to cover all aspects relevant to the linguistic competence under investigation. The decision to rely on an expert-curated database reflects a HITL approach to the automated L2 assessment. So, rather than delegating item generation entirely to automated systems, the platform grounds test construction in materials controlled by professionals. The database may include items authored, adapted, reviewed, or validated by internal experts, namely trained linguists, with M.A. and Ph.D. qualifications in linguistics and, in some cases, specific expertise in SLA. The set of questionnaires constitutes the overall structure of the exam.

Regardless of the chosen question types, each questionnaire is constructed by inserting items with a specified balance of expected complexity. In particular, each question is assigned a difficulty level aligned with the CEFR. Each questionnaire includes 10% of questions at A1 and 10% at C2, while all other levels (A2, B1, B2, C1) are uniformly represented with 20% of questions each. It is possible to set both the number of questions and the test duration according to user needs or the specific requirements of instructional and assessment contexts. Each question is identified via a set of features describing its type and complexity. Items are therefore catalogued according to the following parameters: assessed skill via an identifying label (reading, listening, writing, speaking); test language (here, English); CEFR difficulty level (A1, A2, B1, B2,

C1, C2); epistemic object targeted (e.g., definite articles, passive voice, etc.). Items can also be grouped into two macro-categories: closed-ended questions and open-ended questions, distinguished by the type of production required from the user and by the scoring procedure. In both cases, participants have a limited amount of time, measured in seconds, to formulate and submit their response.

Closed-ended questions generally assume a single correct answer or a predefined set of possible answers. Their scoring is therefore based on a pre-prepared answer key rather than on LLM-based interpretation. For this reason, during item creation particular attention was paid to anticipating and including, where relevant, alternative valid responses, such as spelling variants and synonymous forms. These answer lists were reviewed by internal experts with the aim of making them as exhaustive as possible. However, as in any closed-ended test format, the system can only recognize as correct the answers that have been specified in advance. If a candidate provides a valid but unforeseen answer, it would not be automatically accepted as correct. This constraint is not specific to digital assessment but is inherent to closed-ended formats more generally, including paper-based tests, where scoring depends on the predefined marking scheme. The types of closed-ended questions are briefly described below:

- Multiple-choice questions, where a prompt is associated with several possible answers, only one of which is correct, while the others function as distractors.
- Gap-fill (completion) questions, where a sentence is presented with one or more blanks that the user must fill with one or more words.
- Dictation, in which the test-taker writes what they can understand from the audio associated with the item.
- Sentence reconstruction, in which the user hears an audio recording containing fragments of a sentence presented in a scrambled order and must reconstruct them correctly.

In building the tests, we tried to include as few multiple-choice questions as possible, in order to reduce the likelihood of guessing. Instead, greater emphasis was placed on gap-fill items, which provide more authentic linguistic input by requiring active production from the user.

Open-ended questions require an even more active and autonomous linguistic elaboration by the candidate and target both writing and speaking. These questions make it possible to assess complex and integrative language skills. This category includes:

- Short compositions, consisting in the production of a brief text on various topics and across different registers.
- Image descriptions, where an image is provided as input and the participant should produce a detailed description.
- Expression of one's point of view, by writing a text that makes the candidate's position on a topic explicit (e.g., the use of AI in their daily life).
- Summaries, where the participant should produce a summary of a text just read or an audio recording just heard.

For open-ended writing items, responses must contain a minimum of 50 words, whereas for speaking items, at least 20 seconds of real-time audio recording are required. The inclusion of both question types is motivated by the need to obtain a more complete and balanced assessment of language proficiency. As noted in the literature, effective language assessment should integrate tasks that measure both formal linguistic knowledge and the ability to deploy that knowledge effectively in real-world contexts (Bachman & Palmer, 1996). This approach is consistent with the CEFR conceptualization, which characterizes L2 competence as action-oriented, viewing learners as social agents. Through language, speakers must be able to interact effectively and communicate across diverse social, cultural, and professional contexts.

3.2. Scoring and mapping to CEFR

On the Talketet tests, each question is assigned a weight (1, 2, or 3 points) representing the score obtainable when answering correctly. Questions requiring greater re-elaboration and cognitive effort receive higher weights. Closed-ended items, such as multiple-choice or gap-fill questions, are weighted 1 point; dictations and sentence reconstructions are weighted 2 points; oral productions and written compositions are weighted 3 points. Closed-ended items are scored using a binary correct/incorrect evaluation. As detailed in Section 3.1, for written and spoken production, scoring is delegated to a prompt-engineered LLM capable of generating AES and ASA scores. The AES component evaluates grammatical accuracy, lexical choice, coherence and topic adherence, comprehension, textual cohesion, and the content of the user's response. The model's 0-100 score is distributed across five parameters as follows: structure and grammar = 20%, content and argumentation = 30%, vocabulary = 20%, comprehension and adherence to the topic = 20%, and pragmatics and cohesion = 10%.

The ASA model, after producing a transcription of the audio recorded in real time by the user, adopts an approach similar to Chau and Huensch's (2025)

operationalization of the Complexity-Accuracy-Fluency (CAF) paradigm[6]. In fact, it uses the Azure engine to analyze the speaker's oral production through four parameters: fluency, accuracy, completeness, and pronunciation. The values associated with these four parameters account for 5% of the final score for oral production questions. The remaining 95% derives from the parameters discussed above for AES, redistributed as follows: structure and grammar = 20%, content and argumentation = 25%, vocabulary = 20%, comprehension and adherence to the topic = 20%, and pragmatics and cohesion = 10%.

Once the user completes the test, the system calculates the score for each question as described above and reports the final score obtained in each section. The score is computed as the ratio between points obtained and total obtainable points for the relevant language skill–see the following equation:

$$= \frac{\quad}{\quad}$$

The final test score is computed as the mean of the scores obtained for each language skill, thereby equalizing their importance–see the following equation, where *n* is the number of assessed skills[7]:

$$= \frac{\sum = 1}{\quad} \%$$

To align the test output with that of certification bodies, the percentage score is mapped onto the CEFR bands. In addition to the percentage score, the user

[6] The CAF paradigm refers to the assessment of L2 performance in terms of complexity, accuracy, and fluency, three dimensions traditionally used to capture different aspects of learner competence and development. Its rationale is that L2 proficiency cannot be reduced to a single global score, because learners may, for instance, produce fluent speech that is not highly accurate, or accurate language that remains lexically or syntactically simple. In pronunciation-oriented assessment, however, CAF is often complemented by listener-based constructs such as fluency, intelligibility, comprehensibility, and accentedness, which focus more directly on how speech is perceived and understood. This distinction is important because communicative effectiveness may depend less on native-like accent than on whether speech is intelligible and comprehensible to listeners (Combei, 2023).

[7] This type of weighting is not intended as a fixed model. As mentioned earlier in the text, the platform allows scoring procedures to be adapted according to the aims of a specific test, examiner, or end users, meaning that in some cases, receptive and productive skills may be weighted differently. In the pilot study presented in Section 4, however, equal weighting was adopted because the main aim was to evaluate the functioning of the tool across skills rather than to privilege a specific construct of proficiency. At the same time, this approach offered a neutral baseline for comparing performance across the skills tested.

also receives a CEFR level (A1-C2) for each assessed competence and an overall CEFR level for the test as a whole (see Figure 1).

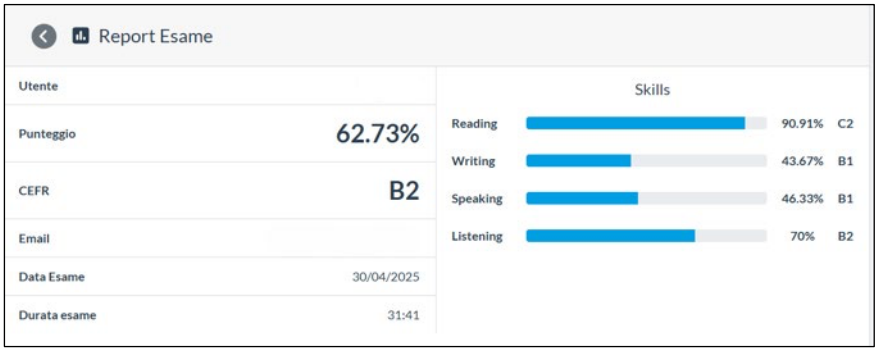


Figure 1. Talketet Back-end.

At the end of the test, the platform automatically generates a personalized descriptive feedback report which, through generative technologies, provides an overall analysis of the user's performance. The tool is supplied with the test results and with the responses to open-ended writing and speaking questions. These data are analyzed, and the generated feedback indicates both the strengths of the test-taker's linguistic competence and the areas in which greater difficulty was observed, also offering suggestions and recommendations for improving the assessed skills.

4. Case study

In order to build an initial validity argument for the 2024/2025 version of Talketet and to evaluate its reliability under ecologically plausible test administration conditions, we conducted a case study involving 43 Italian L1 participants with heterogeneous English proficiency profiles. Participants were recruited through a snowball procedure, beginning with contacts available to the team involved in the development and testing of the web-app, who were then asked to circulate the invitation further among potentially interested participants. Full demographic information is reported later in this section, after the description of the test administration procedure, so that the sample profile can be interpreted in relation to the assessment conditions under which the data were collected.

All participants completed the assessment individually via the Talketet web-app following standardized instructions designed to minimize administration-related variance. The Talketet test battery contained 60 items sampling the four

language skills: writing (14 items), speaking (10 items), reading (21 items), and listening (15 items). Item selection and assembly followed the blueprint described in Section 3, with explicit attention to (i) balancing item difficulty across the test and (ii) ensuring a representative distribution of task formats, including both closed- and open-ended questions (and their respective subcategories as defined in Section 3.1). The test was designed to provide adequate construct coverage while remaining feasible for web-based deployment. Accordingly, the targeted completion time was approximately 30-45 minutes.

Prior to beginning the test, participants received an information brief explaining the study aims, the assessment workflow, and how their data would be processed and used for research and tool development purposes. In line with established ethical standards in linguistics, NLP, and SLA research, participation proceeded only after participants had been informed of the procedure and data-handling practices. All data were anonymized prior to analysis to prevent the identification of individual participants. Responsibility for data collection and personal-data governance rested with the company operating Talketet, which acted as the Data Controller for Data Protection. Participation proceeded only after informed consent had been obtained.

To reduce technical confounds in the listening and speaking components, participants were required to complete a short training procedure before the test: the web-app guided them through a sound and microphone check and a recording trial to verify audio input functionality and to familiarize them with the recording interface.

In addition to performance data, we collected participant information to contextualize scores and support validation analyses. Socio-demographic variables included age, gender, educational level, and occupation. Linguistic background measures included age of first exposure to English, cumulative duration of English learning, primary learning contexts and methods, time spent in English-speaking countries, and self-reported proficiency mapped onto the CEFR levels. These variables served both descriptive purposes (sample characterization) and interpretive purposes (identifying potential sources of systematic variance in performance and user experience) for future research.

The participant sample consisted of 43 Italian-speaking respondents who completed a background questionnaire prior to the L2 English test. The participants were predominantly young adults, with 51.2% of the sample falling in the 25-34 age band and a further 32.6% in the 18-24 band. Smaller subsets were represented in older age groups (35-44: 4.7%; 55-64: 11.6%). In terms of self-identified gender, the sample included 24 women (55.8%), 18 men (41.9%), and one non-binary participant (2.3%). Educational level indicated a relatively highly educated sample. Participants reported having Master's degrees (37.2%)

and Bachelor's degrees (30.2%), and smaller proportions reported upper-secondary education (20.9%), a postgraduate specialization (7.0%), or a PhD (4.7%). Occupational status mirrored this profile: the sample was composed primarily of students (53.5%), followed by employed participants (37.2%), with a small number reporting unemployment (4.7%), entrepreneurship (2.3%), or PhD candidates (2.3%).

With respect to linguistic background, the large majority reported Italian as their only first language (95.3%), while two respondents (4.7%) reported being bilingual. Some participants reported studying, teaching, or doing research in the field of linguistics (25.6%). All participants (100%) reported speaking at least one additional language besides Italian, 62.8% indicated two or more languages, and 34.9% three or more. Across the languages listed, besides English, the most frequently reported were French, Spanish, and German. The reported age of onset for English study was generally early: the mean age of onset was 7.15 years (standard deviation = 2.27; median = 6). The duration of English study was comparatively long: participants reported an average of 13.76 years of study (standard deviation = 3.73; median = 14). In classroom experience, explicit pronunciation instruction was common (69.8%). Extended immersion experience was less prevalent: only 16.3% reported having lived in an English-speaking country for at least three months. Current English engagement was heterogeneous, with both high- and low-use profiles. Nearly half of the sample reported using English daily for work or study (46.5%), whereas the remainder reported occasional use (23.3%) or more limited engagement (at least weekly: 16.3%; at least monthly: 14.0%). Self-reported speaking frequency in oral interaction showed a similarly varied distribution, ranging from daily speaking (37.2%) to no speaking at all (14.0%), with intermediate categories including occasional (27.9%), at least weekly (16.3%), and at least monthly (4.7%). Participants also provided self-assessments of English proficiency in terms of the CEFR levels, both for overall proficiency and for oral production. In both cases, the distribution was weighted toward higher proficiency bands (B2-C2). For the self-reported overall CEFR levels, the sample is distributed as follows: A2 (7.0%), B1 (14.0%), B2 (18.6%), C1 (46.5%), C2 (14.0%). The self-reported CEFR oral skills are represented as follows: A2 (9.3%), B1 (11.6%), B2 (25.6%), C1 (39.5%), C2 (14.0%). Formal language certification was not used as a criterion in this pilot study, as the sample was expected to be non-homogeneous in this respect, with some participants likely holding certificates and others not. Self-assessed CEFR level was therefore chosen as a common proficiency indicator available for the whole sample. Its limitations as a subjective measure are acknowledged in the interpretation of the results.

The case study addressed two complementary aims: (1) to examine convergent evidence by comparing Talketet's automated proficiency estimates with participants' self-reported CEFR-aligned proficiency, and (2) to assess user experience by eliciting participant feedback on usability and perceived adequacy of the web-app (e.g., clarity of instructions, interface navigability, perceived task representativeness, and time burden).

As regards aim (1), we compared participants' self-reported CEFR levels with the CEFR classifications produced by the automated tool. The pattern that emerged was one of stronger convergence for overall proficiency than for oral proficiency, with the speaking-related output appearing systematically more conservative, in the sense that the tool tended to assign lower oral CEFR levels than those self-reported by participants. This may partly reflect the stricter constraints of automated speech processing, including the dependence of the ASR component on pronunciation models, as well as the fact that oral performance is affected by acoustic conditions.

For overall proficiency, exact agreement between self-report and test classification was observed in 53.5% of the cases, and, in general, they aligned within one CEFR band in 95.3% of the cases. In fact, only 2 participants (4.7%) showed a discrepancy of two bands or more, and just one case diverged by three bands. In other words, despite the difficulty of mapping a complex construct such as overall L2 proficiency onto a single categorical CEFR label, the overall estimates of Talketet were typically close to the participants' self-reported level (the mean absolute deviation is of 0.53 CEFR bands). This impression is also consistent with an ordinal agreement index; in fact, using quadratic-weighted Cohen's κ, agreement between overall test levels and overall self-reports was of 0.72, a value generally interpreted as substantial agreement for ordered categories (Cohen, 1960; Landis & Koch, 1977). At the same time, the direction of the residual mismatch indicates that overall test levels were higher than self-reports in 27.9% of the cases, lower in 18.6% of the cases, and identical in the remainder (53.5%).

The picture changes when turning to the oral dimension. Exact agreement between oral self-reported CEFR and the tool's oral test level occurred in 27.9% of the cases. Agreement within one CEFR band occurred in 53.6% of cases, while discrepancies of two bands or more were found in 18.5%. The mean absolute deviation was 1.21 CEFR bands. Moreover, the mismatch was strongly directional, in the sense that Talketet assigned a lower oral CEFR level than the one self-reported by the participant in 69.8% of the cases, and a higher level in only 2.3% of the cases. The value of the quadratic-weighted Cohen's κ indicating moderate ordinal agreement. This asymmetry is consistent with the documented limitations of self-assessment of oral skills (Ma & Winke, 2019). As

a matter of fact, even when CEFR descriptors are provided, respondents may calibrate their judgments to receptive abilities or general communicative confidence rather than to the full CEFR construct for spoken production (Heilenman, 1990; Jankowska & Zieliński, 2015). We also b participants may overestimate speaking competence, for example if their real-world speaking opportunities are limited, producing self-ratings that are aspirational. At the same time, the oral component of the tool may be stricter and more sensitive to dimensions that lay raters underweight (e.g., segmental and prosodic accuracy, discourse organization under time pressure, etc.), and/or it may be affected by construct-irrelevant variance typical of web-based speaking tasks (e.g., misalignment with the ASR pronunciation model, microphone quality, recording anxiety, etc.), which can affect performance and lead to lower predicted CEFR bands.

In general, the results presented above suggest that the overall proficiency output of Talketet shows good alignment with participants' own global CEFR positioning, whereas the oral module appears to require further calibration and triangulation, ideally through comparison with expert human ratings of the same recordings, refinement of cut scores, and longer or more varied speaking elicitation, before strong claims about oral-level classification accuracy can be made.

As for aim (2), based on the post-test questionnaire, the participants' feedback was broadly positive regarding usability and interface clarity, but critical with respect to timing and (to a lesser extent) technical issues. A large majority reported that the quiz instructions were clear and intuitive (85.7%) and that font size and readability were adequate (88.1%). Similarly, most participants judged the item wording to be clear overall (81.0%). Ratings of the platform's usability and graphics were high (mean = 4.0/5; standard deviation = 0.88; median = 4), and the perceived naturalness of the listening voices was also evaluated favorably (mean = 4.05/5; standard deviation = 0.96; median = 4). Many respondents appreciated Talketet's emphasis on spoken production (83.3% found the oral focus useful) and felt that it covered relevant aspects of oral communication (88.1%).

The most salient weakness concerned time pressure. All respondents said that the time available per item was inadequate, often describing the countdown as anxiety-inducing and as constraining the planning phase for both speaking and writing. This, in fact, may have forced respondents to focus on responding quickly, rather than on planning and formulating their answers. In addition, 42.9% reported technical difficulties, generally linked to initiating or authorizing microphone recording, sound and audio recording interruptions, or platform navigation issues (e.g., not managing to send an answer before timeout).

Perceived level fit was mixed (50.0% judged the test appropriate for their level), with some lower-proficiency participants saying that tasks were too demanding, and a minority of higher-level users noting that some prompts felt too short or insufficiently complex for advanced skills (C1-C2 levels). Despite these concerns, the number of questions was generally considered appropriate by 90.5% of the participants, and their overall rating of the test usefulness for English communicative ability was moderately positive (mean = 3.38/5; standard deviation = 1.13; median = 3.5), with 83.3% agreeing with the general usefulness of this type of test for L2 speakers.

The feedback above enabled the company to optimize the graphical user interface, the ASR model, and system response times. Consequently, the most recent version of the tool has been designed to mitigate, among other things, the issues identified by participants in this case study. The most significant modification concerns the assessment of oral skills. In the version described in this chapter, the system relied on a two-component architecture: an ASR engine and speaking-assessment module processed the oral input, while an LLM evaluated the resulting transcript and the other skills. In its new iteration, the web application is moving towards an architecture based on a single multimodal LLM capable of processing spoken input directly, without relying on an intermediate transcription stage. This may allow for a more integrated evaluation of oral performance and of the other language skills, although it does not remove the need for systematic validation, calibration, and human monitoring.

The case study also points to some critical dimensions of AI-mediated language assessment. Technical improvements alone cannot address issues such as construct under-representation, ASR-related bias, over-reliance on automated scores, accessibility barriers, or the anxiety and stress generated by tightly timed tasks. These considerations should consistently inform the development and use of such tools, especially in high-stakes contexts of language assessment, in which automated outputs may influence important decisions (e.g., admission to schools or university programs, as well as visa requirements or residence permits). For this reason, Talketet is best understood not as a replacement for human assessment expertise, but as an AI-based assessment environment that requires constant reviewing and HITL procedures for test design, item validation, score interpretation, and the monitoring of error patterns. This orientation is also consistent with the company's broader commitment to developing and deploying the system in line with the requirements of the European AI Act, especially with respect to transparency, risk management, and meaningful human oversight.

5. Concluding remarks

This chapter discussed both the promises and constraints of AI-mediated L2 assessment and introduced Talketet, a web-app designed to test L2 English listening, reading, writing, and speaking. The 2024/2025 version of the tool combines ASR and LLM-based scoring to handle closed- and open-ended responses and deliver real-time feedback. Scores are weighted by task type and difficulty, reported on a 0-100 scale, and mapped to CEFR levels, providing information on both global and skill-specific performance. In a case study with 43 Italian-speaking learners, automated outputs of Talketet were compared with CEFR self-assessments and complemented by feedback on usability data. The tool showed good convergence for overall proficiency (identical or within one CEFR band) but weaker agreement for oral proficiency, where automated estimates appeared to be more conservative, consistent with known limitations of self-reports and the greater sensitivity of speaking tasks to timing and technical issues. User feedback was positive, but it also highlighted actionable priorities, especially time allocation as well as sound and recording robustness, which informed subsequent interface and performance refinements.

Finally, this work highlights the importance of approaching AI-mediated language assessment critically rather than solutionistically. In this sense, the contribution of AI lies not in replacing human judgment, but in supporting L2 assessment workflows that should remain linguistically grounded, empirically validated, and accountable.

Author contributions

This chapter is the result of a collaborative effort between the two authors, who share responsibility for its content. For the purposes of the Italian academic system, we specify the contributions as follows: Sections 1, 2, and 5 were written by C.R. Combei; Sections 3 and 4 were written by C.R. Combei and F. Zappulla.

REFERENCES

Arronte Álvarez, A., & Xie Fincham, N. (2025). Automated L2 proficiency scoring: Weak supervision, Large Language Models, and statistical guarantees. In E. Kochmar, B. Alhafni, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds.), *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)* (pp. 384–397). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.bea-1.30

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford University Press.

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge University Press.

Chau, T., & Huensch, A. (2025). The relationships among L2 fluency, intelligibility, comprehensibility, and accentedness: A meta-analysis. *Studies in Second Language Acquisition,* 47(1), 282–307. https://doi.org/10.1017/S0272263125000014

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Combei, C. R. (2023). *Speaking Italian with a twist: A corpus study of perceived foreign accent*. Franco Angeli.

Donati, N., Periani, M., Di Natale, P., Savino, G., & Torroni, P. (2024). Generation and evaluation of English grammar multiple-choice cloze exercises. In F. Dell'Orletta, A. Lenci, S. Montemagni, & R. Sprugnoli (Eds.), *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)* (pp. 325–334). ACL Anthology & CEUR Workshop Proceedings.

Fajardo-Ramos, D. C., Chiappe, A., & Mella-Norambuena, J. (2025). Human-in-the-loop assessment with AI: Implications for teacher education in Ibero-American universities. *Frontiers in Education,* 10, 1710992. https://doi.org/10.3389/feduc.2025.1710992

Feccia, P. (2026). *Prompt-engineering e LLMs-as-Judges: ETET come caso di studio* [Unpublished master's thesis]. University of Pavia.

Fulcher, G. (2000). Computers in language testing. In P. Brett & G. Motteram (Eds.), *A special interest in computers* (pp. 93–107). IATEFL Publications.

Heift, T. (2021). Intelligent computer-assisted language learning. In H. Mohebbi & C. Coombe (Eds.), *Research questions in language education and applied linguistics* (pp. 114–126). Springer. https://doi.org/10.1007/978-3-030-79143-8_114

Heilenman, L. K. (1990). Self-assessment of second language ability: The role of response effects. *Language Testing,* 7(2), 174–201.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Penguin Books.

Jankowska, A., & Zielińska, U. (2015). Assessment instrument for developing the speaking skill at the advanced level. In M. Pawlak & E. Waniek-Klimczak (Eds.), *Issues in teaching, learning and testing speaking in a second language* (pp. 251–265). Springer. https://doi.org/10.1007/978-3-642-38339-7_16

Lado, R. (1961). *Language testing*. Longman Group Limited.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.

Lavanya, B., Naga Madhuri, J., Pagidipati, B., Mythili, M., Thiyagarajan, K., & Prema, S. (2024). Adaptive testing with machine learning: Customizing English proficiency assessments based on learner performance. In *2024 International Conference on Communication, Control, and Intelligent Systems (CCIS)* (pp. 1–5). IEEE. https://doi.org/10.1109/CCIS63231.2024.10931968

Lynch, S. (2022). Adapting paper-based tests for computer administration: Lessons learned from 30 years of mode effects studies in education. *Practical Assessment, Research & Evaluation,* 27(1), article 22. https://doi.org/10.7275/pare.1317

Ma, W., & Winke, P. (2019). Self-assessment: How reliable is it in assessing oral proficiency over time? *Foreign Language Annals,* 52, 66–86. https://doi.org/10.1111/flan.12379

Nebhi, K., & Szaszák, G. (2023). Automatic assessment of spoken English proficiency based on multimodal and multitask transformers. In R. Mitkov & G. Angelova (Eds.), *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing* (pp. 769–776). INCOMA Ltd. https://aclanthology.org/2023.ranlp-1.83/

Pack, A., Barrett, A., & Escalante, J. (2024). Large Language Models and automated essay scoring of English language learner writing: Insights into

validity and reliability. *Computers and Education: Artificial Intelligence,* 6, 100234. https://doi.org/10.1016/j.caeai.2024.100234

Shin, D., Lee, J. H., & Kim, K. (2025). An exploratory study on two automated item generators for generating L2 reading test items. *RELC Journal*. Advance online publication. https://doi.org/10.1177/00336882251326284

Stockwell, G. (2018). *Computer-assisted language learning*. Cambridge University Press.

Suvorov, R., & Hegelheimer, V. (2014). Computer-assisted language testing. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 594–613). Wiley.

Vignoli, A., Combei, C. R., & Zappulla, F. (2025). Verso la valutazione automatizzata dell'italiano L2: ETET tra LLM e tecnologie vocali. In C. Bosco, E. Jezek, M. Polignano, & M. Sanguinetti (Eds.), *Proceedings of the 11th Italian Conference on Computational Linguistics (CLiC-it 2025)* (pp. 282–291). ACL Anthology & CEUR Workshop Proceedings.

Vignoli, A. (2026). *Progettazione e validazione di un modello automatizzato per la valutazione dell'italiano come L2: il sistema ETET* [Unpublished master's thesis]. University of Pavia.

Walker, A., & White, G. (2013). *Technology enhanced language learning: Connecting theory and practice*. Oxford University Press.

Winke, P. M., & Isbell, D. R. (2017). Computer-assisted language assessment. In S. Thorne & S. May (Eds.), *Language, education and technology* (pp. 1–13). Springer. https://doi.org/10.1007/978-3-319-02328-1_25-1

Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing Large Language Models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, 56, 150–166. https://doi.org/10.1111/bjet.13494

Zechner, K., & Evanini, K. (2019). *Automated speaking assessment: Using language technologies to score spontaneous speech*. Routledge.

CONTRIBUTORS (ALPHABETICAL ORDER)

Farideh Amirfarhangi holds bachelor's and master's degrees in Applied Linguistics from Azad University in Iran. She is currently a PhD candidate in Linguistics, Terminology, and Intercultural Studies at the University of Parthenope, Naples–Italy, where her doctoral research focuses on eighteenth-century maritime travel accounts. Her research interests include terminology studies, lexical borrowing, cultural and historical discourse, and the representation of Persian culture in European travel literature. She has published a few relevant scholarly works.
Email: farideh.amirfarhangibonab@phds.uniparthenope.it

Claudia Roberta Combei is a tenure-track researcher in Linguistics at the University of Rome "Tor Vergata", where she teaches computational linguistics, phonetics, and sociolinguistics. Previously she worked at the Universities of Bologna and Pavia. In 2019, she earned her PhD from the University of Pisa with a dissertation on human perception and computational processing of foreign-accented speech. Her research combines computational, quantitative, and corpus-based methods, with recent work on figurative language, political discourse, multimodality, and phonetics. She has authored two monographs and numerous journal articles and book chapters. She co-coordinates the RadioCast-It corpus and has contributed to projects funded by UNA Europa, PON DM 1062/2021, and Cariplo. She serves on editorial and advisory boards including *MetBib* (Benjamins), *Humanities and Social Sciences Communications* (Nature), *Materiali Linguistici* (FrancoAngeli), and the *Oxford English Dictionary* (OUP).
Email: claudia.roberta.combei@uniroma2.it

Antonella Giacosa (PhD) has been an EFL teacher in secondary schools for over 20 years and has collaborated with the University of Turin since 2017. She is an educational linguist with an interest in innovative, intercultural,

transdisciplinary and critical approaches to ELT. She has recently published her first book, *Effective Multimodal Interaction for Online and Hybrid Teaching: Data-driven Insights* (2024, Palgrave Macmillan).
Email: antonellamaria.giacosa@unito.it

Paola Giorgis (BA, MA in English and North American Literature and Studies) holds a PhD in Anthropology of Education and Intercultural Education and has a formation in critical sociolinguistics. She has a long experience of teaching English language, literature, and the visual arts. Her main interest regards a critical, intercultural, and interdisciplinary approach to (foreign) languages and foreign language education able to develop an awareness on how words are used and mobilized in education, as well as in public and political discourse. She has published monographs, articles, chapters in collective works, collaborates with many international journals, and has participated to many international conferences. Further info: https://www.iowdictionary.org/author/paola.giorgis
Email: paola.giorgis@iowdictionary.org

Antonella Luporini, PhD in English Linguistics, is Associate Professor in English Language, Translation, and Linguistics at the University of Bologna (LILEC Department), where she teaches English linguistics at undergraduate level and English-Italian translation at graduate level. Since the beginning of her academic career, her research has focused on critical discourse analysis, corpus linguistics, metaphor studies, and the use of new technologies in ELT. She is the Bologna unit coordinator and a member of the Scientific Committee of the Interuniversity Center for Research on Metaphors (CIRM). She was a member of the Editorial Board of *Metbib–Bibliography of Metaphor and Metonymy* (Benjamins) from its founding. Among her institutional roles, she serves as her Department's Delegate for Equity, Diversity, and Inclusion.
https://www.unibo.it/sitoweb/antonella.luporini/
Email: antonella.luporini@unibo.it

Anna Romagnuolo, PhD in Translation Studies, is Associate Professor of English Language and Translation at the University of Naples "L'Orientale", where she serves as President of the University Language Centre. A licensed English and French translator and interpreter, she has coordinated and participated in several EU-funded Erasmus+ projects focusing on digital innovation in education and professional development. She serves on the scientific boards of the international journals *Ácoma*, *e-Pedagium*, and *ALLIeD*. Her research interests include English and corpus linguistics, translation studies, language pedagogy, particularly English for Specific Purposes (ESP), and

discourse analysis, with a focus on political discourse and contemporary narratives.
Email: anna.romagnuolo@unior.it

Andrea C. Valente is a research scholar and educator, working as a sessional lecturer in the Faculty of Education at York University, Canada. Her interdisciplinary research brings together critical discourse analysis, language and education, and approaches informed by studies of the mind and brain. Her current work explores language, pedagogy, and technology, with a particular interest in GenAI. Her research has been published across fields including applied linguistics, education, discourse studies, and the humanities. She is also a licensed schoolteacher from Brazil, having taught across diverse educational contexts.
Email: prof.acval@gmail.com

Francesco Zappulla obtained his Master's degree in Management and Technology from the Technical University of Munich in 2023, following a Bachelor's degree in Business Administration from the University of Verona in 2020. During his graduate studies, he gained professional experience in B2B software startups and venture capital funds. He later joined Etet Srl Società Benefit, where he served as sole director and co-founder. The company focuses on developing AI-based solution for second language (L2) assessment.
Email: francesco@talketet.com

How can English Language Teaching remain critical, human-centered, and pedagogically meaningful in a world increasingly shaped by algorithms? As generative artificial intelligence (GenAI) transforms the ways we teach and learn language, this volume invites readers to reconsider not only the tools available to educators, but also the purposes and responsibilities of language education itself.

Bringing together research, classroom experiences, and project outputs, *Rethinking English Language Teaching in an Algorithmic World* examines the opportunities and challenges of technology-enhanced education across diverse learning and professional contexts. Its contributions address machine translation and post-editing in translator training; corpus-based instruction; AI-assisted language assessment, and the emerging communicative demands of digital freelancing; they also investigate how learners and teachers engage with GenAI, exploring questions of knowledge, authority, and agency in the classroom.

Rather than celebrating technology or rejecting it outright, the volume advocates a critical approach to its integration into English Language Teaching. Ethical concerns surrounding data privacy, bias, fairness, and unequal access are central to this discussion, alongside the need for data literacy: the ability to critically evaluate and responsibly use data and machine-generated outputs. Across the chapters, digital literacy emerges as the capacity to question algorithmic outputs, acknowledge their limitations, and, ultimately, make responsible pedagogical choices. Against this background, the book offers researchers, educators, and practitioners a framework for rethinking what it means to teach and learn English in the algorithmic world we are all, willing or not, living in.

Antonella Luporini, PhD in English Linguistics, is Associate Professor in English Language, Translation, and Linguistics at the University of Bologna (LILEC Department), where she teaches English linguistics at undergraduate level and English-Italian translation at graduate level. Since the beginning of her academic career, her research has focused on critical discourse analysis, corpus linguistics, metaphor studies, and the use of new technologies in ELT. She is the Bologna unit coordinator and a member of the Scientific Committee of the Interuniversity Center for Research on Metaphors (CIRM). She was a member of the Editorial Board of *Metbib–Bibliography of Metaphor and Metonymy* (Benjamins) from its founding. Among her institutional roles, she serves as her Department's Delegate for Equity, Diversity, and Inclusion.

Antonella Giacosa (PhD) has been an EFL teacher in secondary schools for over 20 years and has collaborated with the University of Turin since 2017. She is an educational linguist with an interest in innovative, intercultural, transdisciplinary and critical approaches to ELT. She has recently published her first book, *Effective Multimodal Interaction for Online and Hybrid Teaching: Data-driven Insights* (2024, Palgrave Macmillan).