Verso la Valutazione Automatizzata dell'Italiano L2: ETET tra LLM e Tecnologie Vocali

Anna Vignoli[1,*,†], Claudia Roberta Combei[2,*,†] e Francesco Zappulla[3,†]

[1] *Università degli Studi di Pavia, Corso Strada Nuova 65, 27100 Pavia, Italia*
[2] *Università degli Studi di Roma Tor Vergata, Via Columbia 1, 00133 Roma, Italia*
[3] *ETET S.r.l., Piazza Pinelli 1/7, 16124 Genova, Italia*

Abstract

This paper presents ETET, a web-based application for the automated assessment of L2 proficiency. The main contribution of this work lies in its focus on Italian – a language for which no comparable tools currently exist. Another novelty is the departure from traditional assessment models. In fact, the theoretical framework is grounded in CEFR and Processability Theory, allowing an assessment that reflects the natural developmental sequences of the learners' interlanguage. ETET is not intended to replace human raters, but rather to serve as a complementary tool, since it ensures rapid scoring. Additionally, it has customizable and diversified test formats and items, making it a resource suitable for educational contexts, certification, and selection purposes.

Keywords

valutazione linguistica, italiano L2, CALA, CALT, ICALL

1. Introduzione

L'inserimento delle nuove tecnologie nell'insegnamento e nell'apprendimento delle lingue seconde (L2) rappresenta ormai una pratica consolidata. Molti studi hanno evidenziato come il *Technology-Enhanced Language Learning* (TELL) [1] abbia trasformato il rapporto tra insegnanti e apprendenti e le modalità con cui questi ultimi affrontano il processo di apprendimento linguistico [2]. I cambiamenti portati dal TELL riguardano la progettazione della didattica, la gestione delle lezioni e la valutazione delle competenze acquisite [3]. Alcuni studi sostengono, inoltre, che il TELL rende l'apprendimento più flessibile e favorisce una maggiore autonomia dell'apprendente durante il percorso formativo [4].

L'uso delle nuove tecnologie ha portato a cambiamenti non solo a livello didattico e metodologico, ma anche a livello terminologico; infatti, sono emersi concetti nuovi, quali *Computer/Mobile-Assisted Language Learning* [3], [5], *Digital Language Learning* [6], *Computer-Assisted Language Testing/Assessment* (CALT/CALA) [7], [8] e, più recentemente, *Intelligent Computer-Assisted Language Learning* (ICALL) [9].

L'interesse verso l'ICALL è confermato anche dalla creazione di un gruppo di interesse (SIG ICALL) all'interno del *Computer Assisted Language Instruction Consortium* (CALICO), per favorire lo sviluppo di strumenti didattici basati sull'intelligenza artificiale (IA) e su tecnologie di trattamento automatico del linguaggio (NLP), tra cui *Large Language Model* (LLM), riconoscimento vocale (*Automatic Speech Recognition*, ASR) e sintesi vocale (*Text-to-Speech*, TTS). Alcuni studi recenti hanno mostrato, infatti, che l'IA può essere usata con discreto successo sia nella generazione dei quesiti per i test di lingua [10] sia nella valutazione automatizzata delle competenze linguistiche scritte [11], e orali [12]. L'impiego delle tecnologie NLP e IA nella valutazione linguistica sta ricevendo un'attenzione crescente anche nel contesto italiano. A testimonianza di ciò, il recente volume di Cinganotto e Montanucci [13] sull'IA per l'educazione linguistica dedica un intero capitolo agli approcci automatizzati di quello che le due autrici definiscono *language testing*.

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
* Corresponding authors.
† L'articolo è il risultato della collaborazione tra i tre autori: i paragrafi §1 e § 2 sono scritti da C. R. Combei, i paragrafi §3 e §5 da A. Vignoli, il paragrafo da §4 C. R. Combei, A. Vignoli, e F. Zappulla.

anna.vignoli01@universitadipavia.it (A. Vignoli); claudia.roberta.combei@uniroma2.it (C. R. Combei); francesco@talketet.com (F. Zappulla)
0009-0003-4121-1672 (A. Vignoli); 0000-0003-1884-8205 (C. R. Combei); 0009-0004-1229-0233 (F. Zappulla)

Il nostro lavoro va nella stessa direzione e presenta ETET[2], una web-app commerciale progettata per valutare in maniera automatizzata le competenze linguistiche in una L2. Sebbene ETET sia una piattaforma multilingue – attualmente disponibile per l'inglese e per l'italiano – il presente studio si concentra esclusivamente sul modulo dedicato all'italiano L2.

L'articolo è strutturato come segue: il paragrafo 2 illustra le motivazioni e gli obiettivi della ricerca; il paragrafo 3 delinea il quadro teorico di riferimento; il paragrafo 4 è dedicato alla descrizione di ETET, sia dal punto di vista tecnico sia per quanto riguarda la progettazione dei quesiti, le modalità di assegnazione dei punteggi e la validazione; infine, il paragrafo 5 presenta le prospettive future e alcune considerazioni conclusive.

2. Motivazioni e Obiettivi

Per quanto a nostra conoscenza, ETET rappresenta il primo strumento per la valutazione completamente automatizzata delle competenze linguistiche in italiano L2, una lingua parlata da circa 3.287.300 parlanti non-nativi, secondo Ethnologue[3]. Questo numero riflette la presenza di numerose comunità di parlanti di origine italiana di seconda o terza generazione all'estero (i cosiddetti *heritage speakers)* [14], così come l'ampia rete di promozione linguistica e culturale coordinata dal Ministero degli Affari Esteri, composta da 88 Istituti Italiani di Cultura, 8 istituti statali omnicomprensivi, 43 scuole italiane paritarie, 7 sezioni italiane presso scuole europee, 79 sezioni italiane presso scuole straniere internazionali, 2 scuole non paritarie, 422 comitati della Società Dante Alighieri, attivi in tutto il mondo[4]. In questi contesti culturali, spesso frequentati da apprendenti e da parlanti di italiano L2, emerge l'esigenza di strumenti affidabili per la valutazione delle competenze linguistiche, sia a fini didattici che certificativi.

Infatti, negli ultimi anni, la domanda di strumenti per la valutazione delle competenze linguistiche in italiano ha registrato un aumento, anche in risposta a specifici interventi normativi [15], [16]. Ad esempio, l'art. 14, comma 1, lett. a-bis), del D.L. 4 ottobre 2018, n. 113 (c.d. Decreto Sicurezza, in vigore dal 5 ottobre 2018), convertito, con modificazioni, dalla legge 1 dicembre 2018, n. 136, ha inserito l'art. 9.1 nella legge n. 91/1992, subordinando la concessione della cittadinanza italiana al possesso di un'adeguata conoscenza della lingua italiana, non inferiore al livello B1 del Quadro comune europeo di riferimento per la conoscenza delle lingue (QCER). Per ottenere una certificazione linguistica di italiano L2 di livello B1 del QCER, i richiedenti devono superare uno degli esami di lingua riconosciuti (ad es., CELI 2, CILS B1, ecc.)[5]. Sono previsti alcuni casi di esonero per chi possiede un titolo di studio conseguito in Italia o per i titolari di permesso di soggiorno UE di lungo periodo, i quali devono comunque aver superato in precedenza un esame di lingua italiana di livello A2.

Analogamente, gli studenti universitari provenienti da paesi non appartenenti all'Unione Europea sono tenuti a superare una prova linguistica per accertare il livello B2 del QCER se intendono immatricolarsi a corsi erogati in lingua italiana [17].

In ambito lavorativo, la valutazione delle competenze linguistiche di italiano riveste un ruolo importante nei settori del Business Process Outsourcing (BPO), compresi i call center delocalizzati, dove lavorano numerosi parlanti L2 [18]. In questi casi, la qualità del servizio offerto dalle aziende BPO dipende anche dalla competenza linguistica dei candidati.

La presenza di tutte queste situazioni sociali, culturali, didattiche e professionali in cui è richiesta una certificazione delle competenze linguistiche in italiano L2, insieme ai tempi di attesa spesso lunghi per sostenere i relativi esami, evidenzia la necessità di costruire strumenti di valutazione che siano più accessibili, rapidi e facilmente distribuibili su larga scala. In questo contesto, ETET si propone come strumento di supporto alla valutazione tradizionale, con l'obiettivo di fornire stime automatiche della competenza linguistica in italiano L2 che siano affidabili, immediate e coerenti con i livelli del QCER. Lungi dal voler sostituire la valutazione umana, ETET mira a supportarla, offrendo una soluzione utile in contesti ad alta richiesta o come complemento ai percorsi didattici e certificativi (ad es., situazioni in cui è richiesto il risultato in tempo reale, *placement test*, prove intermedie, simulazioni, esercitazioni, ecc.).

3. Quadro Teorico

Dall'analisi dei documenti scientifici prodotti dagli enti certificatori dell'italiano come L2 emerge una discrepanza significativa tra le modalità di valutazione adottate nei test ufficiali e quanto descritto nella letteratura sull'acquisizione di una L2, in particolare in riferimento alle fasi di sviluppo dell'interlingua. Tale incongruenza si traduce in una potenziale discontinuità tra i livelli di competenza linguistica effettivamente raggiunti dagli apprendenti e quelli formalmente

[2] Sito web ETET: https://www.talketet.com/ (accesso 06/06/2025).
[3] Informazioni dettagliate disponibili qui: http://ethnologue.com/language/ita/ (accesso 06/06/2025).
[4] Informazioni dettagliate disponibili qui: https://www.esteri.it/it/diplomazia-culturale-e-diplomazia-scientifica/cultura/promozionelinguaitaliana/ (accesso 06/06/2025).
[5] Informazioni dettagliate disponibili qui: https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legge:2018-10-04;113~art14 (accesso 06/06/2025).

certificati. Ne deriva la necessità di elaborare strumenti di valutazione fondati su un inquadramento teorico solido e coerente, in grado di giustificare e sostenere i criteri adottati nella misurazione della competenza linguistica.

Il framework teorico di riferimento per il nostro lavoro è rappresentato dal QCER, standard europeo per la valutazione delle competenze linguistiche promosso dal Consiglio d'Europa. Gli obiettivi principali del QCER sono quelli di promuovere la diffusione del plurilinguismo in Europa, fornire strumenti comuni a chi opera nell'ambito dell'educazione linguistica e della valutazione linguistica e favorire il riconoscimento e l'equiparazione dei titoli e dei certificati linguistici [19]. Essendo concepito come strumento valido per tutte le lingue d'Europa (ma è sempre più usato anche nel resto del mondo), il QCER non si delinea come uno strumento prescrittivo, bensì propone una descrizione qualitativa delle competenze linguistiche che caratterizzano ogni livello [19].

Dopo aver fornito un'impalcatura generale, i singoli stati hanno sentito la necessità di trasporre gli indicatori del QCER nei contesti specifici delle varie lingue. Da questa esigenza sono nate pubblicazioni come [20], per l'italiano, con l'obiettivo di identificare descrittori linguistici specifici per ogni livello di competenza. Accanto all'iniziativa italiana sono sorti progetti anche per l'inglese, lo spagnolo, il francese e il tedesco.

Il quadro teorico su cui si fonda ETET è quindi il risultato delle indicazioni generali contenute nel QCER, nella sua trasposizione specifica teorizzata per l'italiano [19], [20] e dal confronto e dalla disamina dei documenti scientifici prodotti dagli enti certificatori di italiano come L2. È stato inoltre valutato di affiancare questa cornice teorica alla teoria psicolinguistica della Processabilità (Processability Theory, PT) – in particolare per quanto riguarda gli aspetti morfosintattici della lingua – la quale si propone di spiegare le sequenze evolutive che si verificano all'interno di una L2. L'allineamento tra questi due approcci consente di definire un sistema valutativo capace di rilevare non solo il livello raggiunto, ma anche la plausibilità evolutiva delle competenze espresse.

La PT [21], teorizzata nel 1998 da Manfred Pienemann, sostiene che esiste un insieme universale e gerarchicamente ordinato di procedure di elaborazione dell'output che vengono acquisite nel tempo e che non sono influenzate dalla L1 [22], [23]. Tali procedure si presentano in ordine gerarchico implicazionale, ovvero, la procedura di un livello più basso è un prerequisito necessario per il funzionamento della procedura del livello successivo [21]. Le procedure sono attivate nel seguente ordine: procedura lemmatica, procedura categoriale, procedura sintagmatica, procedura frasale, procedura subordinante.

Il primo livello di acquisizione è rappresentato dalla procedura lemmatica che prevede un apprendimento di tipo formulaico. In questa fase vengono identificati elementi lessicali singoli e invariabili, senza far ricorso a processi cognitivi specifici se non a quello della memoria lessicale che porta all'acquisizione di *chunk* e *type* (ad es., ciao). Successivamente si passa al secondo livello, ovvero alla procedura categoriale, in cui l'apprendente inizia a distinguere le categorie lessicali e grammaticali (ad es., nome, verbo, ecc.) degli elementi che ha già imparato e a produrre alcune marche morfologiche. Tuttavia, non vi è ancora comunicazione tra i vari elementi della frase. Il terzo livello è quello della procedura sintagmatica che si divide in due sottolivelli: l'accordo entro il sintagma nominale e l'accordo entro il sintagma verbale. Il primo sottolivello prevede una forma iniziale di accordo all'interno del sintagma, ovvero, l'apprendente riconosce la testa del sintagma e inizia a marcare i tratti grammaticali al suo interno. Nel secondo sottolivello, l'apprendente incomincia a costruire sintagmi verbali sempre più complessi. Raggiunto il quarto livello, ovvero quello della procedura frasale, lo scambio di informazioni avviene tra sintagmi diversi. Infine, la procedura subordinante rappresenta l'ultimo livello, cioè dove avviene lo scambio di informazioni tra frase principale e frase subordinate.

La validità universale del framework proposto dall'unione del QCER e della PT si dimostra particolarmente adatta per il progetto di ETET, il cui l'obiettivo a lungo termine è quello di riuscire a coprire e valutare in maniera congruente e scientificamente motivata un numero sempre maggiore di lingue. In effetti, la progettazione teorica di ETET si basa sull'integrazione di due prospettive teoriche complementari: da un lato il QCER e le sue attuazioni più pratiche, che offrono una cornice descrittiva per la valutazione delle competenze linguistiche, dall'altro la PT, che fornisce un modello psicolinguistico per comprendere le tappe evolutive dell'interlingua. L'unione di questi due approcci consente di costruire un sistema di valutazione che tenga conto sia del livello di competenza manifestato, sia della sua coerenza, seguendo le naturali traiettorie di acquisizione dell'interlingua.

4. Descrizione ETET

La web-app ETET è stata progettata per offrire una valutazione automatizzata delle competenze linguistiche scritte e orali, sia attraverso domande chiuse che aperte. Le competenze valutate riguardano la produzione e l'interazione orale, l'ascolto, la comprensione del testo, la produzione scritta e la grammatica. Per quanto riguarda la produzione orale, il sistema valuta anche l'intelligibilità e la pronuncia [24]. ETET restituisce

punteggi su scala 0–100, pesati in base alla difficoltà e alla tipologia della domanda; i punteggi sono mappati sui livelli del QCER sia a livello globale sia a livello della singola abilità valutata.

4.1. Caratteristiche Tecniche

La piattaforma integra algoritmi di feedback in tempo reale, LLM e tecnologie di ASR, prendendo spunto da lavori recenti nell'ambito dell'*Automated Essay Scoring* (AES) [11] e dell'*Automated Speaking Assessment* (ASA) [25].

Dal punto di vista dell'architettura, l'applicativo ETET utilizza un ambiente Linux (Debian). Il back-end è realizzato in Ruby on Rails 8, mentre l'interfaccia di gestione (back-office) è realizzata in Vue.js. Il modulo front-end per l'utente finale si basa anch'esso su Vue.js. L'autenticazione avviene tramite e-mail e password e la sessione viene verificata tramite token JWT.

Il sistema ASR è basato su AzureAI[6], che sfrutta il modello Whisper di OpenAI[7]. Gli input vocali vengono acquisiti tramite browser in formato .ogg o .wav (canale mono) e non vengono normalizzati. I parametri di decodifica non sono attualmente configurabili.

Le domande aperte sono valutate tramite il GPT-4o di OpenAI[8], un modello accessibile via API, con prompt personalizzati per ciascun tipo di test e domanda. La valutazione è asincrona, avviene in background e viene restituita solo al termine del test.

Tutti i dati relativi agli esami, come ad esempio, i testi e i file audio delle risposte prodotti dagli utenti, sono salvati in un database PostgreSQL, ospitato in cloud sull'infrastruttura Azure, in conformità con il GDPR. I dischi sono criptati con chiavi gestite dalla piattaforma e viene applicata una policy di snapshot periodico per garantire la sicurezza e la conservazione dei dati.

4.2. Test e Domande

La web-app ETET permette di creare diverse tipologie di test a seconda delle necessità dei singoli esaminatori. Il punto di partenza per la realizzazione di un test è la definizione di quale sia l'obiettivo della valutazione, il che implica caratteristiche specifiche in termini di contenuto e tempistiche di somministrazione [26]. I test possono essere costruiti per testare tutte e quattro le abilità linguistiche fondamentali (lettura, ascolto, scrittura e parlato) oppure possono essere personalizzati per andare ad indagare le abilità linguistiche che maggiormente interessano all'esaminatore.

La procedura di creazione di un test è un processo incrementale: per ciascuna abilità fondamentale viene creato un questionario che si compone di domande scelte da un database predisposto da esperti interni e selezionate in modo da coprire tutti gli aspetti che si vogliono indagare inerentemente a quella competenza linguistica. L'insieme dei questionari andrà a costituire la forma complessiva dell'esame (v. Figura 1).

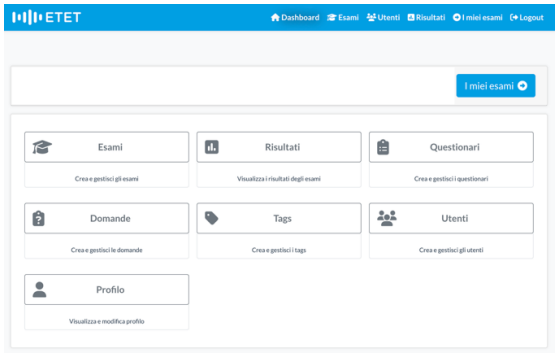

Figura 1: Lato back-office di ETET.

A prescindere dalle tipologie di domande prescelte, ciascun questionario viene costruito inserendo item con un determinato bilanciamento di complessità attesa. In particolare, ogni domanda è caratterizzata da un livello di difficoltà associato ai livelli proposti dal QCER. In ogni questionario si trovano il 10% di domande del livello A1 e del livello C2 mentre tutti gli altri livelli (A2, B1, B2, C1) sono uniformemente rappresentati da un 20% di domande.

Una caratteristica di ETET è la possibilità di impostare il numero di domande e la durata del test in base alle esigenze dell'utente o alle specificità del contesto didattico e valutativo. Ciascuna domanda viene identificata a partire da una serie di caratteristiche che ne descrivono tipologia e complessità. Queste vengono dunque catalogate secondo i seguenti parametri: abilità linguistica testata – attraverso un'etichetta identificativa (lettura, ascolto, scrittura e parlato); lingua del test (in questo caso, l'italiano); livello di difficoltà secondo il QCER (A1, A2, B1, B2, C1, C2); oggetto epistemico indagato (ad es., articoli determinativi, forma passiva ecc.).

I quesiti possono essere, inoltre, divisi in due macrocategorie: domande a risposta chiusa e domande a risposta aperta, distinte in base al tipo di produzione richiesta all'utente e alla modalità di valutazione. In entrambi i casi, i soggetti dispongono di un tempo

[6] Informazioni dettagliate disponibili qui: https://azure.microsoft.com/en-us/solutions/ai (accesso 06/06/2025).
[7] Informazioni dettagliate disponibili qui: https://openai.com/index/whisper/ (accesso 06/06/2025).
[8] Informazioni dettagliate disponibili qui: https://platform.openai.com/docs/models/gpt-4o (accesso 06/06/2025).

limitato, misurato in secondi, per formulare e inserire la propria risposta.

Le domande a risposta chiusa (v. Figura 2) presuppongono generalmente una sola risposta esatta o un numero limitato e comunque predefinito di risposte possibili. Fanno parte di questa tipologia di domande:

- *Domande a scelta multipla*, dove ad una domanda vengono associate più risposte possibili, di cui una soltanto è corretta e le altre svolgono la funzione di distrattori.
- *Domande a completamento*, in cui viene presentata una frase caratterizzata da uno o più spazi vuoti in cui l'utente deve inserire una o più parole. In questa tipologia di domande, nella fase di creazione, è stata prestata molta attenzione ad inserire, laddove necessario, tutti i possibili sinonimi che possono essere indicati dalla persona testata.
- *Dettato*, nel quale la persona testata deve scrivere ciò che riesce a comprendere dall'audio presente nella domanda.
- *Ricostruzione di una frase*, in cui l'utente sente un audio in cui sono contenuti vari pezzi di frase presentati in ordine sparso e li deve ricostruire correttamente.

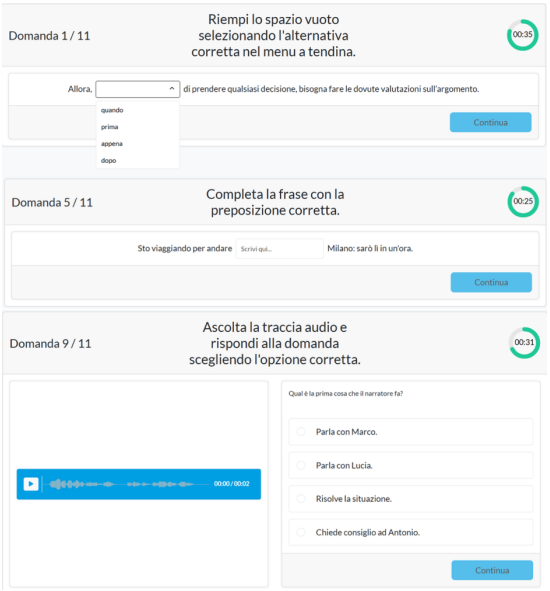

Figura 2: Alcuni esempi di domande a risposta chiusa.

Nella costruzione dei test si è cercato di inserire il minor numero possibile di domande a scelta multipla per diminuire la possibilità che l'utente indovini la risposta. Si è cercato invece, laddove possibile, di sostituire questa tipologia con le domande a completamento, le quali permettono inoltre di ottenere un input linguistico più autentico, essendo necessaria una produzione da parte dell'utente.

Le domande aperte (v. Figura 3) prevedono un'elaborazione linguistica attiva e autonoma da parte del candidato e riguardano le abilità di scrittura e di parlato. Queste domande permettono di valutare competenze complesse e integrative. Fanno parte della tipologia di domande aperte:

- *Componimenti brevi*, che prevedono la scrittura di un breve testo su argomenti vari e con differenti variazioni diafasiche.
- *Descrizioni di immagini*, ovvero viene dato come input un'immagine e il soggetto deve fornirne una descrizione dettagliata dello stimolo.
- *Esposizione del proprio punto di vista*, tramite la scrittura di un testo che evidenzi la propria posizione su un tema (ad es., cambiamento climatico).
- *Riassunto*, la persona testata deve produrre un riassunto del testo che ha appena letto o dell'audio che ha appena sentito.

Figura 3: Alcuni esempi di domande a risposta aperta.

Nelle domande aperte di scrittura viene richiesta come risposta un minimo di 50 parole, mentre in quelle di parlato sono richiesti almeno 20 secondi di registrazione audio in tempo reale.

La presenza di entrambe le tipologie di domanda è giustificata dall'esigenza di ottenere una valutazione più completa e bilanciata della competenza linguistica (v. Figura 4). Infatti, come osservano nella letteratura [27], una valutazione linguistica efficace dovrebbe integrare quesiti che misurano sia la conoscenza linguistica formale sia la capacità di utilizzare efficacemente tale conoscenza in contesti reali. Questa definizione è in linea con la teorizzazione del QCER [28] che vede le competenze linguistiche dei parlanti di una L2 come

orientate all'azione, intendendo gli apprendenti come agenti sociali. Attraverso la lingua, i parlanti devono essere in grado di interagire efficacemente e comunicare in vari contesti sociali, culturali e professionali.

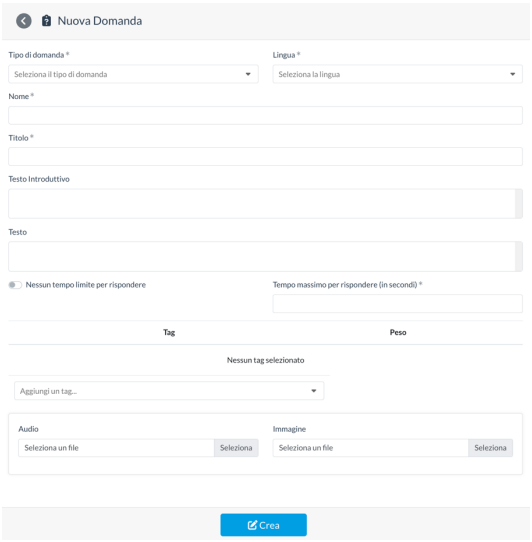

Figura 4: Impostazione domande.

Partendo da questo assunto, si è delineata la necessità di proporre compiti, domande, materiale multimediale (ad es., registrazioni audio, immagini) e testi che l'informante potesse incontrare nell'uso reale della lingua e nelle più varie situazioni comunicative.

A ciascuna domanda è associato un peso (1, 2, 3 punti) che rappresenta il punteggio che si può ottenere rispondendo correttamente. Domande con maggior necessità di rielaborazione e sforzo cognitivo da parte del soggetto avranno pesi maggiori. Domande a risposta chiusa, come ad esempio, quelle a risposta multipla o di completamento, avranno un peso di 1 punto, dettati o ricostruzioni di frasi 2 punti mentre produzioni orali o elaborati scritti 3 punti.

Le domande a risposta chiusa prevedono una valutazione booleana del tipo corretto/errato. Per la produzione scritta e parlata, invece, la valutazione è affidata ad un LLM in grado di generare punteggi di AES e ASA (v. paragrafo 4.1). Il primo valuta la correttezza grammaticale, la scelta lessicale, la coerenza e l'adesione al tema, la comprensione, la coesione testuale e il contenuto della risposta fornita dall'utente. Il punteggio dato in centesimi dal modello si ripartisce nel seguente modo tra 5 parametri: *structure and grammar* = 20%, *content and argumentation* = 30%, *vocabulary* = 20%, *comprehension and adherence to the topic* = 20% e *pragmatics and cohesion* = 10%. Diversi studi hanno evidenziato come i sistemi AES mostrino un'elevata concordanza con i giudizi di valutatori umani esperti [29], [30].

Il modello ASA – dopo aver prodotto una trascrizione fedele dell'audio registrato in tempo reale dall'utente – usa come metriche di valutazione il paradigma *Complexity Accuracy Fluency* (CAF) [24], analizzando la produzione orale del parlante, attraverso 4 parametri: *fluency*, *accuracy*, *completeness*, *pronunciation*. I valori relativi a questi 4 parametri rappresentano il 5% del punteggio finale nelle domande di produzione orale. Il restante 95% è rappresentato dalle valutazioni relative ai parametri discussi prima, redistribuiti come segue: *structure and grammar* = 20%, *content and argumentation* = 25%, *vocabulary* = 20%, *comprehension and adherence to the topic* = 20% e *pragmatics and cohesion* = 10%. La letteratura [31] sostiene che le tecnologie ASA mostrano numerosi vantaggi rispetto alla valutazione orale tradizionale, tra cui una maggiore efficienza e rapidità nei processi di somministrazione del test e di elaborazione dei risultati, ma anche una riduzione dell'incidenza di *bias* umani, con conseguente incremento della coerenza dei punteggi assegnati.

Una volta che l'utente completa il test, il sistema calcola i punteggi per ciascuna domanda, come descritto sopra, e riporta il voto finale ottenuto in ciascuna sezione. Il voto è calcolato come rapporto tra i punti ottenuti e i punti totali ottenibili relativi a quell'abilità linguistica (v. formula (1)).

$$score_i = \frac{pt_{fatti,i}}{pt_{tot,i}} \% \quad (1)$$

Il punteggio finale del test, invece, si ottiene come media dei punteggi ottenuti per ciascuna abilità linguistica, in modo da uniformarne l'importanza (v. formula (2), dove n è il numero di abilità testate).

$$score_{finale} = \frac{\sum_{i=1}^{n} score_i}{n} \% \quad (2)$$

Per allineare l'output del test con quello degli enti certificatori, il punteggio in percentuale ottenuto viene mappato sulle fasce stabilite dal QCER. Oltre al punteggio in percentuale, l'utente finale otterrà, quindi, un voto espresso come livello QCER (da A1 a C2) per ciascuna competenza testata e un voto finale per lo svolgimento complessivo del test (v. Figura 5).

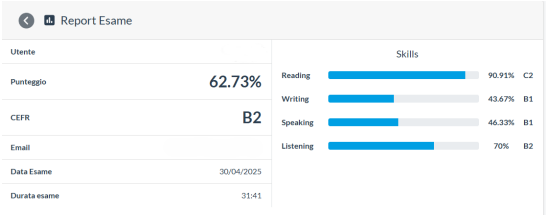

Figura 5: Esempio di feedback.

Al termine della prova, la piattaforma prevede la produzione automatica di un feedback descrittivo personalizzato il quale, tramite le tecnologie generative, restituisce un'analisi complessiva della performance dell'utente. Allo strumento vengono forniti i risultati ottenuti nel test e le risposte alle domande a composizione libera di scrittura e di parlato. Questi dati vengono analizzati e il feedback che viene prodotto evidenzia sia i punti di forza della competenza linguistica della persona testata sia le aree in cui si sono riscontrate maggiori difficoltà, fornendo anche spunti e consigli per un miglioramento delle abilità testate.

4.3. Validazione

Come evidenziato nel paragrafo 1, negli ultimi anni, i LLM hanno mostrato un potenziale crescente nell'ambito della valutazione automatica delle competenze linguistiche nelle L2. In particolare, i LLM si confermano strumenti promettenti, soprattutto in contesti didattici e valutativi a basso rischio. Tuttavia, nella letteratura recente sono stati riscontrati alcuni limiti legati alla capacità dei LLM di cogliere aspetti discorsivi complessi, alla variabilità delle loro prestazioni nel tempo e alla presenza di *bias* inferenziali, socioculturali e sociodemografici [32], [33]. Considerata la mancanza di un consenso nella letteratura circa la validità e l'affidabilità dei LLM come strumenti di valutazione linguistica, e alla luce dell'assenza di protocolli standardizzati e condivisi per la loro validazione, si è ritenuto opportuno intraprendere un primo tentativo di validazione dello strumento ETET per l'italiano L2.

Per verificare la validità del modello e la coerenza nell'assegnazione dei punteggi, sono state condotte alcune sessioni di prova su risposte a domande aperte. Sono state selezionate 3 domande di produzione scritta (PS1, PS2, PS3) e 3 di produzione orale (PO1, PO2, PO3) e per ciascuna domanda è stata formulata una risposta. Ogni risposta è stata valutata dal sistema per 10 iterazioni consecutive, mantenendo, quindi, invariato l'input, al fine di osservare la stabilità dei punteggi assegnati a parità di contenuto. Per ciascuna iterazione sono stati registrati i punteggi relativi a tutti i parametri considerati. Le domande di produzione orale sono state testate sia su un parlante di genere maschile (PO1m, PO2m, PO3m) sia su una parlante di genere femminile (PO1f, PO2f, PO3f), allo scopo di verificare l'eventuale presenza di *bias* di genere nei punteggi assegnati dal modello. Successivamente sono stati calcolati il valore medio e la deviazione standard per ciascun parametro di valutazione associato a ogni domanda: X1 = *structure and grammar*, X2 = *content and argumentation*, X3 = *vocabulary*, X4 = *comprehension and adherence to the topic*, X5 = *pragmatics and cohesion*, X6 = *fluency*, X7 = *accuracy*, X8 = *completeness*, X9 = *pronunciation*. Come discusso nel paragrafo 4.2, per la valutazione della produzione scritta (PS1, PS2, PS3) sono stati considerati i primi 5 parametri, mentre per la produzione orale (PO1, PO2, PO3) tutti e 9 i parametri.

La coerenza dei punteggi è stata valutata attraverso il coefficiente di variazione (CV), ottenuto dal rapporto percentuale tra deviazione standard e valore medio, che descrive la dispersione relativa per ogni oggetto indagato (v. formula (3)).

$$sCV = \frac{\sigma}{\mu}\% \tag{3}$$

Coefficienti di variazione alti sono sintomatici di un'elevata dispersione nei dati, dunque, una limitata coerenza del modello nel giudicare i diversi parametri; viceversa, coefficienti di variazione bassi sono indicativi di una ridotta dispersione relativa. La Tabella 1 mostra i coefficienti di variazione per ciascun tipo di domanda aperta.

È possibile notare come tutte le misurazioni, ad eccezione del parametro X1 nella seconda domanda di produzione orale (con voce femminile), si trovino al di sotto del limite del 10%, fissato come soglia di accettabilità dei risultati (v. Tabella 1). Ne consegue una buona coerenza da parte del modello nell'assegnazione dei punteggi nella nostra sessione di valutazione.

Inoltre, l'esperimento di validazione ha evidenziato che il genere non sembra avere un ruolo rilevante nei valori ottenuti. Si può quindi ipotizzare che il modello sia indifferente a questa variabile, oppure che il genere, andando a sommarsi ad una serie di altri fattori, come il tono di voce, la distanza dal microfono, o la velocità di eloquio, non assuma un'importanza determinante nel calcolo dei punteggi. Questo esito risulterebbe coerente con l'obiettivo di costruire uno strumento robusto e non soggetto a *bias* algoritmici.

Tabella 1
Valori CV per ciascun tipo di domanda

Tipo	X1	X2	X3	X4	X5	X6-9
PS1	4,9%	3,8%	3,9%	6,5%	2,6%	\
PS2	5,1%	3,8%	4,4%	4,1%	2,6%	\
PS3	2,5%	2,8%	3,6%	2,7%	3,9%	\
PO1f	2,5%	3,2%	7,0%	3,9%	4,8%	0,0%
PO2f	10,6%	4,0%	8,1%	4,6%	8,8%	0,0%
PO3f	4,1%	4,1%	5,6%	3,9%	2,8%	0,0%
PO1m	6,7%	4,0%	5,9%	2,3%	6,0%	0,0%
PO2m	7,8%	2,7%	6,5%	3,4%	4,9%	0,0%
PO3m	7,4%	3,9%	5,5%	2,8%	6,0%	0,0%

5. Conclusioni e Sviluppi Futuri

Il presente studio ha illustrato la progettazione e lo sviluppo di ETET, una web-app commerciale per la

valutazione automatizzata delle competenze linguistiche nelle L2. In particolare, la portata innovativa della ricerca è rappresentata dal fatto che la lingua presa in esame sia l'italiano, lingua per la quale – almeno nella conoscenza degli autori – non esistono strumenti di questo tipo. Ulteriore novità è rappresentata dalla scelta di non seguire le modalità di valutazione "tradizionali", ma di sviluppare un quadro teorico – fondato sui descrittori del QCER e sulla PT – che tenesse effettivamente in considerazione l'uso pratico della lingua e le sequenze evolutive naturali dell'interlingua degli individui.

La piattaforma non si propone di sostituire i valutatori umani, ma nasce dalla volontà di essere uno strumento di supporto. ETET, grazie alle tecnologie impiegate, consente infatti un'elevata efficienza nella somministrazione delle prove e grande rapidità nella loro valutazione; allo stesso tempo, la possibilità di personalizzare i test, la varietà delle domande proposte e il feedback personalizzato lo renderebbero uno strumento adatto sia a contesti didattici sia a fini certificativi e pre-selettivi (ad es., in ambito lavorativo o universitario). Infine, l'utilizzo di tecnologie AES e ASA permette di ridurre l'incidenza di *bias* umani e allo stesso tempo di incrementare coerenza e affidabilità nell'assegnazione dei punteggi.

Finora, gli sforzi della nostra ricerca si sono concentrati prevalentemente sullo sviluppo e sull'implementazione della piattaforma ETET per l'italiano. Per questo motivo, non è stata ancora condotta una valutazione strutturata e sistematica dello strumento su un campione di parlanti non nativi di italiano. Oltre alla validazione descritta nel paragrafo 4.3, le uniche ulteriori osservazioni preliminari sono state raccolte da un numero molto ristretto di persone, coinvolte in un test esplorativo della durata di circa mezz'ora. Questo test aveva l'obiettivo di ottenere i primi riscontri sul funzionamento generale della web-app.

Tra le prospettive future del lavoro è prevista la realizzazione di uno studio pilota, attualmente in fase di progettazione nell'ambito di una tesi magistrale, finalizzato a una valutazione più approfondita e sistematica dello strumento. Il protocollo sperimentale relativo alla fase di valutazione di ETET prevederà la somministrazione del test a un campione di 50 informanti con diversi livelli di competenza linguistica in italiano L2. I dati raccolti dallo studio pilota saranno usati per definire un *benchmark* di riferimento, mediante il confronto con un *gold standard* elaborato da esperti valutatori dell'italiano L2, con le autovalutazioni fornite dagli stessi 50 partecipanti e con i dati raccolti da un gruppo di controllo costituito da 5 parlanti nativi di italiano.

Parallelamente, verrà condotta un'analisi qualitativa dei feedback ricevuti dagli informanti sull'usabilità della piattaforma ETET.

Ringraziamenti

Gli autori desiderano esprimere la loro gratitudine a Maurizio Olivieri, a Enrico Reboscio e all'intero gruppo di sviluppo di DotVocal Innovation per il prezioso supporto tecnico fornito.

Bibliografia

[1] S. C. Yang, Y. J. Chen, Technology-enhanced language learning: A case study, Computers in Human Behavior 23.1 (2007) 860–879. doi:10.1016/j.chb.2006.02.015.

[2] A. Walker, G. White, Technology Enhanced Language Learning: Connecting Theory and Practice, Oxford University Press, Oxford, UK, 2013.

[3] G. Stockwell, Computer-Assisted Language Learning, Cambridge University Press, Cambridge, UK, 2018.

[4] J. M. Howard, A. Scott, Any time, any place, flexible pace: Technology-enhanced language learning in a teacher education programme, Australian Journal of Teacher Education 42.6 (2017) 51–68. doi:10.14221/ajte.2017v42n6.4

[5] J. Burston, MALL: The pedagogical challenges, Computer Assisted Language Learning 27.4 (2014) 344–357. doi:10.1080/09588221.2014.914539.

[6] P. Li, Y. J. Lan, Digital language learning (DLL): Insights from behavior, cognition, and the brain, Bilingualism: Language and Cognition 25.3 (2022) 361–378. doi:10.1017/S1366728921000353.

[7] R. Suvorov, V. Hegelheimer, Computer-Assisted Language Testing, in: A. J. Kunnan (Ed.), The Companion to Language Assessment, Wiley, Hoboken, New Jersey, USA, 2014, pp. 594-613.

[8] P. M. Winke, D. R. Isbell, Computer-Assisted Language Assessment, in: S. Thorne, S. May (Eds.), Language, Education and Technology, Springer, Cham, Switzerland, 2017, pp. 1–13. doi:10.1007/978-3-319-02328-1_25-1.

[9] T. Heift, Intelligent Computer Assisted Language Learning, in: H. Mohebbi, C. Coombe (Eds.), Research Questions in Language Education and Applied Linguistics, Springer, Cham, Switzerland, 2021, pp. 655–658. doi:10.1007/978-3-030-79143-8_114.

[10] N. Donati, M. Periani, P. Di Natale, G. Savino, P. Torroni, 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), Pisa, Italy, 2024, pp. 325–334.
[11] F. Yavuz, Ö. Çelik, G. Yavaş Çelik, Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments, British Journal of Educational Technology 56.1 (2025) 150–166. doi:10.1111/bjet.13494.
[12] K. Nebhi, G. Szaszák, 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, INCOMA Ltd., Shoumen, Bulgaria, 2023, pp. 769–776.
[13] L. Cinganotto, G. Montanucci, Intelligenza Artificiale per l'educazione linguistica, UTET Università, Torino, 2025.
[14] I. Caloi, J. Torregrossa, Home and school language practices and their effects on heritage language acquisition: A view from heritage Italians in Germany, Languages 6.1 (2021). doi:10.3390/languages6010050.
[15] L. Cinganotto, Language testing online: Sperimentazioni sulla lingua italiana, Italiano LinguaDue 16.1 (2024) 292–310. doi:10.54103/2037-3597/23842.
[16] M. Mezzadri, P. Vecchio, Accessibilità e inclusività nella certificazione linguistica: Uno studio di caso nell'italiano L2, Italiano LinguaDue 15.2 (2023) 304–327. doi:10.54103/2037-3597/21952.
[17] B. Samu, S. Scaglione, Il requisito della conoscenza della lingua italiana e la sua certificazione, in: M. Benvenuti, P. Morozzo della Rocca (Eds.), Università e studenti stranieri. Un'analisi giuridica dell'accesso all'istruzione superiore in Italia da parte dei cittadini di Paesi terzi, Editoriale Scientifica, Napoli, 2024, pp. 159–174.
[18] C. R. Combei, Speaking Italian with a Twist: A Corpus Study of Perceived Foreign Accent, Franco Angeli, Milano, 2023.
[19] Consiglio d'Europa, Quadro Comune Europeo di Riferimento per le Lingue: Apprendimento, Insegnamento, Valutazione, Consiglio d'Europa, Strasburgo, Francia, 2001.
[20] B. Spinelli, F. Parizzi, Profilo della lingua italiana. Livelli di riferimento del QCER A1, A2, B1, B2, La Nuova Italia/RCS Libri, Firenze, 2010.
[21] M. Pienemann, Language processing and second language development: Processability Theory, John Benjamins, Amsterdam, Netherlands, 1998.
[22] B. Van Patten, M. Smith, A. G. Benati, Key Questions in Second Language Acquisition: An Introduction, Cambridge University Press, Cambridge, UK, 2019.
[23] B. VanPatten, G. D. Keating, S. Wulff, Theories in Second Language Acquisition: An Introduction, Routledge, London, UK, 2020.
[24] T. Chau, A. Huensch, The relationships among L2 fluency, intelligibility, comprehensibility, and accentedness: A meta-analysis, Studies in Second Language Acquisition 47.1 (2025) 282–307. doi:10.1017/S0272263125000014.
[25] K. Zechner, K. Evanini, Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech, 1st. ed., Routledge, Abingdon, UK, 2019.
[26] M. Vedovelli, Manuale della certificazione dell'italiano L2, Carrocci Editore, Roma, 2005.
[27] L. F. Bachman, A. Palmer, Language Testing in Practice, Oxford University Press, Oxford, UK, 1996.
[28] P. Deane, On the relation between automated essay scoring and modern views of the writing construct, Assessing Writing 18.1 (2013) 7–24. doi:10.1016/j.asw.2012.10.002.
[29] B. Bridgeman, C. Trapani, Y. Attali, Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country, Applied Measurement in Education 25.1 (2012) 27–40. doi:10.1080/08957347.2012.635502.
[30] A. Housen, F. Kuiken, I. Vedder, Complexity, accuracy and fluency, in: A. Housen, F. Kuiken, I. Vedder (Eds.). Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA, John Benjamins, Amsterdam, Netherlands, 2012, pp. 1–20. doi:10.1075/lllt.32.01hou.
[31] N. Khabbazbashi, J. Xu, E. D. Galaczi, Opening the Black Box: Exploring Automated Speaking Evaluation, in: B. Lanteigne, C. Coombe, J. D. Brown (Eds.), Challenges in Language Testing Around the World, Springer, Singapore, 2021, pp. 333-343.
[32] A. Arronte Alvarez, N. Xie Fincham, 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), Association for Computational Linguistics, Vienna, Austria 2025, pp. 384–397.
[33] A. Pack, A. Barrett, J. Escalante, Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability, Computers and Education: Artificial Intelligence 6 (2024) 100234. doi:10.1016/j.caeai.2024.100234.