Università per Stranieri di Perugia

Dottorato di Ricerca
in Scienze linguistiche, filologico-letterarie e politico-sociali Indirizzo: Linguistica e didattica delle lingue
Ciclo: XXXVI

Automated Essay Scoring e valutazione umana: un'indagine comparativa

Dottoranda

Talia Sbardella

Relatore

Prof. Roberto Dolci

Correlatore

Prof. Valentino Santucci

A.A. 2022/2023

Indice

Ringraziamenti

Desidero ringraziare di cuore tutti coloro che hanno reso possibile il raggiungimento di questo importante e desiderato traguardo.
Un sentito ringraziamento va al mio relatore, Prof. Roberto Dolci, per la fiducia che mi ha accordato sin dal principio e per avermi guidato con competenza, disponibilità e sensibilità. Il constante stimolo alla riflessione critica e i suoi preziosi consigli sono stati indispensabili, specialmente nelle fasi più difficili di questo percorso.
Vorrei inoltre ringraziare il mio correlatore, Prof. Valentino Santucci, per il sostegno costante, per la grande disponibilità e l'infinita pazienza. I suoi suggerimenti e la sua attenzione hanno rappresentato un contributo decisivo per il perfezionamento dello studio.
Ringrazio il Centro per la Valutazione e le Certificazioni Linguistiche (CVCL), nella persona della Direttrice, Prof.ssa Giovanna Scocozza, per la preziosa quanto indispensabile collaborazione, senza la quale la realizzazione di questa ricerca non avrebbe potuto concretizzarsi.
Un ringraziamento speciale è rivolto alla Prof.ssa Stefania Spina per il fondamentale contributo scientifico offerto alla mia indagine. La sua competenza, la sua disponibilità e la sua generosità nel condividere ricerche, conoscenze ed esperienze sono state per me un punto di riferimento costante. Grazie al materiale da lei messo a disposizione, ho avuto la possibilità di sviluppare la ricerca su basi solide, trovando spunti che hanno arricchito e orientato costantemente il mio studio.
La mia riconoscenza va anche alla Dott.ssa Luciana Forti che, oltre a rappresentare un importante riferimento scientifico, è stata una preziosa presenza umana. La sua capacità di ascolto e il suo sostegno hanno avuto un grande valore nei momenti più impegnativi.
Un ulteriore ringraziamento va alle mie colleghe e ai miei colleghi, in particolare alla Dott.ssa Agnieszka Pakuła e alla Dott.ssa Giorgia Montanucci, con cui ho condiviso momenti fondamentali, arricchiti da riflessioni e scambi stimolanti.

Infine, il ringraziamento più profondo va a mio padre, che continua a guidarmi nel cuore e nel pensiero, alla mia famiglia e a tutti coloro che arricchiscono la mia vita e la completano con la loro presenza.

Introduzione

La diffusione dell'intelligenza artificiale (IA) e dei modelli linguistici generativi, come ChatGPT, sta influenzando in maniera profonda e pervasiva moltissimi aspetti della società contemporanea, trasformando processi, strumenti e pratiche consolidate. Nell'ambito della valutazione della produzione scritta degli apprendenti di lingua seconda o straniera (L2/LS), in particolare, si osserva un interesse sempre più vivo verso l'integrazione di strumenti automatizzati che possano garantire valutazioni rapide, scalabili e potenzialmente più oggettive.

In un contesto in cui le pratiche valutative manuali si confrontano costantemente con criticità quali la soggettività, la variabilità dei criteri applicati e il notevole dispendio di tempo e risorse (Hamp-Lions, 2001; Myers, 2003; Page, 2003), l'adozione di strumenti di *Automated Essay Scoring* (AES) rappresenta una prospettiva innovativa, capace di standardizzare i processi di valutazione e di rendere possibile un'analisi più tempestiva e oggettiva degli elaborati scritti. Tuttavia, l'introduzione di tali strumenti non può non sollevare interrogativi di natura più profonda riguardo alla loro capacità di cogliere la complessità e le dinamiche che contraddistinguono le dimensioni stesse della valutazione linguistica.

Muovendosi all'interno di questo quadro di riflessione, il presente studio si concentra su un'analisi comparativa tra i punteggi assegnati dalla valutazione umana e da quella automatica agli elaborati scritti prodotti da apprendenti di italiano L2/LS in contesti di prove di certificazione linguistica. L'indagine si sviluppa a partire da una selezione delle produzioni scritte raccolte nel Corpus CELI (Spina et al., 2022), che documenta gli elaborati scritti prodotti da apprendenti di italiano come lingua non materna sottoposti alle prove di certificazione CELI, suddivise secondo i livelli B1, B2, C1 e C2 del Quadro Comune Europeo di Riferimento per le Lingue (QCER). La scelta del corpus è motivata dalla sua notevole ricchezza in termini di rappresentazione della varietà e della complessità delle produzioni autentiche degli apprendenti in un contesto certificativo reale.

La riflessione che ha guidato lo studio abbraccia diversi aspetti: da un lato, viene analizzato il quadro teorico alla base della valutazione linguistica, con particolare riferimento ai princìpi alla base della sua evoluzione, distinguendo in maniera critica tra valutazione diretta e indiretta, nonché tra approcci olistici e analitici, questioni che hanno storicamente caratterizzato il campo della valutazione e che sono alla base dell'esigenza di definire standard affidabili ed equi per la valutazione e certificazione delle competenze linguistiche. Dall'altro lato, la ricerca esplora l'evoluzione storica e metodologica degli strumenti AES, illustrando come i modelli linguistici avanzati possano offrire una valutazione automatizzata capace di intercettare aspetti quali la competenza lessicale, grammaticale, la coerenza e la coesione testuale nonché la correttezza e l'adeguatezza sociolinguistica.

Un aspetto centrale dell'indagine riguarda lo studio sperimentale, suddiviso in due fasi complementari. In un primo momento, alcune produzioni rappresentative sono state estratte dal Corpus CELI e sono state sottoposte a ChatGPT mediante un *prompt* appositamente realizzato, che ha guidato il modello nell'assegnazione di un punteggio complessivo, richiedendo una valutazione, in maniera analoga a quella degli esperti umani, relativa a una delle dimensioni valutative previste dalla griglia CELI. Al fine di verificare l'affidabilità del modello e la sua stabilità interna, ogni testo è stato valutato per diversi cicli consecutivi, garantendo così l'indipendenza statistica delle misurazioni e permettendo un'analisi precisa di eventuali oscillazioni nei giudizi. Questa fase ha rappresentato un passaggio preliminare essenziale per sondare la capacità del modello di operare valutazioni coerenti e riproducibili, fungendo da base per il successivo confronto sistematico con la valutazione umana.

La seconda fase della ricerca si concentra sul confronto tra i punteggi assegnati da ChatGPT e quelli risultanti da valutazioni esperte, effettuate su un campione più ampio di 800 testi scritti, sempre selezionati dal Corpus CELI e distribuiti equamente tra i livelli B1, B2, C1 e C2 del QCER. Le valutazioni umane sono state condotte da valutatori qualificati CELI, che hanno applicato la griglia ufficiale di riferimento utilizzata nelle prove di certificazione. L'analisi ha preso in

esame quattro dimensioni della competenza scritta: lessicale, grammaticale, coerenza e coesione testuale, adeguatezza sociolinguistica, in linea con i criteri di valutazione adottati nel Corpus CELI, da cui sono stati selezionati i testi oggetto di studio. L'obiettivo principale è stato quello di rilevare e quantificare le differenze tra i punteggi assegnati dal modello linguistico e quelli attribuiti dai valutatori umani, per poi sviluppare una riflessione critica sulle possibili ragioni delle discrepanze osservate, cercando di esplorare in che misura le valutazioni divergenti possano derivare da differenti modalità di interpretazione, sensibilità o priorità attribuite ai vari aspetti della competenza linguistica.

Infine, lo studio esplora l'idea di una possibile complementarità tra valutazione automatica e umana, ipotizzando un approccio ibrido che integri i punti di forza di entrambi i metodi. I risultati mirano ad offrire spunti per una riflessione critica sulle implicazioni pratiche dell'uso di strumenti AES e per la formulazione di prospettive di ricerca future.

1. La valutazione degli elaborati scritti nella didattica delle lingue

Scrivere in una lingua non materna implica un processo di negoziazione simbolica tra dimensioni cognitive interne e sistemi linguistici esterni, in cui l'errore costituisce una traccia significativa del processo di costruzione di un sistema linguistico personale in evoluzione (Selinker, 1972). Alla luce queste considerazioni, la valutazione della produzione scritta degli apprendenti di una lingua non materna si configura come un momento altamente complesso, in cui il testo prodotto assume il valore di evidenza tangibile di dinamiche cognitive, affettive, linguistiche e culturali spesso non immediatamente osservabili. Gli elaborati scritti, infatti, riflettono le strategie individuali con cui ciascun discente si confronta con la lingua oggetto di apprendimento e con cui modella in modo il proprio percorso formativo. Tale complessità richiede una valutazione articolata, capace di tener conto di diversi fattori, come la coerenza del testo, la padronanza delle convenzioni discorsive della lingua target, nonché l'efficacia delle strategie di compensazione adottate per superare difficoltà espressive (Corder, 1981).

La produzione scritta in lingua non materna, del resto, richiede l'attivazione simultanea di molteplici abilità: dall'organizzazione logica del contenuto alla padronanza delle strutture morfosintattiche, dalla coesione testuale all'adeguatezza pragmatica. Implica pianificazione concettuale, selezione lessicale consapevole, rispetto delle norme discorsive e capacità di adattare il registro al contesto d'uso e alle aspettative culturali dell'interlocutore (Weigle, 2002; Byram, 2000). In quest'ottica, l'elaborato scritto non rappresenta solo un prodotto finale, ma anche uno specchio delle dinamiche profonde dell'apprendimento linguistico, che la valutazione è chiamata a cogliere e interpretare.

Diverse sono state le definizioni che, nel tempo, hanno contribuito a delineare la natura complessa e articolata della valutazione linguistica. Una selezione rappresentativa è stata raccolta da Glenn Fulcher, curatore del sito https://languagetesting.info/, come osserva anche Barni (2023) nel suo contributo sulla valutazione delle competenze nelle lingue seconde. Secondo

Priscilla Allen (2009) "Language Testing is the practice and study of evaluating the proficiency of an individual in using a particular language effectively". Questa definizione, semplice ma efficace, sottolinea due aspetti centrali: da un lato, la natura duplice del language testing come ambito sia pratico sia teorico; dall'altro, la finalità comunicativa della valutazione, che punta a verificare non solo le conoscenze formali, ma la reale capacità dell'apprendente di usare la lingua in modo funzionale.

Una visione più articolata è proposta da Alan Davies (1999), il quale afferma che "The activity of developing and using language tests. As a psychometric activity, language testing traditionally was more concerned with the production, development and analysis of tests. Recent critical and ethical approaches to language testing have placed more emphasis on the uses of language tests. The purpose of a language test is to determine a person's knowledge and/or ability in the language and to discriminate that person's ability from that of others. Such ability may be of different kinds, achievement, proficiency or aptitude. Tests, unlike scales, consist of specified tasks through which language abilities are elicited. The term language assessment is used in free variation with language testing although it is also used somewhat more widely to include for example classroom testing for learning and institutional examinations". L'autore sottolinea come, accanto alla dimensione tecnico-costruttiva della valutazione, si sia progressivamente consolidato un interesse anche per gli scopi e le modalità di impiego dei test linguistici.

Questa prospettiva amplia il focus dal funzionamento interno dello strumento valutativo alla sua applicazione concreta in contesti educativi, istituzionali e sociali. In tale cornice, la valutazione non è più concepita come un semplice atto di misurazione oggettiva, ma come un processo che implica scelte teoriche, pedagogiche ed etiche e che produce effetti tangibili sulla costruzione delle identità linguistiche, sul riconoscimento delle competenze e sull'accesso a opportunità formative o professionali.

Chapelle e Brindley (2010), dal loro canto, pongono l'accento sulla

dimensione processuale e interpretativa della valutazione. Gli autori affermano infatti che "In the context of language teaching and learning, assessment refers to the act of collecting information and making judgments about a language learner's knowledge of a language and ability to use it." L'attenzione è sull'atto valutativo come pratica contestualizzata, integrata nel processo di insegnamento/apprendimento. L'acquisizione di informazioni non è fine a sé stessa, ma funzionale alla formulazione di inferenze e giudizi sulla competenza linguistica dello studente, ovvero sulla sua capacità di usare la lingua in contesti reali, oltre che sulla sua conoscenza formale.

Infine, Fulcher riporta come sua preferita una citazione da Edgeworth (1888), secondo cui la valutazione "[…] is a species of sortition infinitely preferable to the ancient method of casting lots for honours and offices.", sottolineando come l'attività valutativa debba essere accompagnata da riflessioni teoriche, pedagogiche e sociali consapevoli.

Questi punti di vista evidenziano come anche il modo di concepire la valutazione degli elaborati prodotti da apprendenti di lingua non materna si sia modificato nel tempo, seguendo traiettorie teoriche e metodologiche differenziate. I criteri e gli strumenti valutativi hanno infatti attraversato profondi cambiamenti, strettamente legati all'evoluzione dei modelli teorici di competenza linguistica.

In una fase ancora saldamente legata a un'impostazione strutturalista, Robert Lado (1961) fornisce una prima sistematizzazione teorica della valutazione linguistica, ponendo al centro della riflessione l'esigenza di rendere il testing linguistico un'attività oggettiva, replicabile e scientificamente fondata. L'opera *Language Testing: The Construction and Use of Foreign Language Tests* rappresenta uno snodo cruciale nella storia della disciplina, in quanto articola in modo esplicito e sistematico i principi guida per la costruzione di test linguistici standardizzati.

Già in *Linguistics Across Cultures*, Lado (1957) aveva rilevato come, nel processo di apprendimento di una lingua straniera, gli apprendenti tendano ad

acquisire con maggiore facilità quegli elementi linguistici che presentano somiglianze con la propria lingua madre, mentre le strutture che si discostano in modo significativo da essa risultano più complesse e tendono a generare maggiori difficoltà. A tal proposito, affermava: "The individuals tend to transfer the forms and meanings and distribution of forms and meanings of their native language and culture to the foreign language and culture both productively and receptively. In the comparison between native and foreign language, lies the key to ease or difficulty in foreign language learning. We know from the observation of many cases that the grammatical structure of the native language tends to be transferred to the foreign language [...] We have here the major source of difficulty or ease in learning the foreign language [...] those structures that are different will be difficult." (Lado, 1957. p. 2).

A partire da tale osservazione, lo studioso sviluppa la teoria dell'analisi contrastiva (AC), secondo cui è possibile prevedere le principali aree di difficoltà per l'apprendente attraverso un confronto sistematico tra lingua madre (L1) e lingua oggetto di apprendimento. Secondo questo orientamento, i test linguistici devono essere progettati per valutare il grado di superamento di tali difficoltà e vengono pertanto concepiti come strumenti finalizzati a rilevare il *transfer* negativo della L1 nel processo di apprendimento della L2. Considerato da Lado la principale fonte di errore nell'apprendimento linguistico, il *transfer* deve essere identificato e misurato con precisione. I test devono quindi essere costruiti in modo da isolarne gli effetti, attraverso item mirati ai punti di divergenza strutturale tra i due sistemi linguistici, così da verificare i progressi compiuti dall'apprendente nel superamento delle interferenze più critiche.

La modalità di testing privilegiata in questo paradigma è il cosiddetto *discrete-point testing*, che consiste nel suddividere la lingua nelle sue singole componenti (ad esempio strutture grammaticali specifiche, singoli vocaboli, tratti fonologici distinti), da sottoporre a verifica in maniera isolata e standardizzata. Gli strumenti più rappresentativi di questo approccio sono gli item a scelta multipla, nei quali la risposta corretta è unica e predefinita, permettendo così una

valutazione facilmente quantificabile e riproducibile. Le prove sono orientate a presentare gli stimoli linguistici in frasi per lo più decontestualizzate, riducendo al minimo l'influenza di variabili pragmatiche o comunicative. Anche le quattro abilità linguistiche — produzione scritta, produzione orale, comprensione scritta e comprensione orale — vengono trattate in modo compartimentato rispetto alle componenti strutturali della conoscenza linguistica, quali la grammatica, il lessico, la fonologia e la grafematica. In tale contesto, la produzione scritta assumeva prevalentemente la forma di esercizi controllati, con margini di espressione individuale estremamente limitati, bassa variabilità e forte attenzione alla correttezza formale.

Negli stessi anni, Carroll (1961) propose una prospettiva alternativa con il cosiddetto *integrative test*, fondato sull'idea che la lingua debba essere valutata nella sua interezza, tenendo conto dell'interazione tra le sue componenti sistematiche e l'effetto comunicativo complessivo di un enunciato. A differenza dei modelli precedenti, tale impostazione mira a valutare simultaneamente le abilità ricettive e produttive, superando la segmentazione strutturale e favorendo una visione unitaria della performance linguistica, più vicina ai processi reali di comprensione e produzione.

Nei test integrativi viene dunque considerato il linguaggio autentico, ma gli strumenti di valutazione non sono ancora pienamente aderenti alle dinamiche della comunicazione reale, limitandosi a prove che, pur integrando diverse abilità, non sempre garantiscono una valutazione effettiva dell'uso funzionale della lingua all'interno di contesti comunicativi reali.

Un ulteriore sviluppo in questa direzione si deve a Oller (1973), il quale contribuì ad avvicinare il concetto di testing alla realtà dell'uso linguistico spostando l'attenzione sui processi psicolinguistici coinvolti nella produzione e nella comprensione della lingua. L'uso del linguaggio fu interpretato come un'attività complessa che comprende sia l'elaborazione del messaggio in tempo reale — tipica della produzione e della comprensione orale — sia la capacità di attingere alla conoscenza del sistema linguistico per costruire significati coerenti e appropriati al contesto, soprattutto nella produzione scritta. In questa

prospettiva, la competenza linguistica non è solo il risultato della memorizzazione di regole, ma un insieme dinamico di processi cognitivi che entrano in gioco in situazioni comunicative specifiche. Per valutare tale competenza, nella sua complessità, Oller individuò come strumenti particolarmente efficaci il *cloze test* e il dettato, due tipologie di prova che si configurano come tentativi concreti di osservare l'integrazione tra conoscenza linguistica e uso contestuale. Il *cloze test* richiede all'apprendente di completare un testo con parole mancanti, ponendolo di fronte alla necessità di ricostruire il significato globale attraverso l'attivazione simultanea di competenze lessicali, morfosintattiche, pragmatiche e discorsive. Il dettato, invece, si configura come una prova apparentemente semplice ma in realtà complessa, poiché coinvolge capacità di percezione e decodifica del messaggio, memoria a breve termine, competenza ortografica e abilità di comprensione testuale, rivelandosi uno strumento utile per valutare la competenza linguistica in maniera globale e integrata.

A livello didattico, si assiste a un cambiamento paradigmatico: l'interesse si sposta progressivamente dall'insegnamento delle strutture linguistiche in senso stretto alla valorizzazione delle pratiche comunicative. Come evidenzia Vedovelli (2002), ciò comporta il superamento di un approccio centrato sulle forme e strutture per abbracciare una prospettiva orientata all'uso contestualizzato della lingua.

Questa trasformazione investe anche il campo della valutazione linguistica, dove il tradizionale focus sulla conoscenza si sposta competenza linguistica. Si consolida l'idea della lingua come pratica sociale ed emerge l'importanza di includere nella valutazione anche gli aspetti socioculturali, interazionali e pragmatici che ne determinano l'uso efficace e appropriato nei contesti reali (Hymes, 1972). L'atto comunicativo viene quindi riconosciuto come un fenomeno intersoggettivo e situato, regolato da convenzioni e norme d'uso condivise che variano in base all'ambiente, allo scopo, ai partecipanti e al ruolo sociale ricoperto da ciascun interlocutore.

Negli anni Ottanta, i contributi di Michael Canale e Merrill Swain (Canale & Swain, 1980; Canale, 1983) proposero una articolazione della competenza

comunicativa in quattro componenti distinte ma interconnesse: la competenza grammaticale si riferisce alla conoscenza delle regole formali del sistema linguistico (morfologia, sintassi, lessico, ortografia), ossia alla capacità di costruire enunciati corretti dal punto di vista strutturale. La competenza sociolinguistica riguarda invece l'abilità di adeguare la lingua a diversi contesti comunicativi, riconoscendo il modo appropriato di comunicare in una data situazione in funzione di fattori culturali, relazionali e pragmatici.

A queste prime due componenti, Canale e Swain aggiungono la competenza discorsiva, che consiste nella capacità di produrre testi coesi e coerenti, in grado di veicolare significati globali attraverso l'organizzazione efficace delle informazioni. Infine, si inserisce la competenza strategica, che include l'insieme di strategie verbali e non verbali che l'apprendente può mettere in atto per compensare eventuali carenze linguistiche, per riformulare un messaggio non compreso o per mantenere attivo lo scambio comunicativo, rendendo l'interazione efficace anche in condizioni di difficoltà. Anche la produzione scritta viene così riconsiderata come un atto comunicativo integrato, che coinvolge la competenza grammaticale, ma anche quella discorsiva, sociolinguistica e strategica.

Negli anni Novanta, la riflessione teorica sulla valutazione linguistica si arricchisce in modo significativo grazie al contributo di Bachman (1990), che con il modello della *Communicative Language Ability* (CLA) offre una delle più influenti e strutturate teorie per comprendere, descrivere e valutare la competenza comunicativa. In questo modello, la competenza comunicativa è concepita come un insieme dinamico e interrelato di componenti che comprendono, da un lato, la competenza linguistica in senso stretto e, dall'altro, la competenza pragmatica, suddivisa in competenza illocutiva e sociolinguistica. La prima riguarda la capacità di formulare messaggi che realizzino un'intenzione comunicativa (come descrivere, chiedere, persuadere), mentre la seconda si riferisce all'abilità di usare la lingua in modo socialmente appropriato, in relazione alla situazione comunicativa e alle norme culturali di riferimento. Centrale nel modello è anche la competenza strategica, intesa come la capacità dell'apprendente di pianificare,

monitorare e regolare il proprio comportamento linguistico al fine di portare a termine un compito comunicativo, superando eventuali ostacoli nella comprensione o nella produzione.

L'impostazione delineata si rivela particolarmente utile nella valutazione della produzione scritta in L2/LS, dove il testo rappresenta il prodotto osservabile della competenza comunicativa dell'apprendente. Le strategie di pianificazione, revisione, organizzazione del contenuto e adattamento al genere testuale rientrano pienamente nella dimensione strategica della competenza, e devono quindi essere riconosciute e valutate in modo sistematico.

Il modello CLA si è successivamente evoluto in un quadro teorico che integra in maniera esplicita la competenza linguistica con le caratteristiche della situazione comunicativa e con i tratti individuali dell'apprendente. Mentre nella versione del 1990 la CLA si concentrava prevalentemente sulle componenti interne della competenza comunicativa (linguistica, pragmatica, strategica), la riformulazione proposta da Bachman e Palmer (1996) enfatizza le interazioni dinamiche tra l'individuo, il compito linguistico e il contesto d'uso. In questa nuova impostazione, la *performance* non è più considerata una semplice espressione della competenza interna, ma il risultato dell'interazione tra le capacità dell'apprendente, le sue conoscenze tematiche, le disposizioni affettive e le caratteristiche specifiche del compito. Tali componenti interagiscono con la competenza linguistica e strategica e ne modulano l'efficacia nella realizzazione del compito comunicativo. Il modello, dunque, assume una configurazione sistemica e contestualizzata, fornendo una cornice teorica adatta a progettare prove che valutino l'uso autentico della lingua, in particolare nella produzione scritta, dove il testo è l'esito tangibile di una negoziazione complessa tra intenzione comunicativa, genere testuale, scopo, contesto e destinatario. L'enfasi è posta su una concezione integrata dell'uso linguistico, in cui le caratteristiche del compito, la rilevanza dei contenuti attivati e le variabili affettive - come la motivazione, l'ansia o l'attitudine nei confronti del compito - interagiscono con le componenti della competenza linguistica e strategica.

A questi studi si è affiancata l'influenza delle riflessioni proposte da Celce-

Murcia, Dörnyei, Thurrell (1995), secondo cui comunicare non significa solo trasmettere un contenuto, ma anche collocarsi all'interno di un universo culturale e sociale, interagendo con le norme discorsive, le aspettative pragmatiche e le convenzioni testuali della comunità linguistica di riferimento.

I tre autori delineano un modello pedagogico di competenza comunicativa caratterizzato da una struttura multilivello, in cui la competenza discorsiva funge da snodo tra le altre sotto-competenze: la competenza linguistica, relativa alla padronanza delle risorse formali del sistema linguistico; la competenza socioculturale, che riguarda l'adeguatezza del messaggio rispetto a norme sociali e culturali condivise; la competenza strategica, intesa come l'insieme dei meccanismi compensatori e regolatori dell'interazione; la competenza grammaticale, intesa come componente esplicita e strutturale del sapere linguistico. Il modello evidenzia come l'efficacia comunicativa non derivi dalla padronanza isolata di singole abilità linguistiche, bensì dall'interazione dinamica e sinergica tra le diverse componenti della competenza. In tale prospettiva, la valutazione della produzione linguistica – e in particolare quella scritta – deve tener conto della capacità dell'apprendente di costruire un testo culturalmente appropriato, coerente con il genere e la situazione comunicativa.

Tale orientamento trova un punto di sintesi nel Quadro Comune Europeo di Riferimento per le Lingue (QCER), che nella sua prima edizione (Council of Europe, 2001a) introduce un modello centrato sulla performance osservabile e sull'uso funzionale della lingua. Il QCER rappresenta oggi uno strumento fondamentale sai in ambito valutativo, che didattico e curricolare, in quanto fornisce una base comune per la descrizione degli obiettivi, dei contenuti e dei livelli di competenza linguistica in contesti educativi formali e non formali. La sua importanza è duplice: da un lato, consente una trasparenza e comparabilità internazionale delle certificazioni linguistiche, e dall'altro promuove un approccio orientato all'azione, in cui la lingua è considerata come un mezzo per l'agire sociale. In questa prospettiva, la valutazione linguistica si fonda su una visione dell'apprendente come soggetto capace di agire con la lingua all'interno di contesti significativi, concettualizzato come attore sociale.

Come evidenziato da Little (2006), il QCER non deve essere interpretato come un sistema prescrittivo o normativo, bensì come uno "schema descrittivo" aperto, flessibile, volto a sostenere la progettazione curricolare, l'insegnamento e la valutazione secondo criteri di coerenza, trasparenza e adattabilità. Il documento si fonda su una struttura articolata lungo due assi fondamentali: una dimensione verticale e una dimensione orizzontale, concepite in modo complementare e interdipendente. La prima dimensione consiste nella definizione di sei livelli globali di competenza linguistica – A1, A2, B1, B2, C1 e C2 – che delineano una progressione generale dal principiante assoluto al soggetto pienamente competente. I livelli non vanno intesi come categorie rigide, bensì come punti di riferimento all'interno di una struttura in cui è possibile identificare anche livelli potenziati (*plus levels*) in corrispondenza di particolari intervalli di sviluppo: A2+ (livello di sopravvivenza potenziato), che si colloca tra l'A2 e il B1, il B1+ (livello soglia potenziato) posto tra B1 e B2, e B2+ (livello progresso potenziato), che si inserisce tra B2 e C1. Il sistema dei livelli, concepito come "albero flessibile", può inoltre essere adattato localmente attraverso la creazione di sottolivelli ulteriori (ad esempio, A2.1, A2.2), purché venga mantenuto l'ancoraggio al QCER.

Un aspetto fondamentale che caratterizza le scale di livello è la natura centrata sull'apprendente e sull'uso reale della lingua: esse descrivono infatti comportamenti comunicativi osservabili, ovvero ciò che il soggetto è in grado di fare nella lingua target. Tale impostazione rende i descrittori doppiamente accessibili: da un lato, fornisce agli attori istituzionali come progettisti di curricoli, autori di materiali didattici, docenti ed esaminatori, uni strumento condiviso per la programmazione, la didattica e la certificazione, e dall'altro permette agli apprendenti stessi di riconoscersi nei profili descritti, di autovalutare con maggiore consapevolezza il proprio livello di competenza e di orientarsi in modo autonomo e riflessivo nel percorso di apprendimento.

A completamento di questa architettura descrittiva si affianca la dimensione orizzontale del QCER, che amplia l'analisi della competenza linguistica prendendo in considerazione le condizioni, i contesti e le modalità d'uso della

lingua. Tale struttura è modulata attraverso le attività di produzione orale e scritta, ricezione orale e scritta, interazione orale e scritta e mediazione, le quali riflettono la varietà delle pratiche comunicative che caratterizzano l'agire linguistico nei diversi domini della vita sociale. La produzione scritta, in particolare, è descritta come una delle attività linguistiche fondamentali. Il suo sviluppo è rappresentato attraverso una progressione articolata che delinea le capacità dell'utente di pianificare, strutturare e realizzare testi scritti adeguati allo scopo comunicativo, al destinatario e al contesto. I descrittori relativi alla produzione scritta evidenziano la capacità dell'apprendente di esprimersi con coerenza, pertinenza e controllo stilistico crescente, passando da produzioni semplici e prevedibili a testi articolati, ricchi di contenuti e connotati da padronanza retorica e linguistica. In questo senso, la valutazione della scrittura diventa un'osservazione di performance autentiche, in cui la lingua è utilizzata in modo dinamico, strategico e contestualizzato, riflettendo l'interazione tra finalità comunicativa e risorse linguistiche disponibili. Insieme alla progressione verticale, la dimensione orizzontale del *Quadro* definisce quindi un sistema descrittivo integrato, che restituisce la complessità intrinseca del comunicare in lingua straniera inteso come insieme di pratiche sociali, finalizzate e contestualizzate. La visione dell'apprendimento linguistico è ancorata all'azione e alla pluralità dei contesti, in linea con un approccio centrato sul soggetto comunicante come agente dotato di intenzionalità e capacità di adattamento.

Sulla base di questa concezione dell'apprendente come protagonista attivo del proprio percorso formativo, si colloca l'introduzione del *Portfolio Europeo delle Lingue* (PEL) (Council of Europe, 2001b), con l'intento di tradurre in pratica i principi fondamentali del QCER – tra cui la trasparenza, l'autovalutazione, l'autonomia e la valorizzazione dei repertori plurilingui. Il PEL si configura come uno strumento pedagogico e metacognitivo volto a documentare i progressi linguistici dell'individuo, a stimolare la riflessione critica sull'apprendimento e a riconoscere formalmente le competenze acquisite in contesti formali, non formali e informali. Nel contesto europeo, tale strumento assume anche una funzione sociale e politica, in quanto consente agli apprendenti di autodichiarare il proprio

livello di competenza linguistica in maniera trasparente e strutturata, fornendo visibilità alle competenze linguistiche possedute e contribuendo allo stesso tempo alla costruzione di un profilo linguistico riconoscibile e interpretabile all'interno dello spazio europeo dell'istruzione e del lavoro. La funzione rappresentativa del PEL si inserisce coerentemente nella strategia del Consiglio d'Europa volta alla promozione del plurilinguismo come risorsa fondamentale per una cittadinanza attiva, inclusiva e interculturale. L'intento è dunque quello di promuovere un'educazione linguistica orientata all'autoconsapevolezza, alla responsabilità individuale e alla mobilità educativa e professionale. In questa prospettiva, l'atto di documentare, riflettere e valutare le proprie competenze linguistiche costituisce un processo formativo coerente con i paradigmi dell'autovalutazione formativa e dell'apprendimento permanente (Serragiotto, 2016).

In continuità con l'approccio descrittivo e orientato all'azione promosso dal QCER, emerge ancora il *Profilo della lingua italiana*, pubblicato da Spinelli e Parizzi (2010), il cui obiettivo principale è da un lato, adattare i descrittori del QCER alla realtà strutturale, lessicale e pragmatica della lingua italiana; dall'altro, fornire una sistematizzazione dettagliata ed empiricamente fondata delle competenze comunicative attese per ciascun livello di padronanza. La struttura generale si fonda sul principio che la competenza linguistica non debba essere intesa in termini esclusivamente astratti o normativi, ma vada descritta in riferimento a comportamenti comunicativi osservabili e situati, coerentemente con la visione del QCER. In riferimento alla scrittura, il *Profilo* adotta una prospettiva funzionale che mira a rendere espliciti i legami tra risorse linguistiche, generi testuali e scopi comunicativi. L'articolazione proposta comprende non solo l'individuazione dei generi tipici per ciascun livello (e-mail, brevi narrazioni, lettere personali, testi argomentativi), ma anche un repertorio articolato di funzioni pragmatiche (informare, descrivere, esprimere opinioni, richiedere, giustificare), che costituiscono l'orizzonte d'uso delle strutture linguistiche. Particolarmente rilevante è la presenza di inventari linguistici graduati, che includono repertori morfosintattici specifici per ogni livello, strutture sintattiche ricorrenti e compatibili con la complessità dei compiti richiesti, meccanismi di coesione

e coerenza testuale e lessici selezionati e suddivisi per categorie semantiche, in base alla frequenza d'uso e alla funzionalità comunicativa. L'impostazione mira a descrivere quindi modo dinamico ciò che un apprendente sa effettivamente produrre e comprendere per iscritto, rendendo trasparente il nesso tra abilità comunicative e risorse linguistiche disponibili. L'apprendente è così concepito come soggetto attivo, in grado di mobilitare consapevolmente le risorse linguistiche necessarie per realizzare intenzioni comunicative, costruire testi coerenti e interagire con interlocutori reali o immaginati.

Da un punto di vista didattico, il *Profilo* assume un ruolo altamente strategico. Gli inventari graduati offrono un riferimento empiricamente fondato per la progettazione curricolare, la scelta di *input* linguistici adeguati e la costruzione di rubriche valutative coerenti con i livelli del QCER. Inoltre, rispetto all'impostazione generale del Quadro, la formulazione proposta da Spinelli e Parizzi si distingue per un ulteriore livello di granularità descrittiva, finalizzato a precisare in modo articolato le competenze linguistiche attese nei livelli iniziali dell'apprendimento dell'italiano L2/LS. Mentre il QCER propone descrittori di carattere prevalentemente globale, volti a rappresentare la competenza comunicativa in termini sintetici e funzionali, il *Profilo* adotta un modello analitico, che scompone la competenza in sottocomponenti linguistiche specifiche, collegate alle strutture morfosintattiche, lessicali e testuali proprie della lingua italiana. Questa articolazione consente una più chiara definizione della corrispondenza tra i livelli del *Quadro* e le risorse linguistiche effettivamente attivate dall'apprendente, offrendo allo stesso tempo un riferimento solido per la valutazione, in quanto fornisce una base strutturata per l'osservazione sistematica e l'analisi delle competenze acquisite.

Tornando al QCER, è opportuno soffermarsi su alcuni aspetti teorici e metodologici che hanno suscitato nel tempo osservazioni e riflessioni critiche nel panorama della didattica delle lingue. Sebbene questo strumento abbia contribuito in modo decisivo alla sistematizzazione dei livelli di competenza e alla diffusione di un metalinguaggio condiviso in ambito europeo, non sono mancate critiche alla sua impostazione. Diversi autori, infatti, hanno sollevato interrogativi

sulla sua solidità empirica. Secondo North (2000), molti descrittori sono stati elaborati a partire da intuizioni di esperti più che da dati osservabili, mentre Hulstijn (2007) ha evidenziato come i livelli di competenza non siano stati derivati da analisi sistematiche delle produzioni autentiche degli apprendenti, ma da giudizi soggettivi basati sull'esperienza professionale di insegnanti e valutatori. Tali considerazioni hanno sollevato dubbi sulla capacità del *Quadro* di rappresentare in modo accurato e generalizzabile lo sviluppo reale della competenza in L2, anche per quanto riguarda la produzione scritta. Allo stesso tempo, le osservazioni hanno messo in luce la necessità di rafforzare la dimensione empirica della valutazione linguistica e di promuovere approcci fondati sull'analisi diretta delle produzioni reali degli apprendenti.

In questa direzione si è affermato un orientamento metodologico un fondato sull'uso dei corpora linguistici autentici per analizzare le produzioni scritte e orali in modo sistematico e documentato. Un esempio interessante di applicazione di tale approccio è l'indagine di Casani (2020), basata sull'analisi di 400 testi scritti da apprendenti adulti di italiano L2/LS selezionati dal Corpus MERLIN[1]. L'obiettivo dello studio è quello di contribuire a una validazione empirica dei livelli del QCER, con un focus particolare sulla descrizione della competenza morfosintattica. Gli studi hanno fornito dati empirici robusti che hanno evidenziato chiaramente come gli errori degli apprendenti non siano distribuiti in maniera casuale ma seguano precise traiettorie evolutive che si collegano ai livelli del QCER. L'impiego di strumenti analitici avanzati, come la Computer-aided Error Analysis (CEA), introdotta da Granger (1998) e approfondita da Dagneaux, Denness e Granger (1998), ha reso possibile una analisi dettagliata degli errori linguistici prodotti in contesti autentici. Lo studio ha permesso di identificare tipologie specifiche di errori comuni a determinati livelli di competenza e di comprendere più a fondo le dinamiche interne dell'interlingua, rivelando

[1] Il Corpus MERLIN, di natura trilingue (italiano, tedesco e ceco), è composto da 2286 testi scritti da apprendenti adulti di lingua seconda o straniera. I testi, raccolti prevalentemente in contesti di certificazione TELC (The European Language Certificates), sono stati annotati morfosintatticamente e rivalutati da esperti sulla base delle griglie del QCER. Il Corpus è liberamente consultabile online all'indirizzo: https://www.merlin-platform.eu/.

meccanismi sistematici quali omissioni, sovraestensioni e tipologie di errori grammaticali frequenti. I risultati confermano l'importanza di fondare la valutazione linguistica su dati osservabili, rilevati in contesti d'uso reali, anziché su intuizioni teoriche o giudizi esperti isolati.

In linea con le prospettive orientate all'uso, si inserisce l'approccio *task-based* proposto da Norris (2009), che amplia l'orizzonte valutativo spostando il focus dalla competenza formale alla capacità di usare la lingua per agire in contesti autentici o semi-autentici. La valutazione della produzione scritta si concentra dunque sulla capacità dell'apprendente di portare a termine un compito significativo (ad esempio, scrivere una e-mail formale, una lettera di reclamo, un articolo), prendendo in esame non solo la correttezza linguistica, ma anche l'efficacia comunicativa, l'organizzazione testuale, la coerenza e l'aderenza al genere discorsivo richiesto.

Il modello *scenario-based* (Turner & Purpura, 2016) propone un ulteriore approccio alla valutazione partendo dal presupposto che la performance dell'apprendente non possa essere interpretata in maniera esaustiva se astratta dal contesto in cui si realizza. La valutazione, pertanto, si articola attraverso scenari simulati, costruiti per riflettere situazioni comunicative verosimili e intrinsecamente legate a specifici riferimenti culturali. All'interno di questi contesti, l'apprendente è chiamato a fronteggiare richieste complesse, che lo inducono ad attivare simultaneamente competenze linguistiche, processi cognitivi e risorse interazionali, in vista della risoluzione di compiti dotati di coerenza interna e pertinenza funzionale. Le prove progettate secondo questo modello assumono la forma di eventi comunicativi autentici, nei quali la lingua non viene utilizzata solo in riferimento alla correttezza morfosintattica o lessicale, ma come strumento strategico per la negoziazione del significato, la gestione del discorso e l'adeguamento alle variabili contestuali e culturali dell'interazione (Purpura, 2016). Purpura fornisce una definizione chiara di valutazione linguistica: "Language assessment is a broad term referring to a systematic procedure for eliciting test and non-test data (e.g., a teacher checklist of student performance) for the purpose of making inferences or claims about certain

language-related characteristics of an individual." (Purpura, 2016, p. 303). In quest'ottica, il modello *scenario-based* si configura come una concreta applicazione di una modalità valutativa volta a integrare dati osservabili, compiti autentici e processi inferenziali complessi.

Nel 2020, il *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*, in italiano *Quadro comune europeo di riferimento per le lingue: apprendimento, insegnamento, valutazione – Volume complementare* (QCER-VC) (Council of Europe, 2020) arricchisce e aggiorna il QCER, ampliando in modo significativo la prospettiva proposta nel documento del 2001. Tra le principali innovazioni introdotte, emergono l'aggiunta del livello di competenza pre-A1, la valorizzazione della mediazione linguistica, dell'interazione online e della competenza plurilingue e pluriculturale, intesa come repertorio integrato, e dinamico. Il quadro teorico e descrittivo sulla produzione presentato nel testo evidenzia con chiarezza la complessità della produzione linguistica, collocandola non solo all'interno di una prospettiva funzionale — legata all'atto comunicativo — ma anche come espressione di competenze formali, cognitive e socio-pragmatiche.

La produzione, nelle sue varie manifestazioni, è considerata come un processo articolato che implica pianificazione, strutturazione, selezione, nonché controllo e adattamento in tempo reale. Lo spostamento dal prodotto al processo, parzialmente anticipato nel QCER, trova nel QCER-VC uno sviluppo più ampio, utile sia per la didattica che per la valutazione: la competenza di esprimersi in modo coerente e articolato, soprattutto nei contesti formali e accademici, viene descritta come il frutto di un apprendimento consapevole, strutturato e continuo, che implica anche la familiarizzazione con i generi testuali e le loro convenzioni. Le strategie di produzione, come "pianificazione", "compensazione" e "controllo e riparazione", agiscono come regolatori metacognitivi che permettono all'apprendente di mantenere il flusso comunicativo anche in presenza di ostacoli linguistici o cognitivi, e sono fondamentali per esprimersi in contesti autentici (QCER- VC, pp. 64–65). Osservando la figura 1, la quale illustra le attività e

strategie di produzione linguistiche, risulta evidente una chiara distinzione tra due assi fondamentali: da un lato, le attività di produzione, suddivise in produzione orale e scritta, dall'altro le strategie di produzione, che agiscono trasversalmente a entrambe le modalità. Soffermandosi sulla sezione relativa alla produzione scritta, è possibile notare tre sottosezioni: produzione scritta generale, scrittura creativa e relazioni e saggi. L'articolazione rivela l'attenzione verso la molteplicità dei generi testuali e delle funzioni della scrittura, che vanno dalla narrazione personale alla produzione accademica e professionale.

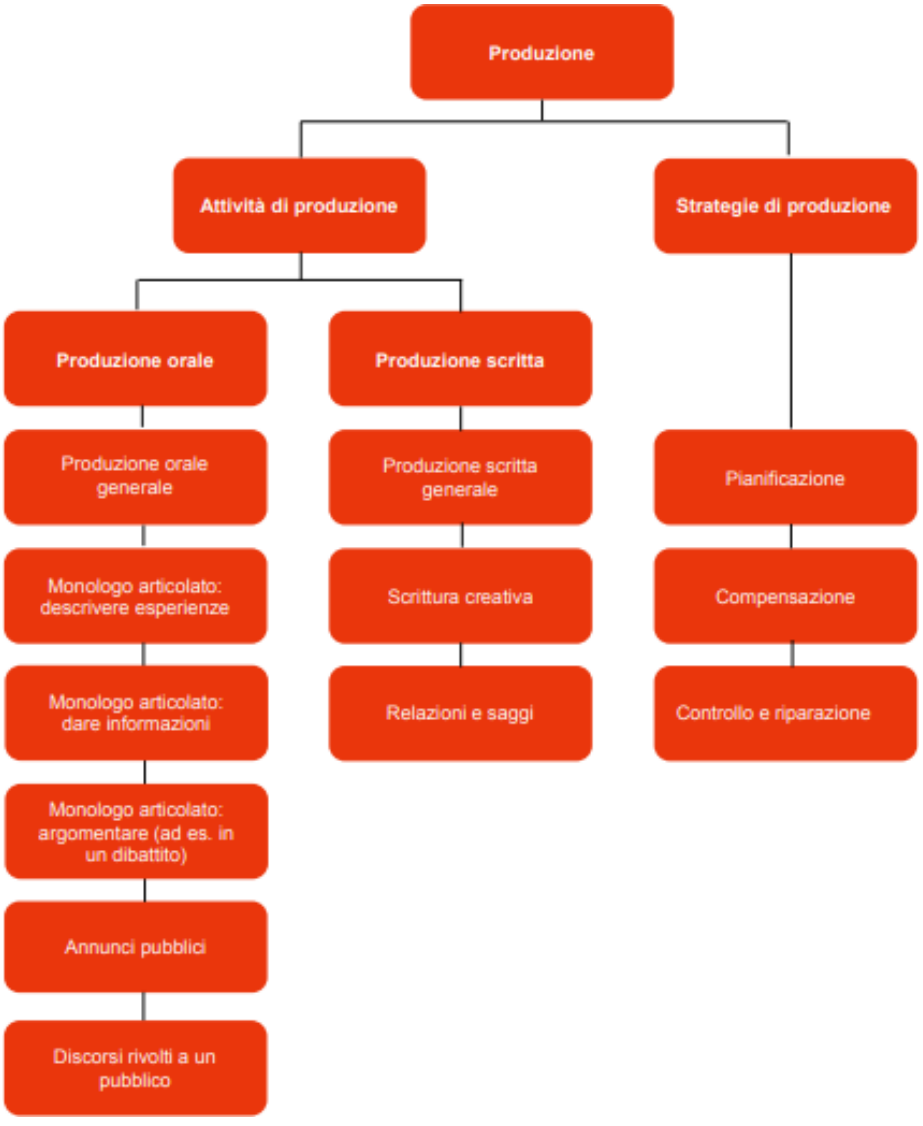


Figura 1. Attività e strategie di produzione (QCER-VC 2020, p. 65).

In questo quadro, la valutazione della produzione scritta si delinea come un processo sempre più sfaccettato, che richiede un'attenta considerazione dei costrutti sottesi all'atto di scrivere, nonché delle modalità con cui questi possono essere osservati e interpretati in chiave formativa. L'analisi dell'elaborato scritto, infatti, non può prescindere da una lettura che tenga conto non solo del prodotto finale, ma anche delle scelte linguistiche e strategiche che ne hanno guidato la realizzazione. Ciò implica una maggiore attenzione alla qualità e alla trasparenza degli strumenti valutativi, così come alla loro capacità di sostenere processi di apprendimento consapevoli e orientati allo sviluppo progressivo delle competenze.

Nelle sezioni che seguono, queste riflessioni troveranno ulteriore approfondimento attraverso l'analisi di alcune dimensioni fondamentali della competenza scritta, lette anche alla luce delle indicazioni offerte dal QCER-VC.

1.1 La dimensione lessicale

La competenza lessicale in L2 rappresenta oggi una delle aree più rilevanti e discusse nell'ambito della linguistica educativa e della glottodidattica, sia per la sua centralità nel processo di acquisizione della lingua, sia per le implicazioni metodologiche e valutative che essa comporta.

Come evidenziato da Jafrancesco e La Grassa (2021), la riflessione sulla dimensione lessicale è uscita progressivamente da una posizione ancillare rispetto alla grammatica, assumendo un ruolo di primo piano nei percorsi di insegnamento/apprendimento dell'italiano L2. Tale prospettiva deriva dalla constatazione, ormai largamente condivisa nella letteratura scientifica, che la costruzione della competenza comunicativa passa primariamente attraverso il lessico, e che l'apprendente L2 si affida inizialmente proprio a un repertorio di parole chiave per avviare lo sviluppo della propria interlingua (Selinker, 1972; Vedovelli e Casini, 2016).

Villarini (2021) riprende questo orientamento evidenziando come, nella fase iniziale dell'apprendimento, alcune parole ad alta frequenza e funzionalità assumano il ruolo di "supernodi" lessicali, da cui si dirama l'espansione della rete lessicale dell'apprendente, con evidenti implicazioni per la didattica e per la valutazione.

Nel passaggio dai metodi grammaticali-traduttivi a quelli comunicativi, si è assistito a una ridefinizione del ruolo del lessico, che non è più concepito come semplice lista di unità atomiche da apprendere in modo meccanico, ma come insieme dinamico di forme e significati, di collocazioni, espressioni idiomatiche e *chunk*.

L'approccio lessicale, teorizzato da Lewis (1993), propone una concezione della lingua come "lessico grammaticalizzato", in cui la grammatica non rappresenta un insieme di regole preordinate, ma emerge progressivamente dall'uso lessicale e dalle regolarità che lo caratterizzano. In tale quadro teorico, la competenza lessicale si configura come una componente complessa della competenza comunicativa, articolata nella capacità di riconoscere, selezionare e utilizzare in modo appropriato unità poli-lessicali, combinazioni frequenti, collocazioni e forme pragmaticamente marcate. Essa presuppone un processo di interiorizzazione fondato su un'esposizione ampia, autentica e significativa all'*input* linguistico, sostenuta da un intervento didattico intenzionale, guidato e riflessivo (Lo Cascio, 2007; Maggini, 2021).

Questa complessità si riflette anche sul piano valutativo: secondo Gallina (2022), valutare la competenza lessicale significa osservare e descrivere la capacità dell'apprendente di mobilitare in modo flessibile e pertinente il proprio repertorio lessicale all'interno di compiti comunicativi autentici, in cui entrano in gioco fattori cognitivi, testuali e pragmatici.

L'autrice propone dunque un modello valutativo orientato all'osservazione in situazione, che si fonda sull'analisi della produzione linguistica in compiti d'uso realistici, e che integra in modo sistematico criteri quantitativi e qualitativi. Un elemento di particolare rilievo nel modello di Gallina è la distinzione tra ciò che l'apprendente potenzialmente conosce — il repertorio latente — e ciò che

effettivamente utilizza in un dato compito. Tale distinzione, che richiama le nozioni di lessico passivo e attivo, evidenzia la necessità di costruire strumenti di valutazione capaci di sollecitare in modo autentico e differenziato l'attivazione del repertorio lessicale. Per questo motivo, Gallina propone l'impiego di compiti comunicativi autentici, quali narrazioni, descrizioni di immagini, lettere, e-mail, testi regolativi e argomentativi, che permettano di osservare il comportamento lessicale in relazione a scopi specifici e generi testuali determinati. Inoltre, l'autrice suggerisce una prospettiva olistica e processuale sulla valutazione, che non si limiti al prodotto finale, ma che tenga conto anche delle strategie messe in atto durante la pianificazione e revisione del testo.

L'analisi testuale diventa così lo strumento privilegiato per valutare la competenza lessicale, poiché consente di rilevare come l'apprendente struttura il proprio discorso, quali parole sceglie, come le combina, e in che misura è in grado di variare, elaborare o negoziare il significato. L'osservazione si focalizza non solo sull'*output* lessicale, ma anche sull'efficacia comunicativa e sulla coerenza globale dell'enunciato.

Infine, Gallina ribadisce la necessità che la valutazione del lessico sia coerente con gli obiettivi formativi e calibrata sui livelli di competenza attesi, in linea con i descrittori del QCER VC, il quale, per quanto riguarda la competenza lessicale, offre una trattazione più articolata rispetto al documento del 2001. Le griglie descrittive, già presenti nel QCER, si articolano in due sottodimensioni: l'ampiezza del lessico e la padronanza del lessico. La prima fa riferimento alla quantità e varietà di vocaboli che l'apprendente è in grado di utilizzare, mentre la seconda si concentra sulla qualità dell'uso, in termini di proprietà semantica, pertinenza contestuale, flessibilità stilistica e consapevolezza pragmatica.

I descrittori relativi all'ampiezza del lessico, proposti dal QCER-VC e rappresentati nella figura 2 nella pagina seguente, tracciano una progressione graduale e coerente che riflette lo sviluppo della competenza lessicale lungo l'intero *continuum* dei livelli di padronanza linguistica. Tale progressione si estende da una conoscenza minima, limitata a parole di alta frequenza e a espressioni di uso quotidiano, caratteristica dei livelli iniziali (A1-A2), fino al

possesso di un repertorio ampio, vario e articolato, capace di coprire con precisione concetti specialistici, astratti e culturalmente connotati, come richiesto nei livelli più avanzati (C1-C2). L'insieme dei descrittori mette dunque in evidenza una concezione del lessico non solo come quantità di parole note, ma come risorsa dinamica che si amplia, si raffina e si diversifica parallelamente all'aumento della competenza comunicativa complessiva.

Ampiezza del lessico	
C2	Ha buona padronanza di un repertorio lessicale vastissimo che comprende espressioni idiomatiche e colloquiali; dà prova di essere consapevole dei livelli di connotazione semantica.
C1	Ha buona padronanza di un vasto repertorio lessicale che permette di superare prontamente le lacune usando circonlocuzioni; la ricerca di espressioni e le strategie di evitamento sono poco evidenti. È in grado di scegliere tra più possibilità lessicali in quasi tutte le situazioni, utilizzando dei sinonimi anche per parole/segni non comuni. Ha una buona padronanza di espressioni idiomatiche e colloquiali; è in grado di fare dei giochi di parole/segni con facilità. È in grado di comprendere e utilizzare in modo appropriato il lessico tecnico e le espressioni idiomatiche proprie del suo campo di specializzazione.
B2	È in grado di comprendere e di utilizzare la terminologia tecnica generale del campo di specializzazione, quando ne discute con altri specialisti. Dispone di un buon repertorio lessicale relativo al suo settore e a molti argomenti generali. È in grado di variare le formulazioni per evitare un eccesso di ripetizioni; lacune lessicali possono ancora provocare esitazioni e richiedere circonlocuzioni. Nella maggior parte dei contesti è in grado di inserire in modo abbastanza sistematico le parole/i segni appropriati. È in grado di comprendere e utilizzare una gran parte del lessico relativo al suo campo specialistico ma ha delle difficoltà con la terminologia specialistica di altri settori.
B1	Ha un buon repertorio di lessico relativo ad argomenti familiari e situazioni quotidiane. Dispone di lessico sufficiente per esprimersi con qualche circonlocuzione su quasi tutti gli argomenti che si riferiscono alla vita di tutti i giorni, quali la famiglia, gli hobby e gli interessi, il lavoro, i viaggi e l'attualità.
A2	Dispone di lessico sufficiente per sostenere transazioni della *routine* quotidiana in situazioni e su argomenti familiari. Dispone di lessico sufficiente per esprimere bisogni comunicativi di base. Dispone di lessico sufficiente per far fronte a bisogni semplici di sopravvivenza.
A1	Dispone di un repertorio lessicale di base fatto di singole parole/segni ed espressioni riferibili a un certo numero di situazioni concrete.
Pre-A1	*Nessun descrittore*

Figura 2 – Scala dell'ampiezza lessicale (QCER-VC 2020, p. 142).

La seconda scala relativa alla padronanza del lessico, non presenta descrittori per i livelli pre-A1 e A1 e introduce un livello ulteriore di osservazione qualitativa. Essa valuta, per esempio, la capacità dell'apprendente di evitare ambiguità semantiche, di controllare la polisemia, di scegliere termini tecnicamente appropriati, di distinguere registri e toni, e di adottare strategie compensatorie in caso di lacune.

In particolare, a partire dal livello B2, si osserva come l'efficacia comunicativa dipenda progressivamente meno dall'estensione quantitativa del repertorio e sempre più dalla capacità di combinare le risorse lessicali disponibili, di riformularle, di collocarle con pertinenza e di adattarle ai vincoli e agli scopi dell'interazione. La figura 3 riassume e visualizza questa progressione, evidenziando il passaggio da una competenza prevalentemente quantitativa a una qualitativa e funzionale.

La valutazione della padronanza lessicale, dunque, richiede strumenti di analisi sensibili al contesto e alla funzione, evidenziando la necessità di interpretare il lessico dentro il discorso e non al di fuori di esso.

Padronanza del lessico	
C2	Uso del lessico costantemente corretto e adeguato.
C1	Usa un lessico meno comune in modo idiomatico e appropriato. Occasionali sbagli di minore entità, ma nessun errore lessicale significativo.
B2	La correttezza lessicale è generalmente elevata, anche se si può presentare qualche confusione e qualche scelta lessicale scorretta, ma non pregiudizievole per la comunicazione.
B1	Mostra una buona padronanza del lessico elementare, ma continuano a verificarsi errori gravi quando esprime pensieri più complessi o affronta argomenti e situazioni non familiari. Usa in modo appropriato un ampio repertorio lessicale di base quando parla di argomenti familiari.
A2	Dispone di un repertorio ristretto, funzionale ad esprimere bisogni concreti della vita quotidiana.
A1	*Nessun descrittore.*
Pre-A1	*Nessun descrittore.*

Figura 3. Scala di padronanza del lessico (QCER-VC 2020: 142).

1.2 La dimensione grammaticale

Il passaggio da approcci strutturalisti e grammaticali-traduttivi a modelli comunicativi e orientati all'azione ha modificato profondamente la concezione della grammatica: da sistema normativo da memorizzare, essa è oggi intesa come un insieme di risorse a supporto dell'interazione comunicativa (Coonan et al., 2018; Balboni, 2015).

Nel contesto della didattica dell'italiano L2, la grammatica assume una funzione duplice: da un lato, costituisce il fondamento formale della costruzione dell'enunciato, e dall'altro, contribuisce alla coerenza testuale e all'efficacia pragmatica della comunicazione. Secondo Vedovelli (2011), l'interlingua dell'apprendente si sviluppa anche attraverso una ristrutturazione progressiva delle conoscenze grammaticali, nella quale l'errore viene interpretato non come una deviazione da correggere, ma come un indicatore significativo del processo di acquisizione in corso.

Il QCER-VC include una riflessione dettagliata sulla correttezza grammaticale, che indica che la "[…] la capacità di chi usa/apprende la lingua sia di ricordare correttamente espressioni "prefabbricate", sia di concentrarsi sulle forme grammaticali mentre sta articolando il suo pensiero. Ciò non è semplice perché, quando si formulano dei pensieri o si realizzano dei compiti ancora più impegnativi, chi usa/apprende la lingua deve dedicare la maggior parte delle sue capacità di elaborazione mentale all'esecuzione del compito." (QCER-VC, 2020: 143). Tale capacità è fortemente influenzata dalla complessità del compito comunicativo: quando l'attenzione dell'apprendente è assorbita dall'esecuzione di attività cognitive impegnative, il livello di correttezza grammaticale tende a diminuire. Questo fenomeno si riflette nei descrittori ufficiali, rappresentati nella figura 4, che non seguono una progressione lineare, ma tengono conto di tali fluttuazioni. I concetti chiave operativi adottati nella scala del sono: il controllo di un repertorio grammaticale specifico (dal livello A1 al B1), la predominanza degli errori (dal livello B1 al B2) e il grado di controllo esercitato sulle strutture grammaticali (dal livello B2 al C2).

Correttezza grammaticale	
C2	Mantiene costantemente il controllo grammaticale di forme linguistiche complesse, anche quando la sua attenzione è rivolta altrove (ad es. nella pianificazione di quanto intende dire e nell'osservazione delle reazioni altrui)
C1	Mantiene costantemente un livello elevato di correttezza grammaticale; gli errori sono rari e poco evidenti.
B2	Ha una buona padronanza grammaticale; nella struttura delle frasi possono ancora verificarsi sbagli occasionali, errori non sistematici e difetti minori, che sono però rari e vengono per lo più corretti a posteriori.
	Mostra una padronanza grammaticale piuttosto buona. Non fa errori che possano provocare fraintendimenti. Ha un buon controllo delle strutture utilizzate in una lingua semplice e di alcune forme grammaticali complesse, anche se tende a utilizzare le strutture complesse in modo rigido con qualche inesattezza.
B1	Comunica con ragionevole correttezza in contesti familiari; la padronanza grammaticale è generalmente buona anche se si nota l'influenza della lingua madre. Nonostante gli errori, ciò che cerca di esprimere è chiaro.
	Usa in modo ragionevolmente corretto un repertorio di formule di *routine* e strutture d'uso frequente, relative alle situazioni più prevedibili.
A2	Usa correttamente alcune strutture semplici, ma continua sistematicamente a fare errori di base – ad es. tende a confondere i tempi verbali e a dimenticare di segnalare gli accordi; ciononostante ciò che cerca di dire è solitamente chiaro.
A1	Ha solo una padronanza limitata di qualche semplice struttura grammaticale e di semplici modelli sintattici, in un repertorio memorizzato.
Pre-A1	È in grado di utilizzare principi molto semplici che regolano l'ordine delle parole/dei segni in frasi brevi.

Figura 4 – Scala di correttezza grammaticale (QCER-VC 2020: 144).

In questo contesto, la riflessione metalinguistica assume un ruolo crescente, in linea con quanto evidenziato da Ellis (2008), secondo cui la consapevolezza linguistica – ovvero la capacità di riflettere sulla lingua e di monitorare la propria produzione – favorisca lo sviluppo della competenza grammaticale e contribuisca all'autonomia dell'apprendente. A livello metodologico, la ricerca ha evidenziato l'efficacia dell'insegnamento implicito e dell'apprendimento grammaticalmente sensibile attraverso compiti comunicativi, come evidenziato da Long (2015) e da Ellis (2009). In tale approccio, le strutture grammaticali vengono acquisite in modo incidentale, ma consolidate mediante attività di riflessione e feedback formativo, privilegiando l'uso funzionale e contestualmente adeguato della grammatica. In ambito valutativo, si evidenzia quindi la necessità di considerare la grammatica non come un dominio isolato, ma come componente integrata della performance linguistica complessiva.

1.3 La dimensione sociolinguistica

Il QCER-VC attribuisce un ruolo centrale alla competenza sociolinguistica**,** definendola come la capacità di gestire consapevolmente la dimensione sociale dell'uso linguistico. Essa si manifesta attraverso l'adattamento del proprio comportamento comunicativo in base al contesto situazionale, agli interlocutori e alle convenzioni culturali e linguistiche della comunità di riferimento. Tale competenza viene descritta e articolata in modo sistematico nella scala di adeguatezza sociolinguistica (figura 5), illustrando con chiarezza la progressione dei livelli di competenza.

Appropriatezza sociolinguistica	
C2	È in grado di mediare efficacemente tra i parlanti della lingua di destinazione e della propria comunità tenendo conto delle differenze socioculturali e sociolinguistiche. Ha buona padronanza di espressioni idiomatiche e colloquiali ed è consapevole dei livelli di connotazione semantica. Coglie pienamente le implicazioni sociolinguistiche e socioculturali della lingua usata da parlanti competenti e reagisce in modo adeguato. È in grado di utilizzare efficacemente, sia oralmente che per iscritto, un'ampia e accurata varietà di lingua per comandare, discutere, persuadere, dissuadere, negoziare e consigliare.
C1	È in grado di riconoscere un'ampia gamma di espressioni idiomatiche e colloquiali e coglie i cambiamenti di registro; può però a volte aver bisogno che venga confermato qualche particolare, soprattutto se non ha familiarità con l'accento. È in grado di comprendere l'umorismo, l'ironia e impliciti riferimenti culturali e di cogliere sfumature di significato. È in grado di comprendere film in cui si fa ampio uso di espressioni gergali e idiomatiche. È in grado di usare la lingua per scopi sociali in modo flessibile ed efficace, includendo anche le dimensioni affettive, allusive e umoristiche. È in grado di regolare il suo livello di formalità (registro e stile) per adattarsi in modo appropriato al contesto sociale formale, informale o colloquiale e mantenere un registro orale coerente. È in grado di cogliere osservazioni critiche o di esprimere con tatto un profondo disaccordo.
B2	È in grado, con qualche sforzo, di intervenire, in una discussione prendendovi parte, anche se gli interlocutori parlano velocemente e in modo colloquiale. È in grado di identificare e interpretare dei codici socioculturali e sociolinguistici e di modificare consapevolmente il suo modo di esprimersi affinché risulti adeguato alla situazione. È in grado di esprimersi in modo sicuro, chiaro e cortese in registro formale o informale, a seconda della situazione e della persona implicata (delle persone implicate).
	È in grado di adattare la sua espressione per distinguere tra registri formali e informali, ma non sempre lo fa in modo appropriato. È in grado di interagire con parlanti la lingua di arrivo senza rendersi involontariamente ridicolo/a o irritarli o metterli nella necessità di comportarsi in modo diverso da come farebbero con un interlocutore competente. È in grado di esprimersi in modo adeguato alla situazione ed evita errori grossolani di formulazione.
B1	È in grado di realizzare un'ampia gamma di atti linguistici e di rispondervi usando le espressioni più comuni in registro neutro. È consapevole delle più importanti regole di cortesia e le rispetta. È consapevole delle più significative differenze esistenti tra usi e costumi, atteggiamenti, valori e credenze prevalenti della comunità in questione e la propria e ne ricerca i segnali.
A2	È in grado di realizzare atti linguistici di base, quali richieste e scambi di informazioni, e di rispondervi e di esprimere in modo semplice opinioni e atteggiamenti. È in grado di socializzare in modo semplice ma efficace, usando le espressioni comuni più semplici e attenendosi alle convenzioni di base.
	È in grado di gestire scambi comunicativi molto brevi, usando formule convenzionali correnti per salutare e rivolgere la parola a qualcuno. È in grado di fare inviti, dare suggerimenti, chiedere scusa e rispondere a mosse analoghe ecc.
A1	È in grado di stabilire contatti sociali di base usando le più semplici formule convenzionali correnti per salutare e congedarsi, presentare qualcuno, dire "per favore", "grazie", "scusi" ecc.
Pre-A1	*Nessun descrittore.*

Figura 5 – Scala di appropriatezza sociolinguistica (QCER-VC 2020: 148).

Ai livelli iniziali, in particolare al livello A1, l'apprendente mostra un uso esclusivamente elementare della lingua in contesti sociali, limitandosi a formule convenzionali basilari, spesso apprese come sequenze fisse, applicate nei saluti o nelle interazioni più prevedibili.

Il livello A2 introduce un ampliamento di questo repertorio minimo: l'utente è in grado di partecipare a scambi brevi e routinari, seguendo le convenzioni più comuni, come formulare inviti o esprimere ringraziamenti e scuse, ma senza ancora capacità autonoma di scelta o adattamento del registro in funzione della situazione.

Il livello B1 segna una soglia intermedia in cui l'apprendente acquisisce maggiore consapevolezza delle norme sociali e dei comportamenti verbali appropriati in differenti situazioni quotidiane. È in grado di utilizzare una gamma più ampia di espressioni linguistiche in contesti generici, pur mantenendo tendenzialmente un registro neutro. Inoltre, riconosce le differenze culturali più evidenti e dimostra rispetto per gli usi e le convenzioni della comunità linguistica, senza però riuscire ancora ad adattarsi pienamente alle varianti stilistiche richieste da contesti più marcati.

A livello B2, questa competenza si consolida: l'utente riesce a distinguere tra registri formali e informali, anche se con occasionali incertezze nell'uso. È in grado di prendere parte a interazioni complesse, intervenendo in modo efficace anche in situazioni spontanee e meno prevedibili, ed è capace di scegliere espressioni linguistiche in funzione del ruolo degli interlocutori e delle caratteristiche situazionali. Inoltre, sa modificare il proprio stile comunicativo con un certo grado di controllo, dimostrando una progressiva abilità nell'adeguarsi a contesti sociali diversificati.

Il livello C1 rappresenta un grado avanzato di padronanza. L'utente sa usare la lingua con flessibilità in situazioni sociali varie, riconosce e impiega correttamente espressioni connotate culturalmente e riesce a modulare il proprio comportamento linguistico in base a criteri di formalità, stile e ruolo relazionale. È in grado di cogliere con precisione elementi tipici della comunicazione

informale, di variare registro con disinvoltura e di gestire l'interazione in maniera coerente con i codici sociali della comunità linguistica, anche in presenza di variabilità interna, come nel caso degli accenti regionali.

Infine, il livello C2 si caratterizza per la piena padronanza dell'adeguatezza sociolinguistica. L'utente è capace di operare scelte linguistiche raffinate e sempre appropriate, sia nella comunicazione scritta che orale, in un'ampia gamma di contesti, anche non familiari. È in grado di cogliere e utilizzare con precisione forme idiomatiche, locuzioni colloquiali e livelli di connotazione stilistica, e possiede la competenza necessaria per mediare tra interlocutori di background linguistici e culturali diversi, dimostrando una profonda comprensione delle dinamiche sociolinguistiche implicate nell'interazione.

1.4 La dimensione della coerenza e coesione testuale

Nel QCER VC, la dimensione della coerenza e della coesione testuale viene definita come il risultato dell'interrelazione tra i diversi elementi linguistici che concorrono alla costruzione del significato complessivo di un testo. Tale interrelazione si realizza attraverso l'uso appropriato di connettivi, dispositivi coesivi, meccanismi referenziali, ellissi, sostituzioni, marcatori testuali e strutture discorsive, i quali garantiscono la continuità tematica e la progressione informativa del discorso. Sia la coesione sia la coerenza si articolano, pertanto, su due livelli simultanei: da un lato, quello della frase/enunciato, in cui si manifesta la connessione sintattico-semantica tra le unità linguistiche; dall'altro, quello della struttura complessiva del testo, dove entrano in gioco l'organizzazione globale delle informazioni, la linearità argomentativa e la coerenza logico-concettuale dell'intero messaggio.

I descrittori della scala riportati nella pagina successiva, nella figura 6, si basano, pertanto, su tre elementi chiave: la capacità di collegare parole o segmenti mediante connettivi, l'organizzazione del testo attraverso la segmentazione in paragrafi, e il controllo e la varietà nell'uso degli strumenti linguistici impiegati per strutturare il discorso.

Coerenza e coesione	
C2	È in grado di realizzare un discorso coerente e coeso, usando in modo appropriato una grande varietà di schemi organizzativi e un'ampia gamma di connettivi e di meccanismi coesivi di altro tipo.
C1	È in grado di realizzare un discorso chiaro, sciolto e ben strutturato, mostrando un uso controllato degli schemi organizzativi, dei connettivi e delle espressioni coesive. È in grado di produrre un testo ben strutturato e coerente, utilizzando una varietà di mezzi di coesione e di schemi orqanizzativi.
B2	È in grado di usare in modo efficace diversi connettivi per esplicitare i rapporti tra i concetti. È in grado di usare un numero limitato di elementi di coesione per collegare i propri enunciati in un discorso chiaro e coerente. In un intervento lungo possono presentarsi dei "salti" logici. È in grado di produrre testi generalmente ben organizzati e coerenti, utilizzando una varietà di parole di colleqamento e dispositivi di coesione. È in grado di strutturare testi più lunghi in paragrafi chiari e logici.
B1	È in grado di introdurre una controargomentazione in un testo discorsivo semplice (ad es. con *ma*, *però*). È in grado di collegare una serie di elementi relativamente brevi e semplici in una sequenza lineare per punti. È in grado di formare frasi più lunghe e collegarle tra loro, utilizzando un numero limitato di dispositivi coesivi, ad es. in una storia. È in grado di creare logiche e semplici interruzioni di paragrafo in un testo molto lungo.
A2	È in grado di collegare frasi semplici usando i connettivi più usuali per raccontare una storia o descrivere qualcosa, realizzando un semplice elenco di punti. È in grado di collegare gruppi di parole con connettivi semplici (ad es. *e*, *ma*, *perché*).
Coerenza e coesione	
A1	È in grado di collegare parole o gruppi di parole con connettivi molto elementari (per es. "e" o "allora").
Pre-A1	*Nessun descrittore.*

Figura 6 – Scala relativa alla coerenza e coesione testuale (QCER-VC 2020: 153).

Al livello A1, l'utente è in grado di collegare parole o gruppi di parole solo tramite connettivi molto elementari**,** come "e" o "allora", dimostrando una competenza ancora incipiente nella strutturazione testuale.

Il livello A2 segna un primo passo in avanti, in quanto l'apprendente riesce a produrre sequenze più estese, utilizzando i connettivi più frequenti (come "e", "ma", "perché") per collegare frasi semplici e dare una forma basilare al racconto o alla descrizione. Tuttavia, la costruzione testuale rimane ancora frammentaria, con strutture poco articolate e priva di una vera organizzazione paratestuale.

Nel passaggio al livello B1 si rileva una maggiore capacità di articolazione: l'utente è in grado di formare frasi più lunghe e di collegarle tramite un numero limitato di dispositivi coesivi, anche se l'organizzazione rimane ancora prevalentemente lineare. È in grado di introdurre semplici controargomentazioni e di suddividere testi più lunghi in paragrafi chiari, aspetto che riflette un primo livello di controllo sulla macrostruttura testuale.

A livello B2 si osserva un ulteriore affinamento: l'apprendente utilizza una varietà di connettivi in modo più efficace, riesce a esplicitare chiaramente i rapporti logici tra i concetti e produce testi ben organizzati e coesi. Pur limitato nell'ampiezza del repertorio linguistico, l'uso degli strumenti coesivi risulta più sistematico e funzionale alla costruzione di una sequenza testuale coerente.

Nei livelli avanzati, C1 e C2, si raggiunge un'elevata padronanza della coerenza e coesione. Al livello C1 l'utente è in grado di realizzare testi strutturati e coesi, mostrando controllo sull'impiego di connettivi, espressioni coesive e schemi organizzativi complessi. Il testo risulta chiaro, ben articolato e capace di guidare il lettore o interlocutore attraverso le diverse fasi del discorso. A livello C2, infine, si osserva la competenza più alta: l'apprendente riesce a gestire un'ampia gamma di dispositivi di coesione e a strutturare un testo con coerenza piena, facendo ricorso in modo appropriato e diversificato a schemi organizzativi e a una vasta gamma di connettivi e meccanismi coesivi. Tale livello implica anche una notevole flessibilità nell'adattare la struttura testuale in base agli scopi comunicativi, garantendo la continuità logica del discorso e la chiarezza nella trasmissione dei contenuti.

1.5 I princìpi della valutazione linguistica

Come sottolineato da Bachman (1990), il punto di partenza imprescindibile di ogni operazione valutativa consiste nella definizione del costrutto, ovvero dell'insieme delle abilità e conoscenze che si intendono osservare e misurare attraverso il compito linguistico. La chiarezza nella definizione del costrutto è condizione necessaria per garantire la coerenza fra obiettivi di apprendimento, progettazione delle prove e criteri di valutazione.

Il principio di validità, nell'accezione proposta da Messick (1989), assume una portata epistemologica ampia e sistemica. Lo studioso definisce la validità come un costrutto integrato che abbraccia diverse componenti: la pertinenza dei contenuti, la coerenza rispetto al modello teorico di riferimento, la relazione con criteri esterni, l'adeguatezza dell'interpretazione dei risultati e l'impatto delle

decisioni che ne conseguono. La validità, dunque, non risiede nello strumento valutativo in sé, ma nella qualità delle inferenze che si è legittimati a trarre dai dati osservati e nel modo in cui tali inferenze sono utilizzate in ambito educativo e sociale.

In particolare, Messick evidenzia che la validità di una valutazione dipende dalla giustificazione teorica delle inferenze che si traggono dai dati, dalla solidità empirica delle prove su cui tali inferenze si basano e dalla consapevolezza delle implicazioni sociali e educative del loro utilizzo. La valutazione, dunque, è valida non solo se misura il costrutto previsto, ma anche se le sue conseguenze sono coerenti con i principi di equità e responsabilità educativa. Questo approccio multidimensionale alla validità ha fortemente influenzato il campo del Language testing and assesment, orientando la progettazione delle prove verso una maggiore attenzione all'impatto etico e formativo delle pratiche valutative.

A ciò si lega il principio di affidabilità, inteso come la capacità di uno strumento valutativo di produrre risultati consistenti, indipendentemente da variabili contestuali, temporali o soggettive. Questo aspetto si rivela fondamentale nella valutazione della produzione scritta, dove la componente qualitativa dell'interpretazione del testo da parte del valutatore introduce un potenziale margine di variabilità. Come rilevano Alderson, Clapham e Wall (1995), la scrittura presenta una delle sfide più significative in termini di affidabilità, soprattutto per la difficoltà di standardizzare il processo valutativo senza ridurne la ricchezza interpretativa.

Secondo Bachman e Palmer (1996), un sistema valutativo efficace deve essere valido, affidabile, equo e orientato all'apprendimento. Questi aspetti sono particolarmente rilevanti per la produzione scritta, che comporta inevitabilmente una componente soggettiva nella valutazione e una molteplicità di variabili da considerare: stile, organizzazione, accuratezza linguistica, ma anche creatività e padronanza discorsiva. Per rispondere alla necessità di garantire maggiore trasparenza, affidabilità e validità nella misurazione delle competenze, la ricerca in ambito glottodidattico ha favorito la diffusione di strumenti strutturati, tra cui le griglie analitiche. Queste ultime permettono di articolare il giudizio su più

dimensioni autonome, ciascuna delle quali riflette un aspetto specifico della competenza scritta, come la padronanza morfosintattica, l'organizzazione testuale, la pertinenza lessicale, la coerenza pragmatica e l'aderenza al genere richiesto. Tali strumenti, affermatisi a partire dagli anni Novanta grazie ai contributi di studiosi come Bachman, Palmer e Weigle (2002), si sono rivelati particolarmente efficaci anche per il loro impiego nei principali test standardizzati internazionali, come TOEFL e IELTS. Oltre a garantire maggiore oggettività e replicabilità, le griglie analitiche si fondano su una logica formativa: attraverso l'assegnazione di punteggi specifici e l'esplicitazione dei criteri di valutazione, esse offrono all'apprendente un feedback articolato, utile alla consapevolezza metalinguistica e al miglioramento delle proprie abilità. Per il docente, invece, possono rappresentare un dispositivo operativo funzionale alla diagnosi e alla programmazione didattica. In una prospettiva autenticamente educativa, la griglia analitica si configura dunque come uno strumento di mediazione tra osservazione valutativa e azione formativa. Quando progettata sulla base di dati autentici, validata empiricamente e allineata ai descrittori del QCER, essa consente di restituire un'immagine attendibile dello sviluppo della competenza scritta in L2/LS, rispondendo efficacemente tanto agli obiettivi certificativi quanto a quelli formativi (Weigle, 2002; Scarino & Liddicoat, 2009).

Hamp-Lyons (1991) sottolinea che l'impiego di rubriche analitiche rappresenta una strategia efficace per guidare l'attenzione dei valutatori verso aspetti specifici della performance scritta, riducendo la tendenza a formulare giudizi globali poco controllabili. L'autrice evidenzia l'importanza della trasparenza e della condivisione dei criteri per favorire la convergenza valutativa tra diversi esaminatori. Lumley (2005), evidenzia come la formazione dei valutatori - unita all'uso di scale descrittive calibrate - migliori in modo significativo l'affidabilità *inter-rater*, cioè il grado di concordanza tra due o più valutatori indipendenti che giudicano la stessa performance linguistica. L'autore sottolinea come questo fenomeno si verifichi in particolare nei cosiddetti *high-stakes contexts*, ossia quelle situazioni valutative in cui l'esito di una prova ha un impatto rilevante sul percorso formativo, accademico o professionale dell'apprendente. Tali contesti

possono includere, ad esempio, esami di certificazione linguistica, prove di ammissione a programmi universitario selezioni per l'accesso a opportunità lavorative, dove la coerenza dei punteggi ha conseguenze rilevanti.

Entrambi gli autori, pur riconoscendo la complessità del compito valutativo, pongono l'accento sul ruolo delle rubriche come strumento di mediazione tra il testo scritto e il giudizio esperto, contribuendo a strutturare il processo interpretativo e a limitare la variabilità soggettiva.

Inoltre, l'uso combinato di rubriche e sessioni di formazione per i valutatori – con attività di *benchmarking*, discussione di casi e definizione dei criteri – si è rivelato essenziale per rafforzare l'affidabilità intersoggettiva. Fulcher e Davidson (2007) sostengono che l'affidabilità non deve essere vista come un obiettivo isolato, ma come parte di una più ampia "validità in uso", che considera l'interazione tra strumenti, contesti e pratiche professionali. In quest'ottica, garantire un'elevata affidabilità non significa soltanto ridurre la variabilità dei punteggi, ma anche assicurare che tali punteggi rappresentino fedelmente e in modo giustificabile la performance dell'apprendente in rapporto al costrutto valutato.

Un ulteriore asse centrale dell'impianto valutativo è il principio di equità, che implica l'eliminazione di *bias* potenziali derivanti da variabili sociolinguistiche e culturali. Come osserva Fulcher (2010), la valutazione linguistica è inserita in un contesto sociale, istituzionale e politico, e produce effetti concreti sugli apprendenti. La costruzione di strumenti valutativi deve dunque mirare a garantire pari opportunità a tutti i candidati, evitando che differenze di *background* o percorsi formativi penalizzino l'espressione autentica delle competenze linguistiche.

Lo studioso evidenzia che un test equo è uno strumento che riconosce e accoglie la diversità, ponendo l'accento sulla pertinenza dei compiti, sull'accessibilità delle istruzioni e sulla formazione critica dei valutatori. L'equità, dunque, si realizza attraverso una progettazione attenta delle prove, una riflessione etica sulle loro conseguenze e una prassi valutativa che consideri la lingua non come un'entità astratta, ma come un mezzo situato e relazionale.

Tali riflessioni si ricongiungono alla visione critica di McNamara (1996, 2000), che ha messo in discussione l'idea del *testing* come pratica neutrale e oggettiva. Secondo l'autore, i test linguistici non misurano solo competenze formali, ma agiscono come pratiche performative che producono effetti identitari, sociali e politici. Ogni valutazione, infatti, riflette scelte culturali su cosa debba essere considerato "corretto", "appropriato" o "valido". In tal senso, lo studioso propone il concetto di giustizia interpretativa (*interpretive justice*), secondo cui la validazione di una prova deve tenere conto non solo della coerenza interna e delle metriche psicometriche, ma anche delle sue conseguenze sociali e del modo in cui contribuisce a includere o escludere gli individui dal riconoscimento delle competenze.

Una posizione critica sul Language Testing è espressa da Shohamy (1997; 2001a; 2001b), che riflette su come i test e i loro risultati possano essere impiegati come strumenti di potere, contribuendo al rafforzamento e alla conservazione di egemonie culturali e linguistiche. L'autrice evidenzia come l'adozione di modelli monolingui, standardizzati e formalmente "neutrali" nasconda in realtà la tendenza a privilegiare una concezione di lingua idealizzata, normata, spesso modellata sulla varietà dominante. Gli apprendenti plurilingui, i parlanti di varietà non standard o coloro che adottano strategie comunicative ibride possono risultare così svantaggiati, non per mancanza di competenza, ma per disallineamento rispetto a un modello imposto.

Le prove standardizzate, secondo Shohamy, possono diventare strumenti di controllo ideologico e culturale, soprattutto nei contesti scolastici e istituzionali. Ad esempio, l'uso dei test per determinare l'accesso a percorsi formativi, certificazioni o cittadinanza linguistica può tradursi in una forma di esclusione sistemica mascherata da oggettività tecnica. Il problema non risiede solo nella struttura del test, ma nell'intero discorso di legittimazione che accompagna la valutazione: l'idea che vi sia un modo "giusto" di scrivere o parlare e che questo modo corrisponda a una norma linguistica invisibilmente legata a contesti di prestigio sociale.

L'autrice elabora, dunque, una concezione alternativa di valutazione

linguistica, fondata su principi di democrazia, inclusività e dialogicità. Una valutazione in grado di riconoscere e valorizzare la diversità linguistica non come deficit da correggere, ma come risorsa comunicativa e culturale. Secondo questa prospettiva, i test dovrebbero essere costruiti in modo da riflettere la pluralità dei repertori linguistici, accogliendo la variabilità come elemento costitutivo della competenza, non come deviazione dalla norma. Ciò implica anche ripensare le modalità di feedback, il ruolo dei valutatori, la costruzione dei compiti e la trasparenza dei criteri, affinché l'apprendente possa essere parte attiva del processo valutativo. Adottare un approccio critico significa, secondo la studiosa, interrogarsi continuamente sulle proprie scelte valutative, sulla costruzione degli strumenti, sull'uso dei punteggi, ma anche sul tipo di soggettività che la valutazione promuove o inibisce.

Infine, la dimensione formativa della valutazione occupa un ruolo strategico all'interno di un approccio didattico orientato allo sviluppo delle competenze comunicative. Come sottolineano Black e Wiliam (1998), la valutazione non dovrebbe assumere una funzione esclusivamente sommativa o certificativa, ma andrebbe integrata in modo dinamico nel processo di insegnamento-apprendimento. Particolare enfasi viene posta sul ruolo del feedback, che, se tempestivo, specifico e ancorato a criteri trasparenti, può incidere in modo significativo sull'autoregolazione dell'apprendente, sulla consapevolezza dei propri processi cognitivi e sullo sviluppo di strategie metalinguistiche efficaci. Secondo tale prospettiva, dunque, il testo scritto non rappresenta un prodotto conclusivo, ma un artefatto intermedio suscettibile di approfondimento e la valutazione formativa si configura come uno strumento epistemologicamente fondato e pedagogicamente efficace (Serragiotto 2016, p. 36), in grado di promuovere l'inclusione, la motivazione intrinseca e la responsabilizzazione attiva degli apprendenti.

1.5.1 Valutazione diretta e indiretta

Nel contesto della valutazione linguistica, una distinzione fondamentale è quella tra valutazione diretta, due modalità che riflettono approcci differenti alla misurazione della competenza linguistica e che comportano scelte teoriche, didattiche e tecniche ben distinte. Questa distinzione risulta particolarmente significativa nella valutazione della produzione scritta, dove la natura del compito valutativo condiziona in modo rilevante sia le performance dell'apprendente, sia le pratiche di insegnamento che ne derivano.

La valutazione diretta implica l'osservazione e l'analisi di un testo scritto effettivamente prodotto dall'apprendente in risposta a una consegna comunicativa autentica o semi-autentica. Questo approccio si fonda sulla performance reale e contestualizzata dell'utente linguistico, e rappresenta la modalità di riferimento nei principali sistemi di certificazione delle competenze linguistiche, compresi gli esami CELI. La valutazione diretta consente una misurazione più fedele della competenza comunicativa scritta, in quanto tiene conto della capacità dell'apprendente di organizzare, sviluppare e articolare significati in maniera autonoma e coerente.

La valutazione indiretta, invece, si basa su strumenti che mirano a inferire la competenza di scrittura attraverso compiti che non prevedono la produzione testuale in senso stretto, ma piuttosto il controllo di conoscenze linguistiche correlate, come la grammatica, il lessico, la punteggiatura o la coesione testuale. Esempi tipici includono esercizi di completamento, trasformazione o correzione di frasi. Sebbene la valutazione indiretta presenti vantaggi in termini di standardizzazione, oggettività e rapidità, essa non è in grado di cogliere appieno la complessità della produzione scritta, poiché esclude la componente compositiva, strategica e pragmatica dell'atto di scrivere. Ne consegue che, nei contesti in cui l'obiettivo è valutare l'effettiva competenza comunicativa scritta, la valutazione diretta è generalmente considerata più valida e significativa, pur comportando maggiori oneri in termini di tempo, formazione e risorse. Tale

distinzione è ampiamente documentata nella letteratura sulla valutazione linguistica (Bachman & Palmer, 1996; Weir, 2005), dove si sottolinea come la valutazione diretta, pur meno standardizzabile, offra un indice più autentico delle competenze comunicative dell'apprendente, in linea con l'orientamento del QCER.

L'approccio riflette le condizioni d'uso autentiche della lingua scritta, in cui l'apprendente è chiamato a mobilitare risorse linguistiche, retoriche e cognitive per costruire un testo coeso e contestualizzato (Cumming, 2001). Tuttavia, la valutazione diretta presenta alcune criticità legate alla difficoltà di garantire un'elevata affidabilità tra valutatori, a causa dell'inevitabile componente interpretativa implicita nel processo di giudizio. La valutazione indiretta, invece, si fonda sull'analisi di sottocompetenze linguistiche attraverso strumenti strutturati come *cloze test*, trasformazioni, esercizi grammaticali, i quali mirano a inferire la competenza scritta dell'apprendente senza ricorrere alla produzione testuale diretta. Sebbene tale approccio favorisca standardizzazione e oggettività, esso presenta limiti intrinseci in termini di validità costruttiva, poiché non consente di osservare la capacità dell'apprendente di gestire l'intero processo compositivo, né di misurare le abilità integrate richieste dalla scrittura autentica (Bachman & Palmer, 1996b).

Il *Manual for Language Test Development and Examining* (Consiglio d'Europa, 2011), redatto in collaborazione con l'Association of Language Testers in Europe (ALTE), sottolinea come la valutazione diretta risulti preferibile nei contesti in cui l'obiettivo sia quello di accertare la competenza comunicativa complessiva, in particolare in relazione a generi testuali specifici e a contesti accademici o professionali. La valutazione indiretta può tuttavia costituire un valido complemento diagnostico, soprattutto nelle fasi preliminari del percorso formativo o all'interno di batterie di test su larga scala, dove prevalgono esigenze di efficienza e replicabilità.

A fronte dei limiti e dei punti di forza di ciascun approccio, si assiste oggi a una crescente adozione di modelli ibridi che integrano elementi di entrambe le modalità valutative. Tali modelli cercano di bilanciare le esigenze di validità, affidabilità e sostenibilità, e risultano coerenti con un'impostazione flessibile e

funzionale della valutazione linguistica. Come osservano Taylor e Galaczi (2011), l'efficacia del sistema valutativo risiede nella capacità di articolare in modo complementare approcci diversi, evitando rigidità dicotomiche e valorizzando la varietà di strumenti a disposizione dei docenti e degli enti certificatori.

1.5.2 Approcci valutativi: metodo olistico e metodo analitico

Una seconda distinzione di rilievo, all'interno delle procedure di valutazione, riguarda gli approcci valutativi adottati: in particolare, il metodo olistico e il metodo analitico. Tali approcci, oltre a riflettere concezioni differenti della competenza linguistica, influenzano direttamente le modalità di correzione, i criteri adottati per l'attribuzione dei punteggi e il tipo di feedback offerto agli apprendenti, incidendo così sulle pratiche didattiche quotidiane e sulla progettazione delle attività valutative in contesto educativo. Questi due approcci riflettono concezioni differenti della competenza linguistica e implicano scelte teoriche, didattiche e operative ben distinte, soprattutto nella valutazione della produzione scritta e orale in contesto L2/LS.

Il metodo olistico prevede l'attribuzione di un punteggio globale all'elaborato sulla base di un giudizio complessivo del valutatore, fondato su impressioni generali e su descrittori sintetici dei livelli di competenza. Questo approccio, spesso impiegato nelle prime fasi di selezione o in contesti in cui è richiesta rapidità, consente una valutazione rapida ed economicamente sostenibile. Tuttavia, la sua efficacia dipende in larga misura dall'esperienza del valutatore e dalla coerenza interna al sistema di correzione, esponendosi al rischio di soggettività e ridotta trasparenza nei criteri adottati.

Il metodo analitico, al contrario, si basa sulla scomposizione della performance scritta in dimensioni specifiche – tipicamente correttezza grammaticale, ampiezza e adeguatezza lessicale, coerenza e coesione testuale, struttura del discorso, aderenza alla consegna – ciascuna delle quali viene valutata separatamente secondo una scala predefinita. Questo approccio, adottato in molti contesti certificativi è particolarmente apprezzato per la sua

trasparenza, replicabilità e valore formativo, poiché consente all'apprendente di ricevere un feedback articolato e mirato. Al tempo stesso, però, richiede un elevato grado di preparazione da parte dei valutatori, un'attenta progettazione delle rubriche e tempi più lunghi per la correzione. In prospettiva glottodidattica, la scelta tra metodo olistico e analitico dovrebbe dipendere dalla finalità della valutazione, dal contesto di applicazione e dalle risorse disponibili, tenendo sempre presente l'equilibrio tra validità, affidabilità e sostenibilità del processo valutativo.

Hamp-Lyons (1991), Cumming (2001) e Hyland (2003) evidenziano come la scelta metodologica influenzi non solo il giudizio, ma anche la natura del feedback e l'intero processo di insegnamento/apprendimento.

1.6 Affidabilità e validità delle procedure di valutazione umana

Garantire alti livelli di validità e affidabilità non è un compito semplice. La valutazione della scrittura implica un certo grado di soggettività, legata alla percezione individuale del valutatore, alla sua esperienza, alla formazione ricevuta e persino a fattori contestuali e cognitivi (Barkaoui, 2007). Anche in presenza di griglie analitiche dettagliate, è possibile osservare variazioni nei punteggi assegnati a uno stesso testo. Per tale motivo, nei principali contesti certificativi – come nel caso degli esami CELI, gestiti dal CVCL dell'Università per Stranieri di Perugia – le istituzioni adottano protocolli rigorosi, che comprendono sessioni di formazione e aggiornamento dei valutatori, l'uso sistematico di rubriche valutative condivise, la doppia correzione incrociata degli elaborati e il monitoraggio statistico dei risultati, in linea, in riferimento al contesti CELI, con gli standard di qualità promossi dall'ALTE, di cui il CVCL è membro.

Nonostante queste misure, permane una tensione strutturale tra validità e affidabilità. Da un lato, un approccio altamente strutturato e rubricato può aumentare la coerenza delle valutazioni, ma rischia di ridurre la sensibilità al contenuto e alla qualità argomentativa dei testi. Dall'altro, una valutazione più flessibile e interpretativa può essere più aderente alla complessità del testo

scritto, ma meno replicabile. Questo equilibrio dinamico rappresenta una delle principali sfide teoriche e operative della valutazione linguistica (Weir, 2005).

1.7 Sfide e limiti della valutazione tradizionale

Nonostante i numerosi sforzi per affinare le procedure di valutazione umana della scrittura, la valutazione tradizionale presenta una serie di limiti strutturali che ne riducono l'efficacia e la sostenibilità, soprattutto in contesti educativi su larga scala. In primo luogo, la correzione di testi scritti richiede tempi lunghi e un alto impegno cognitivo da parte dei valutatori, soprattutto quando si adotta un approccio analitico. La necessità di garantire qualità e coerenza valutativa implica un investimento continuo nella formazione del personale e nella gestione dei processi correttivi.

Un ulteriore limite è rappresentato dalla variabilità dei giudizi, che può compromettere l'equità dell'intero sistema valutativo. Studi empirici hanno documentato come anche valutatori esperti possano divergere significativamente nella valutazione di uno stesso testo (Lumley, 2002), soprattutto in presenza di testi liminari, collocabili tra due livelli del QCER. Tale incertezza può influenzare negativamente non solo l'affidabilità, ma anche la percezione di giustizia da parte degli apprendenti.

Vi sono poi questioni di scalabilità: nei contesti in cui è necessario correggere grandi volumi di testi – ad esempio nelle prove standardizzate o nei test di certificazione – il ricorso esclusivo alla valutazione umana può non essere sostenibile, soprattutto se si vogliono mantenere elevati standard di qualità. Inoltre, in presenza di un elevato numero di prove, la valutazione tradizionale può incontrare difficoltà nel fornire feedback personalizzati, elemento considerato sempre più importante nei processi di apprendimento linguistico.

Infine, uno degli aspetti più delicati e problematici della valutazione tradizionale riguarda la presenza di *bias* impliciti, ovvero influenze cognitive e percettive non consapevoli che possono alterare l'imparzialità del giudizio. Tali fenomeni possono emergere in diversi modi e a più livelli, anche in assenza di

intenzionalità discriminatoria (Chen & Hanning, 1985). Anche in contesti formalizzati e standardizzati, come quelli della certificazione linguistica, questi elementi possono influenzare il giudizio, spesso in modo sottile e difficilmente rilevabile. Il fenomeno chiamato *halo effect*, cioè la tendenza a lasciare che un'impressione iniziale positiva o negativa influenzi il giudizio complessivo, così come l'effetto contrasto, per cui il giudizio su un testo può variare a seconda della qualità del testo precedente, sono ben documentati nei processi valutativi (Eckes, 2012).

Per mitigare tali effetti, le istituzioni preposte alla valutazione adottano misure come la formazione specifica dei valutatori, la revisione incrociata degli elaborati, l'anonimizzazione dei testi e l'uso di rubriche standardizzate. Tuttavia, la complessità della valutazione della scrittura, la quale implica giudizi su aspetti linguistici, retorici, pragmatici e culturali, rende quasi impossibile l'eliminazione totale della componente soggettiva.

In questo scenario, si inserisce con crescente interesse l'ipotesi di integrare i processi valutativi con strumenti basati sulll'IA, i quali, se correttamente progettati, possono contribuire a ridurre alcuni *bias* di tipo cognitivo. I sistemi automatizzati, infatti, valutano sulla base di parametri predefiniti e coerenti, senza essere influenzati da elementi visivi, culturali o affettivi. Tuttavia, anche l'uso dell'IA comporta rischi non trascurabili, che richiedono un'attenta riflessione critica. I sistemi basati sull'apprendimento automatico apprendono infatti schemi e regolarità dai dati su cui vengono addestrati; qualora tali dati riflettano disuguaglianze o pregiudizi impliciti del contesto sociale e culturale di origine, sussiste il rischio che queste distorsioni vengano replicate o addirittura amplificate nel processo valutativo (Blodgett et al., 2020).

In questa cornice dai contorni sempre più articolati, si rende necessaria una riflessione approfondita sull'integrazione delle tecnologie nella valutazione linguistica, affinché essa possa avvenire in modo consapevole, etico e pedagogicamente fondato.

2. Automated Essay Scoring

L'approccio basato sull'AES ha progressivamente acquisito rilievo come una delle innovazioni tecnologiche più significative nella valutazione delle produzioni scritte (Shermis & Barrera, 2002; Shermis & Burstein, 2003; Shermis, Raymat, & Barrera, 2003). Esso si basa sull'impiego su modelli algoritmici che consentono di assegnare un punteggio a testi scritti, simulando – entro certi limiti – i processi decisionali tipici della valutazione umana, con l'obiettivo di ridurre il carico di lavoro per gli esaminatori e a migliorare la coerenza, la rapidità e la scalabilità dei processi di valutazione, soprattutto in contesti caratterizzati da un elevato numero di studenti (Burstein, Kukich, Wolff, Lu, & Chodorow, 1998; Burstein, 2003)[2].

L'avanzamento delle tecnologie informatiche e dei metodi di *machine learning* e *deep learning* ha permesso all'AES di evolversi in modo rilevante, offrendo strumenti capaci di simulare la valutazione umana in maniera sempre più sofisticata, non limitandosi agli aspetti formali dei testi, ma analizzando anche strutture sintattiche, semantiche e stilistiche più complesse (Uto, 2021). Grazie alla disponibilità di ampi corpora annotati e all'affinamento delle tecniche di NLP, l'AES può oggi restituire valutazioni che si avvicinano sempre più a quelle umane, in termini di affidabilità, riproducibilità e precisione (Attali & Burstein, 2006).

L'AES si è sviluppato in sinergia con strumenti progettati per offrire un supporto più diretto agli studenti attraverso correzioni automatiche degli errori nei testi scritti, che rientrano nell'ambito *dell'Automated Writing Evaluation Evaluation* (AWE), e sono impiegati per generare feedback immediati e dettagliati sugli errori linguistici degli studenti, facilitando il processo di revisione e miglioramento dei testi (Cotos, 2018; Woodworth & Barkaoui, 2020).

In ambito italiano, un esempio significativo di implementazione di questo tipo di tecnologia è costituito dal progetto AIDI (AI per l'Apprendimento e il

[2] Per un quadro più ampio sulle potenzialità dell'AES nella risoluzione di problematiche legate al tempo, ai costi, all'affidabilità e alla generalizzabilità della valutazione, si vedano, tra gli altri, Bereiter (2003), Burstein (2003), Chung e O'Neil (1997), Hamp-Lyons (2001), Myers (2003), Page (2003), Rudner e Gagne (2001), Rudner e Liang (2002), Sireci e Rizavi (1999).

Dialogo in Italiano), sviluppato presso l'Università per Stranieri di Perugia in collaborazione con l'Università Telematica IUL di Firenze, che propone l'integrazione di un *chatbot* – chiamato AIDI – nella piattaforma Moodle per supportare l'apprendimento online dell'italiano L2/LS (Cinganotto & Montanucci, 2024a; Cinganotto et al., 2024; Cinganotto & Montanucci 2024b).

AIDI si basa su un'architettura didattico-computazionale che coniuga l'approccio glottodidattico con soluzioni di NLP e *machine learning*, ponendo particolare attenzione alla qualità e all'adeguatezza del feedback linguistico erogato. Istruito sulla base del *Profilo della Lingua italiana*, questo strumento accompagna il discente durante le varie tappe del proprio percorso formativo, stimolandolo a osservare e usare la lingua in contesti che riproducono scenari autentici. Il sistema integrato in AIDI è articolato su due modalità principali: la valutazione del testo scritto e la simulazione dialogica. Nella prima modalità, il sistema è in grado di analizzare in tempo reale le produzioni scritte dell'apprendente, fornendo una restituzione immediata sotto forma di correzioni linguistiche e suggerimenti migliorativi.

L'intervento automatizzato non si limita all'identificazione degli errori ortografici, morfosintattici e lessicali, ma offre proposte di riformulazione contestualizzate, accompagnate da spiegazioni metalinguistiche che favoriscono la consapevolezza linguistica e stimolano la revisione autonoma del testo. Nel secondo caso, l'utente interagisce con il chatbot in situazioni comunicative autentiche simulate, esercitando la scrittura all'interno di contesti d'uso realistici e situati, quali lo scambio informale con un amico, l'ordinazione al ristorante o un colloquio di lavoro. In tali interazioni, il sistema valuta la pertinenza linguistica delle risposte fornite, restituendo un feedback immediato che considera sia la correttezza formale, sia l'adeguatezza comunicativa e la coerenza pragmatica dell'enunciato.

L'evoluzione dell'AES ha seguito un percorso parallelo a quello dell'AWE, condividendone molte delle basi tecnologiche e teoriche, ma con una finalità valutativa più marcata. Se l'AWE si configura primariamente come strumento didattico e formativo, volto a supportare l'apprendente attraverso feedback

personalizzati e orientati al miglioramento progressivo, l'AES nasce con lo scopo di simulare il giudizio umano nella valutazione sommativa di testi scritti, includendo generalmente la coerenza testuale, l'organizzazione argomentativa, l'adeguatezza grammaticale e lessicale, nonché la rilevanza dei contenuti (Shermis & Burstein, 2013); Shermis e Burstein (2013, p. 15) notano come il suo impiego si stia progressivamente estendendo anche ad ambienti di apprendimento come i corsi di lingua L2/LS, in cui la componente valutativa si affianca a quella formativa.

Alla luce di queste considerazioni, l'AES si configura oggi come una tecnologia in continua evoluzione, capace di affiancare e in parte integrare i processi valutativi tradizionali. L'estensione del suo impiego al di fuori dei contesti standardizzati, in particolare nei percorsi di apprendimento linguistico L2/LS, testimonia una crescente attenzione verso il potenziale formativo — oltre che misurativo — di questi strumenti.

Persistono tuttavia diversi punti critici riguardo alla sua applicazione, in particolare per quanto concerne la possibile riduzione della dimensione umana nella valutazione e l'eventuale imprecisione nell'interpretazione di alcuni aspetti contestuali dei testi (Perelman, 2013).

Il presente capitolo intende delineare un quadro teorico e tecnico dell'AES, a partire da una ricostruzione storica delle sue principali tappe evolutive (§2.1), per poi analizzare l'impiego di modelli linguistici avanzati nella valutazione automatica (§2.2) e discutere i principali modelli operativi attualmente utilizzati, con riferimento sia agli approcci computazionali sottostanti sia alle metriche di valutazione (§2.3).

2.1 Storia ed evoluzione degli strumenti AES

Lo sviluppo degli strumenti AES si sviluppa lungo un percorso di oltre mezzo secolo, riflettendo l'evoluzione delle tecnologie linguistiche e dell'AI in parallelo con gli sviluppi della psicometria e della valutazione educativa. I primi tentativi documentati risalgono agli anni Sessanta del Novecento, quando Ellis Page ideò il pionieristico *Project Essay Grade* (PEG), che rappresentò una svolta epocale nell'ambito della valutazione automatizzata (Page, 1966). PEG utilizzava un modello di regressione lineare per predire il punteggio di un elaborato scritto, basandosi su caratteristiche superficiali e misurabili del testo come la lunghezza media delle parole, la densità di punteggiatura, la lunghezza delle frasi e la varietà lessicale (Page, 1968). Tali tratti, noti come *proxy features*, venivano correlati con punteggi assegnati da valutatori umani per addestrare il modello. Per calibrare efficacemente il modello statistico erano necessari inizialmente dai 100 ai 400 elaborati scritti già valutati manualmente; ciò permetteva la stima dei coefficienti necessari per predire successivamente il punteggio di nuovi testi. Pur nella sua apparente semplicità, la metodologia proposta da Page aveva dimostrato come le valutazioni automatiche generate da PEG presentassero frequentemente una forte correlazione con quelle effettuate da valutatori umani, segnalando così una buona affidabilità complessiva. Inoltre, il modello era in grado di fornire un'analisi basilare degli errori grammaticali e strutturali presenti negli elaborati degli studenti.

Nonostante i diversi vantaggi evidenziati dall'autore, il sistema presentava importanti limiti strutturali. Una delle principali criticità risiedeva nel suo focus sugli aspetti superficiali della scrittura, a scapito dell'analisi semantica e contenutistica dei testi. Tale impostazione impediva al sistema di valutare dimensioni fondamentali della produzione scritta, quali la coerenza e la qualità concettuale delle idee, l'organizzazione logica degli argomenti e l'efficacia dell'impianto retorico (Dikli, 2006). Dikli sottolinea inoltre come le prime versioni del sistema PEG presentassero significativi problemi di validità. In particolare, il sistema poteva essere facilmente ingannato attraverso la produzione di testi molto lunghi, privi di coerenza semantica o arricchiti da termini ricercati. Questo

limite derivava dall'eccessiva enfasi posta su indicatori superficiali, che venivano erroneamente interpretati come segnali affidabili della qualità testuale. La studiosa afferma infatti che PEG "[…] has been criticized for ignoring the semantic aspect of essays and focusing more on the surface structures. By failing to detect the content related features of an essay (organization, style etc.), the system does not provide instructional feedback to students. An early version was found to be weak in terms of scoring accuracy. For example, since PEG™ used indirect measures of writing skill, it was possible to "trick" the system by writing longer essays" (Dikli 2006, p. 5). Questa osservazione evidenzia come l'approccio del sistema, basato su misure indirette delle abilità di scrittura, privilegiasse aspetti quantitativi a scapito della profondità contenutistica e dell'effettiva qualità testuale. Ellis Page intervenne successivamente per affrontare queste criticità, introducendo tecniche più sofisticate come *parser* grammaticali e schemi di classificazione più articolati, al fine di migliorare la gamma e la profondità delle caratteristiche analizzate da PEG (Dikli 2006, p. 7).

Negli anni Ottanta e Novanta del Novecento, lo sviluppo dell'elaborazione del linguaggio naturale (NLP) e della psicometria computazionale portò alla creazione di modelli più raffinati. I sistemi cominciarono a integrare tecniche linguistiche e modelli statistici più complessi, culminando nell'*Intelligent Essay Assessor* (IEA) che introdusse l'uso dell'analisi semantica latente (*Latent Semantic Analysis*, LSA) per valutare il contenuto concettuale degli elaborati (Foltz et al., 1999).

Come spiegano Landauer et al. (2003, p. 88) "LSA is a machine learning method that acquires a mathematical representation of the meaning relation among words and passages by statistical computation applied to a large corpus of texts. The underlying idea is that the aggregate of all the word contexts in which a given word does and does not appear provide a set of mutual constraints that largely determines the similarity of meaning of words and sets of words each other. Simulations of psycholinguistic phenomena show that LSA similarity measures are highly correlated with human meaning similarities among words and naturally produced texts. For example, when the system itself, after training,

is used to select the right answer on multiple-choice tests, it scores overlap those of humans on standard vocabulary and subject matter tests. It also closely mimics human words sorting and category judgments, simulates word-word and passage-word lexical priming data and can be used to accurately estimate the learning value of passage for individual student". Questa definizione chiarisce il principio alla base dell'IEA: la capacità di individuare relazioni di significato tra parole e testi attraverso modelli statistici, avvicinando così l'analisi automatica a processi cognitivi di tipo umano.

Lo strumento confrontava il testo prodotto dallo studente con un insieme di testi di riferimento, che potevano comprendere, ad esempio, una serie di testi autorevoli come pubblicazioni scientifiche su un determinato argomento. L'IEA calcolava quindi la somiglianza tra l'elaborato da valutare e i testi di riferimento tramite le metriche LSA, determinando quanto il contenuto dello studente fosse pertinente e completo rispetto al tema assegnato. Nell'assegnazione del punteggio, venivano inclusi feedback sugli aspetti formali della lingua, al fine di produrre un giudizio olistico. Un elemento interessante introdotto da IEA fu la rilevazione del plagio: sfruttando la LSA per rilevare somiglianze anche quando frasi sono parafrasate o riorganizzate, IEA era in grado di individuare produzioni scritte troppo simili tra loro e segnalare potenziali casi di plagio (Landauer et al., 2003, p. 106).

La figura 7, riportata nella pagina successiva, illustra i componenti principali del sistema IEA e il processo di valutazione automatica. L'elaborato dello studente viene analizzato in base a quattro dimensioni: contenuto, stile, meccanica e plagio. Ogni componente contribuisce in percentuale personalizzata alla generazione di un punteggio complessivo, con sottocategorie specifiche come la somiglianza con le fonti, la coerenza testuale e la presenza di errori ortografici. Il sistema include inoltre un modulo di validazione, volto a garantire l'affidabilità della valutazione.

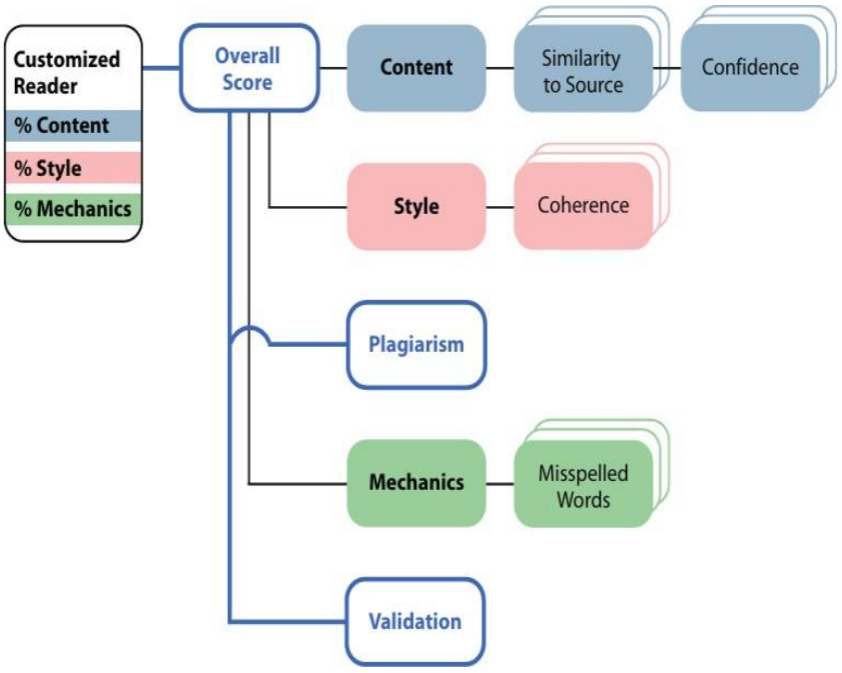


Figura 7. Struttura del sistema IEA (Landauer et al. 2003: 90).

Parallelamente, presso l'Educational Testing Service (ETS) si affermò il sistema *e-rater*. A differenza di IEA, che puntava sulla similarità semantica con LSA, questo modello adottava un insieme di tecniche di NLP tradizionale per analizzare la forma del testo, includendo moduli per l'analisi sintattica, l'analisi del discorso e la valutazione del lessico rispetto al tema fornito (Burstein, Kukich & Wolff, 1998). Come affermato da Burstein e Marcu (2000, p. 3), "The application is designed to identify features in the text that can be linked to writing qualities defined in scoring criteria, used by human readers for manual essay scoring. Each of the modules identifies features that correspond to scoring guide criteria features which can be correlated to essay score, namely, syntactic variety, organization of ideas, and vocabulary usage". In fase di addestramento, *e-rater* viene quindi istruito attraverso centinaia di elaborati scritti già valutati da esaminatori umani (circa 300-500 testi per ogni traccia); da ciascun elaborato, il sistema estrae diverse caratteristiche, come lunghezza media della frase, varietà sintattica, ricchezza lessicale, errori grammaticali o ortografici, presenza di connettivi testuali, struttura del testo, aderenza al tema. Per analizzare le diverse

costruzioni sintattiche e valutare la complessità della frase, lo strumento fa ricorso a dei *parser* grammaticali e impiega un algoritmo vettoriale per rappresentare il contenuto (Burstein, Chodorow, & Leacock, 2003, p. 1). In fase operativa, durante la valutazione di un elaborato scritto, *e-rater* analizza tutte le caratteristiche testuali e le confronta con i modelli statistici appresi. Il punteggio viene quindi assegnato in base all'identificazione della combinazione di caratteristiche più simile a quelle riscontrate nei testi di riferimento (Dikli, 2006, p. 13).

ETS ha sviluppato anche il sistema *Criterion*, una piattaforma AWE/AES che utilizza *e-rater* per fornire punteggi automatici e feedback formativo. Burstein (2003, p. 119) sottolinea che "The advisory component includes feedback to indicate the following qualities on essay response: (a) the text is too brief to be a complete essay (suggesting that the student write more); (b) the essay text do not resemble other essays written about the topic (implying perhaps that the text is off-topic), and (c) the essay response is repetitive (suggesting that the student use more synonyms)". Questa descrizione evidenzia che l'approccio di *Criterion* non è orientato esclusivamente alla valutazione automatica, ma anche fornire un supporto formativo allo studente. Il sistema si configura come uno strumento che promuove un processo di scrittura ricorsivo, in cui la produzione e la revisione del testo si alternano in una prospettiva di miglioramento continuo (Burstein, 2003, p. 120). Il valore pedagogico della piattaforma risiede, dunque, nella possibilità di fornire un riscontro immediato e dettagliato, favorendo l'autonomia dello studente e lo sviluppo di una maggiore consapevolezza linguistica e testuale.

A cavallo tra la fine degli anni Novanta e l'inzio del Duemila, Vantage Learning sviluppa *IntelliMetric*, un sistema AES che combina NLP, reti neurali artificiali e metodi statistici per simulare il processo cognitivo di valutazione (Elliot, 2003). Elliot (2003, p. 72) descrive come "IntelliMetric uses a multistage process to evaluate responses. First, IntelliMetric is exposed to a subset of responses with known scores from which it derives knowledge of the characteristics of each score point. Second, the model reflecting the knowledge derived is tested against a smaller set of responses with known scores to validate the model developed and

confirm generalizability. Third, once generalizability is confirmed, the model is applied to score novel responses with unknown scores". Anche in questo caso, durante l'addestramento, il sistema "assorbe" centinaia di testi già valutati da esperti umani. Gli elaborati vengono poi usati per estrarre la scala di punteggio e le valutazioni associate a ciascun punteggio (Vantage Learning, 2000). Sempre Elliot (2003, p. 72) afferma inoltre che "IntelliMetric has been used to evaluate open-ended, essay type questions in English, Spanish, Hebrew, and Bahsa. Functionality for the evaluation of text in Dutch, French, Portuguese, German, Italian, Arabic, and Japanese is currently available as well". L'autore mette così in luce la portata multilingue del sistema, evidenziando la capacità del sistema di adattarsi a differenti codici linguistici grazie a un'architettura flessibile e a risorse specifiche per ciascuna lingua.

Vantage Learning (2001; 2003) evidenzia come *IntelliMetric* consideri un'ampia gamma di dimensioni relative alla produzione scritta, che raggruppa in cinque categorie principali: focus e coerenza, organizzazione testuale, sviluppo ed elaborazione dei contenuti, struttura della frase, meccanica e convenzioni, riferendosi alla correttezza grammaticale, ortografica e all'uso della punteggiatura. Grazie a tali caratteristiche, lo strumento ha rappresentato un punto di riferimento per lo sviluppo e la diffusione dei sistemi di valutazione automatica degli elaborati scritti. In questa prospettiva, Dikli (2006) osserva come la sua adozione in numerosi contesti istituzionali – dalle scuole e università ai programmi di *testing* standardizzati – abbia contribuito a una progressiva normalizzazione dell'impiego degli AES nell'ambito della valutazione scolastica e accademica.

Nonostante i notevoli progressi compiuti nel tempo, questi sistemi sono stati oggetto di diverse critiche e hanno alimentato un ampio dibattito, in particolare per la tendenza a premiare stili di scrittura artificiali e per la difficoltà nel riconoscere l'originalità argomentativa dei testi (Perelman, 2013; Bennett & Zhang, 2015). Perelman (2013, p. 4) osserva che "[…] even with this biased methodology, however, the data still show that for traditional writing assignments [...] human scorers perform better overall than machines", sostenendo che la

valutazione automatica, pur raggiungendo buoni livelli di affidabilità, risulti comunque inferiore a quella umana nella capacità di cogliere gli aspetti qualitativi e contenutistici della scrittura.

Con l'affermazione del *deep learning*, è emersa una nuova generazione di sistemi AES, basata sull'impiego di reti neurali profonde in grado di apprendere automaticamente rappresentazioni distribuzionali e relazioni sintattico-semantiche più complesse (Taghipour & Ng, 2016; Alikaniotis, Yannakoudakis & Rei, 2016). Questi modelli hanno determinato un significativo miglioramento in termini di accuratezza, potenziando la capacità del sistema di generalizzare su una più ampia varietà di testi.

In anni recenti, l'avvento dei modelli linguistici di grandi dimensioni (LLMs) come GPT-3 e GPT-4 ha aperto nuove prospettive per la valutazione automatica dei testi. LLM come ChatGPT, finemente addestrati su rubriche valutative, si sono dimostrati capaci di generare valutazioni coerenti e ben motivate (Pack, Barrett, & Escalante, 2024). Grazie alla loro capacità di comprensione contestuale e generazione linguistica, i nuovi sistemi si propongono di superare le limitazioni degli AES precedenti, e sono oggi oggetto di sperimentazioni nel contesto educativo e della valutazione assistita. In particolare, l'utilizzo di tecniche di *few-shot learning* — con cui un modello IA impara a fare previsioni accurate addestrando un numero ridotto di esempi — permette a questi sistemi di adattarsi velocemente a diversi contesti valutativi, rendendo possibile l'implementazione su larga scala.

È importante però sottolineare che è proprio questa capacità di rapido adattamento a sollevare diversi interrogativi rilevanti sul piano metodologico, in quanto l'assenza di un ampio set di dati di addestramento rende più difficile garantire la coerenza, la ripetibilità e la trasparenza delle valutazioni generate (Kasneci et al. 2022). Inoltre, Kasneci et al. (2022, p. 10), sottolineano che "While the majority of the research in large language models is done for the English language, there is still a gap of research in this field for other languages. This can potentially make education for English-speaking users easier and more efficient than for other users, causing unfair access to such education technologies for

non-English speaking users. Despite the efforts of various research communities to address multilingualism fairness for AI technologies, there is still much room for improvement". Questa riflessione assume particolare rilievo in rapporto all'impiego dei sistemi AES, in quanto la scarsa rappresentatività delle lingue diverse dall'inglese nei dati di addestramento potrebbe incidere negativamente sulla validità e sull'affidabilità dei punteggi generati dai modelli.

2.2 Modelli linguistici avanzati per la valutazione automatica

I modelli linguistici avanzati costituiscono il nucleo della valutazione automatica degli elaborati scritti, permettendo a questi strumenti di emulare processi cognitivi complessi utilizzati dagli esseri umani nella valutazione del linguaggio scritto. L'evoluzione dell'AES è strettamente legata ai progressi nel campo del *Natural Language Processing* (NLP) e delle tecniche di *machine learning* e *deep learning*, che hanno consentito di sviluppare sistemi sempre più sofisticati e precisi nell'analisi linguistica (Faseeh et al., 2024).

Come evidenziato in precedenza, nelle prime fasi dello sviluppo dell'AES, i modelli si concentravano principalmente su aspetti formali della scrittura, come la grammatica, l'ortografia e la sintassi di base. Questi sistemi, fortemente ispirati al distribuzionalismo di Harris (1954) e al lavoro di Maurice Gross (1997), utilizzavano regole predefinite e lessici grammaticali costruiti manualmente. Attraverso metodi di corrispondenza di pattern, i testi venivano analizzati secondo regole fisse, con un'efficacia limitata nella comprensione semantica (Burstein, 2003).

L'avvento delle reti neurali artificiali e dei modelli basati su deep learning hanno rappresentato un punto di svolta. Le *Long Short-Term Memory* (LSTM) e, successivamente, le architetture *Transformer* hanno permesso di analizzare sequenze linguistiche complesse, mantenendo informazioni a lungo termine e cogliendo dipendenze tra frasi distanti (Vaswani et al., 2017). L'introduzione di modelli linguistici pre-addestrati su larga scala, come BERT e GPT ha segnato un'evoluzione significativa nel campo della valutazione automatica del testo

(AES), ampliandone il potenziale in termini di accuratezza e profondità interpretativa. Questi modelli, addestrati su enormi corpora testuali e caratterizzati da centinaia di milioni di parametri, sono in grado di apprendere rappresentazioni contestuali complesse del linguaggio naturale. In particolare, BERT, grazie alla sua architettura basata sull'attenzione bidirezionale, è in grado di catturare il significato delle parole in funzione sia del contesto precedente che successivo, offrendo così una comprensione linguistica più precisa e articolata. A loro volta, i modelli della famiglia GPT, di natura autoregressiva, sono specializzati nella generazione di testo coerente e nella simulazione di processi di ragionamento linguistico, dimostrandosi particolarmente efficaci nella simulazione degli schemi valutativi umani (Devlin et al., 2019; Brown et al., 2020).

Un tratto distintivo dei modelli linguistici avanzati è la loro capacità di integrare valutazioni olistiche e analitiche. Essi non si limitano più a misurare errori grammaticali, ma interpretano aspetti complessi della scrittura come struttura argomentativa, varietà lessicale, coerenza semantica e creatività. Sistemi come *e-rater* (Attali & Burstein, 2006) utilizzano indici semantici, *n-grams*, misure di leggibilità e complessità retorica per assegnare punteggi globali, avvicinandosi alle modalità di giudizio dei docenti esperti. Inoltre, applicazioni come *Grammarly* o *ProWritingAid* sfruttano reti neurali per offrire feedback personalizzato e contestuale in tempo reale. Oltre alla correzione grammaticale, forniscono suggerimenti su stile, chiarezza e tono, offrendo un supporto prezioso per studenti di L2 e per chi sviluppa testi accademici o professionali (Cotos, 2018).

Tuttavia, nonostante le potenzialità, persistono alcune aree di criticità che richiedono ulteriori approfondimenti. I modelli basati su *deep learning* necessitano di grandi quantità di dati per generalizzare efficacemente (Zhang et al., 2021). La loro opacità decisionale solleva preoccupazioni sulla trasparenza e la spiegabilità (Binns, 2018). Secondo l'Australian Education Union (AEU), l'impiego dell'AES nei test standardizzati, come il NAPLAN, sollecita numerosi interrogativi circa la qualità e l'equità del giudizio automatico, in quanto emerge una tendenza dei modelli a premiare la lunghezza e l'uso di un lessico ricercato

piuttosto che la sostanza argomentativa (Australian Education Union, 2017).

In modo analogo, il National Council of Teachers of English (NCTE) ha formalmente dichiarato la propria contrarietà all'uso di sistemi di valutazione automatica nei contesti scolastici, sottolineando che tali strumenti non possano sostituire il giudizio critico umano, in quanto incapaci di valutare aspetti chiave della scrittura come il tono, l'intento comunicativo e la costruzione retorica (National Council of Teachers of English, 2013). Nel documento in cui argomenta la propria posizione, il NCTE sostiene infatti che "Computers are unable to recognize or judge those elements that we most associate with good writing (logic, clarity, accuracy, ideas relevant to a specific topic, innovative style, effective appeals to audience, different forms of organization, types of persuasion, quality of evidence, humor or irony, and effective uses of repetition, to name just a few). […] some systems gauge the sophistication of vocabulary by measuring the average length of words and how often the words are used in a corpus of texts; or they gauge the development of ideas by counting the length and number of sentences per paragraph. […] Computer scoring favors the most objective, 'surface' features of writing (grammar, spelling, punctuation) […]. Privileging surface features disproportionately penalizes nonnative speakers of English who may be on a developmental path that machine scoring fails to recognize" (National Council of Teachers of English, 2013). Queste considerazioni mettono in luce il rischio che la valutazione automatica riduca la scrittura a un insieme di parametri quantificabili, trascurandone la dimensione cognitiva, comunicativa e culturale.

In prospettiva più ampia, le riflessioni sopra riportate confermano la necessità di un equilibrio tra efficienza tecnologica e profondità interpretativa, affinché l'impiego dei sistemi di valutazione automatica possa integrarsi in modo etico e consapevole nei processi educativi, senza snaturare la complessità del linguaggio umano.

3. ChatGPT: l'intelligenza artificiale nella valutazione automatica dei testi scritti

ChatGPT rappresenta uno dei modelli linguistici generativi più avanzati e ampiamente discussi nel panorama contemporaneo dell'IA. La sua rilevanza si estende anche all'ambito dell'AES in cui costituisce un esempio emblematico dell'evoluzione verso sistemi di valutazione della produzione scritta maggiormente integrati e sensibili alla complessità testuale (Mansour et al., 2024). La crescente attenzione verso l'uso di questi strumenti nei contesti educativi, valutativi e professionali implica una comprensione accurata della loro evoluzione tecnologica e delle implicazioni metodologiche derivanti dal loro impiego nell'AES.

La sezione che segue si concentra dunque sull'evoluzione dei modelli GPT, ripercorrendo le principali tappe che hanno portato alla nascita e all'affermazione di ChatGPT nel panorama AES.

3.1 Evoluzione dei modelli GPT

L'origine di ChatGPT risale all'evoluzione della serie di modelli GPT, sviluppata da OpenAI. Nel giugno 2018 OpenAI pubblicava il report *Improving Language Understanding by Generative Pre-Training*, in cui veniva descritto GPT-1, un modello da 117 milioni di parametri costruito esclusivamente come decoder di un'architettura Transformer a dodici strati[3] (Radford et al., 2018).

L'addestramento di GPT-1 seguiva un protocollo semi-supervisionato: una prima fase di *pre-training* non supervisionato su ampi corpora testuali permetteva al modello di assimilare strutture linguistiche di base, mentre una successiva fase di *fine-tuning* supervisionato su specifici compiti di NLP (ad esempio question answering e analisi del sentiment) raffinava le sue capacità.

[3] Vaswani et al. (2017), definisce l'architettura Transformer a dodici strati come una rete neurale profonda composta da dodici blocchi sequenziali, ciascuno dei quali include meccanismi di self-attention e feed-forward layers. Tale configurazione consente al modello di apprendere rappresentazioni linguistiche complesse, catturando efficacemente relazioni a lungo raggio tra le parole in un testo.

Nel 2019 OpenAI presentò GPT-2 nel report *Language Models are Unsupervised Multitask Learners*. Si trattava di un LLM di seconda generazione costruito esclusivamente con architettura Transformer in modalità *decoder-only* e dotato di 1,5 miliardi di parametri. L'innovazione principale risiedeva nella dimostrazione che l'addestramento non supervisionato su un ampio corpus — il dataset *WebText*, composto da circa 8 milioni di pagine web selezionate in base alla popolarità su Reddit — poteva portare, senza alcun fine-tuning, a prestazioni di livello superiore in compiti di NLP tradizionalmente riservati a modelli supervisionati (Radford et al., 2019a). La decisione di distribuire GPT-2 in modo graduale inizialmente rilasciando soltanto versioni ridotte, per poi rendere disponibile il modello completo da 1,5 miliardi di parametri solo nel novembre 2019 — rifletteva le preoccupazioni di OpenAI circa l'uso potenzialmente improprio di tale tecnologia nella diffusione di disinformazione, e stimolò un ampio dibattito sui meccanismi di governance e sui requisiti di trasparenza per i sistemi di IA (Radford et al., 2019b).

GPT-3, rilasciato nel 2020, ha segnato un avanzamento notevole per la qualità e la versatilità delle risposte, grazie ai suoi 175 miliardi di parametri (OpenAI, 2020). L'addestramento è stato condotto su un corpus ampio, composto da circa 400 miliardi di token, accuratamente selezionati e bilanciati per garantire una varietà di stili, domini e generi testuali (Brown et al., 2020). Come i suoi predecessori, GPT-3 è stato addestrato seguendo un approccio autoregressivo, basato sulla predizione del token successivo: il modello apprende a generare sequenze testuali plausibili proseguendo parola dopo parola, a partire dai contesti e dalle distribuzioni apprese durante la fase di training.

Nel 2022, OpenAI ha rilasciato una serie di modelli noti con il nome GPT-3.5. Pur mantenendo la stessa architettura dei modelli precedenti, GPT-3.5 ha beneficiato di sostanziali ottimizzazioni sia nella fase di pre-addestramento sia in quella di *fine-tuning*. Un elemento centrale di questo progresso è stato l'impiego sistematico della tecnica nota come *Reinforcement Learning from Human Feedback* (RLHF), la quale consente di affinare il comportamento del modello attraverso l'integrazione di preferenze espresse da annotatori umani.

Tale approccio ha contribuito in modo decisivo a migliorare l'allineamento tra le risposte generate dal modello e le aspettative dell'utente (Ouyang et al., 2022).

GPT-3.5 ha costituito inoltre la base tecnica del lancio iniziale di ChatGPT, presentato pubblicamente il 30 novembre 2022 (OpenAI, 2022). Questa versione ha introdotto una modalità di interazione linguistica conversazionale accessibile anche a utenti non esperti, rendendo le capacità generative del modello utilizzabili in una ampissima gamma di contesti applicativi, tra cui l'AES[4].

Nel 2023, OpenAI ha introdotto GPT-4, una versione ancora più avanzata, in grado di gestire input multimodali e migliorare la coerenza logica delle risposte rispetto ai suoi predecessori. GPT-4 rappresenta una significativa evoluzione rispetto a GPT-3.5 non solo per l'aumento della capacità di rappresentazione e comprensione semantica, ma anche per le sue prestazioni nei compiti di ragionamento, comprensione contestuale e coerenza argomentativa (OpenAI, 2023). Secondo la documentazione tecnica di OpenAI, infatti, GPT-4 i distingue per la sua capacità di operare su *windows context* - cioè la quantità massima di testo che il modello può analizzare simultaneamente - estese fino a 32768 token. Ciò consente al modello di analizzare testi molto lunghi in modo continuativo, mantenendo coerenza tra le varie sezioni del discorso e riducendo il rischio di perdita informativa tra l'inizio e la fine dell'elaborazione. Tali sviluppi hanno consentito a GPT-4 di considerare l'intero sviluppo argomentativo di un testo, valutandone la coerenza discorsiva, la coesione sintattica e la strutturazione logica in modo più simile a quanto farebbe un valutatore umano. Queste potenzialità, tuttavia, risultano pienamente efficaci solo quando il modello viene opportunamente guidato attraverso istruzioni precise e contestualizzate. In questo senso, il ruolo del *prompt engineering* — ovvero la progettazione consapevole e strategica degli input forniti al modello — si configura come elemento fondamentale per ottenere risposte rilevanti, mirate e significative.

[4] Cfr. § 3.3.

La tabella 1 illustra alcune tra le principali differenze relative alle versioni sopra citate.

Version	**Uses**	**Architecture**	**Parameter count**	**Year**
GPT-1	General	12-level, 12-headed Transformer decoder (no encoder), followed by linear-softmax with Book Corpus: 4.5GB of text	117 million	2018
GPT-2	General	GPT-1, but with modified normalization with Web Text: 40GB of text	1.5 billion	2019
GPT-3	General	GPT-2, but with modification to allow larger scaling with 570 GB plaintext	175 billion	2020
InstructGPT	Conversation	GPT-3 fine-tuned to follow instructions using human feedback model	175 billion	2022
ProtGPT2	Protein Sequences	As GPT-2, with Protein sequences from UniRef50 of total 44.88 million	738 million	2022

BioGPT	Biomedical Content	As GPT-2 medium (24 layers, 16 heads) with non-empty items from PubMedtota1 1.5 million	347 million	2022
ChatGPT	Dialogue	Uses GPT-3.5, and fine-tuned with both supervised learning and reinforcement learning from human feedback (RLHF)	175 billion	2022
GPT-4	General	Trained with both text prediction and RLHF and accepts both text and images as input, third party data	100 trillion	2023

Tabella 1 - Evoluzione dei modelli GPT con dettagli su utilizzo, architettura e anno di rilascio. Adattato da Ray (2023).

3.2 Prompt engineering e capacità inferenziali

Nel contesto dell'elaborazione automatica del linguaggio naturale, il *prompt* rappresenta una componente fondamentale dell'interazione tra utenti e LLMs. Nei modelli generativi avanzati come GPT-4, il *prompt* si configura come uno spazio semantico e funzionale in cui si attivano, mediante stimoli ben calibrati, le risorse inferenziali latenti del modello, tra cui la capacità di dedurre regole, simulare schemi argomentativi, applicare strutture retoriche e selezionare

registri linguistici adeguati al contesto.

L'insieme delle pratiche metodiche volte alla progettazione strategica del *prompt* è oggi noto come *prompt engineering*, disciplina emergente che si pone l'obiettivo di ottimizzare la coerenza, la pertinenza e l'efficacia dell'*output* generato da sistemi basati su LLMs. Tale approccio consente di modellare in modo controllato il comportamento linguistico del modello attraverso una strutturazione dell'*input*, che può includere la descrizione del compito, l'assegnazione di un ruolo simulato al modello, la definizione esplicita del livello linguistico di riferimento (ad esempio secondo gli standard QCER), l'inserimento di esempi rappresentativi (*few-shot prompting*), e l'articolazione di criteri valutativi in forma di rubriche o scale descrittive.

Yancey et al. (2023) hanno dimostrato come l'impiego di rubriche strutturate basate sul QCER, integrate all'interno del *prompt* insieme a istruzioni esplicite e a richieste di giustificazione del punteggio assegnato, abbia consentito a GPT-4 di fornire valutazioni testuali comparabili a quelle di valutatori umani esperti. Gli autori evidenziano come questa configurazione possa favorire non solo una maggiore trasparenza del processo valutativo, ma anche un incremento significativo della coerenza *inter-rater* tra i punteggi generati dal modello e quelli assegnati manualmente, suggerendo il potenziale dei modelli linguistici di grandi dimensioni come strumenti di supporto affidabili nella valutazione automatica della scrittura in L2/LS. Imperial, Forey e Madabushi (2024), propongono il *framework* STANDARDIZE, concepito per sistematizzare la progettazione dei *prompt* valutativi mediante l'inclusione strutturale di tre artefatti principali: *aspect information* (ossia criteri espliciti di valutazione), *linguistic flags* (indicatori linguistici salienti associati a ciascun livello QCER), ed *exemplars* (esempi di testi rappresentativi). Secondo questo modello, è possibile stabilire un ancoraggio tra l'*output* generato e gli standard valutativi di riferimento, riducendo l'ambiguità interpretativa e aumentando la replicabilità delle risposte del modello. A conferma della flessibilità dei LLMs nel generare contenuti linguistici calibrati, lo studio di Malik et al. (2024) introduce il *Proficiency Control Task* (PCT), un *framework* sperimentale volto a valutare la capacità dei modelli di generare testi coerenti con

specifici livelli QCER, sulla base di *prompt* arricchiti con descrittori qualitativi e lessicali. I risultati del PCT indicano che GPT-4, se guidato da *prompt* contenenti indicazioni esplicite sul livello linguistico desiderato, è in grado di controllare efficacemente la complessità lessicale, la struttura sintattica e la coesione discorsiva del testo prodotto. Di rilievo è anche lo sviluppo del modello CALM, derivato da GPT-4 mediante *fine-tuning* supervisionato e ottimizzazione tramite *Proximal Policy Optimization,* capace di eguagliare e in alcuni casi superare, secondo gli autori, le performance di GPT-4 nel controllo di livello, offrendo allo stesso tempo maggiore efficienza computazionale (Malik et al., 2024). Infine, il contributo teorico e operativo fornito da Phoenix e Taylor (2024) arricchisce il campo del *prompt engineering* con un'architettura metodologica articolata intorno a cinque principi fondamentali: *Give Direction*, *Specify Format*, *Provide Examples*, *Evaluate Quality* e *Divide Labor*. Tali principi, concepiti come linee guida per la progettazione di *prompt* in ambito produttivo e valutativo, trovano applicazione diretta anche nell'ambito AES. Il principio *Give Direction* implica la definizione di ruoli e contesti, elemento chiave per orientare il registro e l'argomentazione del modello; *Specify Format* può garantire coerenza strutturale e facilitare l'integrazione degli output in ambienti automatizzati; *Provide Examples* può attivare meccanismi di generalizzazione controllata tramite strategie *few-shot*; *Evaluate Quality* introduce criteri sistematici per la misurazione dell'adeguatezza e affidabilità dell'output; infine, *Divide Labor* può favorire l'organizzazione modulare dei compiti complessi, migliorando la qualità complessiva della produzione linguistica del modello. L'attribuzione di un ruolo, la definizione del contesto, l'uso di standard come il QCER e scale di assegnazione punteggi sono dunque elementi chiave per ottenere output coerenti e affidabili da GPT-4.

Nella sezione successiva saranno passati in rassegna alcuni recenti studi sulla valutazione automatica dei testi scritti con un focus specifico — ma non esclusivo — sui LLMs. L'obiettivo è quello di offrire una rassegna critica degli approcci proposti, delle metodologie adottate e delle principali implicazioni che tali sistemi comportano nei diversi contesti applicativi.

3.3 Rassegna degli studi sulla valutazione automatica dei testi scritti

Negli ultimi anni, la letteratura scientifica ha dedicato crescente attenzione all'efficacia e all'affidabilità dei sistemi di valutazione automatica dei testi scritti, con particolare riferimento all'integrazione di modelli linguistici avanzati come ChatGPT. Le ricerche convergono su alcuni punti comuni: la capacità degli AES di offrire una valutazione rapida e standardizzata, l'interesse per il loro impiego nella didattica delle lingue, ma anche la presenza di limiti sostanziali legati all'equità, alla trasparenza e alla coerenza delle valutazioni rispetto a quelle umane.

Una delle principali linee di studi riguarda la comparazione tra punteggi generati da modelli automatici e quelli assegnati da valutatori umani esperti. La ricerca condotta da Wang e Brown (2007), ha messo a confronto i punteggi assegnati da *IntelliMetric* con quelli attribuiti da valutatori esperti nell'ambito di due prove di scrittura standardizzate largamente utilizzate negli Stati Uniti: il *WritePlacer Plus* e il *Texas Higher Education Assessment* (THEA). Il campione considerato nello studio comprendeva 107 studenti universitari del Texas. I saggi prodotti sono stati valutati parallelamente da valutatori umani e dal sistema automatico, con l'obiettivo di misurare il grado di concordanza tra le due modalità di valutazione sia in termini di punteggio complessivo sia rispetto a specifiche dimensioni della competenza scritta.

I risultati dell'indagine hanno evidenziato una differenza sistematica nei punteggi assegnati: *IntelliMetric* tendeva ad attribuire valori significativamente più elevati rispetto agli esaminatori umani. L'unica dimensione in cui è stata rilevata una correlazione statisticamente significativa è quella relativa alla struttura della frase. Gli autori interpretano questo dato come indicativo di una maggiore predisposizione del sistema automatico a riconoscere aspetti formali e strutturali della scrittura, quali la correttezza grammaticale e la costruzione sintattica delle frasi, in quanto tali elementi risultano più facilmente codificabili in termini computazionali rispetto a dimensioni più complesse e discorsive, come la coerenza argomentativa, la pertinenza tematica o l'efficacia comunicativa, che

richiedono capacità inferenziali e interpretative più avanzate. Wang e Brown concludono sottolineando la necessità di utilizzare i sistemi di valutazione automatica con prudenza, evidenziando che, se da un lato essi offrono rapidità e coerenza formale, dall'altro rischiano di produrre valutazioni meno rigorose e poco allineate con la complessità del giudizio umano, soprattutto in relazione agli aspetti concettuali, stilistici e pragmatici della scrittura.

In uno studio recente, Kim et al. (2024) hanno analizzato 74 testi argomentativi provenienti dal corpus di testi scritti da studenti del test di livello per la lingua inglese dell'Iowa State University, confrontando i punteggi assegnati da ChatGPT con quelli di valutatori umani. Il modello è stato testato in due condizioni di *prompt* (con o senza input testuale e consegna esplicita) e i risultati hanno mostrato una correlazione solo moderata o bassa con i punteggi umani, variabile in base alla tipologia di traccia. Le principali discrepanze si sono osservate nella valutazione della struttura argomentativa e della coerenza semantica, dove il modello tendeva a penalizzare o sovrastimare determinati aspetti del testo in modo poco trasparente. Gli autori sottolineano la necessità di *prompt* progettati in modo più accurato e fanno riferimento alla responsabilità e all'etica nell'uso dell'AI nella valutazione certificativa.

Un altro studio condotto da Geçkin, Kızıltaş e Çınar (2023), analizza il grado di correlazione tra i punteggi assegnati da cinque valutatori umani e quelli generati da ChatGPT-3.5 in merito a una prova di scrittura in ambito accademico. Il compito prevedeva la redazione di un paragrafo argomentativo da parte di 43 studenti universitari turchi, apprendenti di inglese L2. Per l'assegnazione dei punteggi è stata fornita la stessa scala di livello e competenza a ChatGPT-3.5 e ai valutatori umani. I risultati hanno mostrato che la correlazione tra ChatGPT e i valutatori umani è stata generalmente debole, con coefficienti di Spearman (ρ) compresi tra 0,39 e 0,40, che indicano una relazione limitata tra le due modalità di valutazione. Il livello di accordo è stato valutato anche tramite il coefficiente di Cohen (κ), i cui valori si collocano tra i livelli slight e fair, secondo la classificazione proposta da Landis e Koch (1977). I risultati emersi indicano una corrispondenza solo parziale tra i punteggi prodotti dal sistema automatico e

quelli umani: in particolare, ChatGPT ha mostrato un comportamento più stabile nell'attribuzione dei punteggi a testi di livello intermedio, mentre la sua affidabilità si riduceva nei testi più deboli o più avanzati, segnalando una tendenza a evitare punteggi estremi e a concentrarsi attorno alla media.

La ricerca condotta da Mizumoto et al. (2024) si è concentrata in particolare sulla dimensione dell'accuratezza linguistica. L'obiettivo dello studio era quello di valutare la capacità di ChatGPT, nella versione GPT-4, di individuare errori grammaticali in testi prodotti da apprendenti di inglese L2, confrontandone le prestazioni con quelle di valutatori umani e del correttore automatico *Grammarly*. A tal fine, è stato utilizzato come corpus di riferimento il *Cambridge Learner Corpus – First Certificate in English* (CLC-FCE). I dati analizzati provengono da 232 saggi prodotti da apprendenti asiatici, selezionati in quanto rappresentativi di gruppi con pattern di errore simili. L'accuratezza è stata calcolata come numero di errori per 100 parole, sulla base della codifica manuale presente nel corpus (circa 80 categorie di errore), e successivamente confrontata con le analisi effettuate da ChatGPT e Grammarly.

Dal punto di vista statistico, i risultati mostrano una correlazione molto elevata tra ChatGPT e i valutatori umani nella rilevazione degli errori ($\rho = 0.79$), con valori di affidabilità intra-sistema (Krippendorff's α) pari a 0.96, ottenuti replicando l'analisi su due sessioni distinte. L'output di ChatGPT si è dimostrato stabile e riproducibile, suggerendo una coerenza interna rilevante. Inoltre, le valutazioni fornite da ChatGPT hanno mostrato una correlazione negativa con i punteggi complessivi di scrittura ($\rho = - 0.63$), superiore rispetto a quella ottenuta dai valutatori umani ($\rho = - 0.58$). Ciò suggerisce che, nel contesto specifico del corpus utilizzato, ChatGPT si avvicina — e in alcuni casi supera — il giudizio umano nella capacità di predire la qualità complessiva del testo scritto.

Nel confronto diretto con Grammarly, ChatGPT ha dimostrato una maggiore precisione nella rilevazione degli errori grammaticali, ottenendo una correlazione più alta sia con i giudizi umani ($\rho = 0.79$ vs. 0.69) sia con i punteggi di scrittura ($\rho = -0.63$ vs. -0.55). I dati evidenziano come Grammarly tenda a focalizzarsi su errori di superficie, come ortografia o punteggiatura, mentre

ChatGPT è in grado di riconoscere anche errori più complessi di struttura sintattica e pragmatica.

Tuttavia, gli autori mettono in evidenza alcune limitazioni strutturali. In primo luogo, l'accuratezza è stata misurata esclusivamente tramite il rapporto errori/100 parole, senza considerare altre metriche. Inoltre, gli autori sottolineano che le prestazioni di ChatGPT, pur solide, dipendono fortemente dalla formulazione dei *prompt*, dalla versione del modello e dal dominio del testo, e non possono essere generalizzate senza ulteriori validazioni. La definizione stessa di "errore" rimane un concetto piuttosto flessibile a livello teorico, e la sua variabilità può influire sulle misurazioni.

In merito alla capacità dei sistemi automatici come ChatGPT di assegnare feedback, si segnala l'indagine di Steiss et al. (2024), basata su un campione costituito da 198 studenti delle scuole secondarie statunitensi (gradi 6–12), distribuiti in 12 classi di studi sociali e provenienti da tre diverse scuole pubbliche. I partecipanti includevano sia studenti madrelingua inglese, sia studenti che stavano apprendendo l'inglese come L2 e che partecipavano a regolarmente ai corsi disciplinari. Il compito di scrittura proposto era rivolto all'analisi di eventi storici, con l'obiettivo di far produrre agli studenti un testo argomentativo supportato da fonti primarie. I testi sono stati successivamente utilizzati per generare feedback in due modalità: da un lato, il feedback umano elaborato da valutatori esperti; dall'altro, il feedback prodotto da ChatGPT-3.5, tramite *prompt* costruiti per simulare una revisione autentica focalizzata su miglioramento testuale.

I risultati indicano che, sebbene i valutatori umani abbiano ottenuto punteggi più alti in quattro delle cinque categorie, il feedback prodotto da ChatGPT è stato valutato positivamente, con punteggi medi compresi tra 3,09 e 4,02. L'accordo tra valutatori è risultato elevato con un indice κ compreso tra 0,71 e 0,84, confermando l'affidabilità delle misurazioni. L'unica dimensione in cui ChatGPT ha superato i valutatori umani è stata quella relativa all'adesione esplicita ai criteri valutativi, con un punteggio medio di 3,64 contro 3,40. Tuttavia, nelle dimensioni di accuratezza, priorità e tono, il feedback umano è risultato

significativamente migliore, con differenze medie fino a 0,83 punti a favore dei valutatori.

Un'analisi più dettagliata ha evidenziato come la qualità del feedback generato da ChatGPT diminuisca nei testi più avanzati: i suggerimenti offerti sono stati più precisi e utili nei saggi di livelli più bassi, mentre sono risultati meno pertinenti nei testi di qualità maggiore. Inoltre, gli autori non hanno rilevato differenze significative nella qualità del feedback ricevuto da studenti madrelingua inglese e da apprendenti di inglese L2, né nel caso del feedback umano né in quello automatico. Tuttavia, lo studio ha evidenziato che ChatGPT tendeva a utilizzare un tono meno incoraggiante nei confronti degli studenti più deboli, un aspetto potenzialmente critico per il sostegno alla motivazione degli apprendenti in difficoltà. Gli autori sottolineano pertanto l'importanza di un uso guidato del modello di intelligenza artificiale, in combinazione con una solida formazione all'*AI literacy*, e ne raccomandano l'impiego come strumento di supporto nelle fasi iniziali del processo di revisione, riservando invece quelle più delicate e conclusive alla valutazione umana.

Sempre in merito al feedback, ma con un orientamento più specifico alla dimensione della coerenza e coesione testuale, si riporta il contributo di Yoon, Miszoglad e Pierce (2023), che ha valutato la qualità del feedback fornito da ChatGPT, nella versione GPT-4, su una serie di testi prodotti da studenti apprendenti inglese L2 in contesto scolastico, eterogenei per livelli di competenza, *background* etnici e condizioni economiche. L'indagine ha analizzato 50 testi scritti tratti dal Corpus ELLIPSE[5] e aveva l'obiettivo di valutare la capacità di GPT-4 di fornire riscontri efficaci sulle relazioni logiche e testuali che strutturano un testo scolastico argomentativo. Dal punto di vista metodologico, i testi sono stati

[5] Il Corpus ELLIPSE (English Language Learner Insight, Proficiency and Skills Evaluation) è un corpus liberamente disponibile di circa 6.500 esempi di scrittura di apprendenti di inglese L2, valutati per la competenza linguistica olistica complessiva e per punteggi di competenza analitica relativi a coesione, sintassi, vocabolario, fraseologia, grammatica e correttezza. Il corpus fornisce punteggi di competenza linguistica per i singoli autori dei testi ed è stato sviluppato per promuovere la ricerca sugli approcci basati su corpus e NLP per valutare le caratteristiche generali e più specifiche della competenza. È possibile consultare la risorsa al seguente indirizzo link: https://github.com/scrosseye/ELLIPSE-Corpus

sottoposti a ChatGPT affinché il sistema generasse feedback focalizzati sulla coesione e sulla coerenza testuale.

I feedback prodotti sono stati successivamente analizzati da valutatori esperti, che ne hanno esaminato due aspetti principali: da un lato, l'accuratezza, intesa come la capacità del modello di individuare correttamente punti di forza e criticità a livello coesivo nei testi secondo la scala di valutazione ELLIPSE; dall'altro, l'utilità didattica, valutata in base alla chiarezza dei commenti, alla loro pertinenza rispetto al testo, al grado di specificità e alla presenza di suggerimenti effettivamente orientati al miglioramento.

I dati quantitativi hanno mostrato che i punteggi assegnati da ChatGPT alle produzioni scritte seguendo la stessa scala sono fortemente allineati con quelli dei valutatori umani, a conferma della capacità del sistema di riprodurre coerentemente il giudizio numerico assegnato dai valutatori sulla base dei criteri forniti dalla rubrica analitica ELLIPSE. Tuttavia, l'analisi qualitativa del contenuto dei feedback ha rivelato diverse criticità. In molti casi, i commenti del sistema sono risultati troppo generici, standardizzati, e poco ancorati al testo dell'apprendente. Anche nei casi in cui il modello proponeva esempi tratti direttamente dai testi, le osservazioni erano spesso poco pertinenti o addirittura fuorvianti, in quanto GPT tendeva a focalizzarsi su elementi secondari o a ignorare passaggi realmente critici, senza fornire indicazioni personalizzate o inviti espliciti alla rielaborazione, producendo un feedback decisamente limitato nella sua efficacia formativa. Gli autori evidenziano quindi che, nonostante ChatGPT abbia dimostrato una buona affidabilità nella valutazione numerica, la funzione didattica del feedback sia risultata fortemente limitata dalla mancanza di personalizzazione, dalla superficialità analitica e dall'assenza di indicazioni mirate.

Poole e Polio (2024) hanno indagato l'impatto dell'IA generativa, con particolare riferimento a ChatGPT, sulla scrittura orientata al compito (*task-based writing*) nell'apprendimento linguistico. Gli autori collocano la riflessione in un quadro teorico che coniuga i principi del *Task-Based Language Teaching* (TBLT) con le nuove sfide poste dall'uso di LLMs come strumenti di supporto alla produzione scritta.

Lo studio si sviluppa lungo due direttrici principali: da un lato, un'analisi critica delle potenzialità e dei rischi associati all'uso di ChatGPT nella progettazione e nell'esecuzione di compiti di scrittura autentica in L2; dall'altro, una revisione sistematica della letteratura emergente sull'impiego dell'IA per il feedback e la valutazione nel contesto di compiti comunicativi.

Gli autori evidenziano come ChatGPT venga attualmente impiegato dagli studenti per diverse finalità: dalla generazione iniziale di idee alla riscrittura di frasi, dalla semplificazione lessicale alla verifica grammaticale. Tuttavia, viene segnalato il rischio che il ricorso a tali strumenti possa compromettere l'autenticità del compito comunicativo e la trasparenza nella valutazione delle competenze individuali, soprattutto nei contesti in cui l'*output* dell'IA viene integrato senza una supervisione pedagogica strutturata.

In particolare, viene discussa la tensione tra il principio dell'"output modificabile"[6], centrale nel TBLT, e la natura predittiva e autoreferenziale dei testi generati da ChatGPT, che non si fondano sull'esperienza comunicativa dell'apprendente ma su una modellizzazione probabilistica del linguaggio. Ne consegue il rischio di una riduzione della complessità cognitiva implicata nel processo di scrittura, con un impatto negativo sulla formazione linguistica.

Lo studio mette inoltre in discussione l'impiego non mediato di ChatGPT per la generazione del feedback, evidenziando che, pur potendo offrire correzioni formali corrette, il sistema tende a proporre osservazioni generiche, spesso prive di riferimenti al compito specifico o al contesto comunicativo. Un feedback di questo tipo, se non opportunamente guidato, rischia di essere esclusivamente correttivo, perdendo la sua funzione formativa e processuale. Inoltre, si sottolinea la difficoltà del sistema nel riconoscere l'intenzionalità comunicativa del testo, elemento fondamentale nella valutazione della scrittura orientata al compito.

I due autori sottolineano la necessità di progettare compiti che includano

[6] Il concetto di output modificabile (*modified output*) è stato introdotto da Swain (1995) nell'ambito della sua *Output Hypothesis*, secondo cui la produzione linguistica non ha soltanto una funzione comunicativa, ma anche cognitiva e metalinguistica. L'apprendente, nel tentativo di esprimersi in modo comprensibile, è portato a ristrutturare il proprio output sulla base del feedback ricevuto o di difficoltà percepite, attivando così processi di consapevolezza linguistica (noticing) e di riformulazione

l'interazione con modelli linguistici generativi in modo didatticamente controllato, evitando un uso passivo o sostitutivo dell'IA. In questo quadro, l'integrazione dell'IA deve avvenire in modo coerente con i principi del *task-based learning*, salvaguardando la centralità dell'output autentico dell'apprendente.

Nel complesso, la letteratura recente sul ruolo di ChatGPT nella valutazione automatica della scrittura evidenzia un campo in rapida evoluzione, in cui si intrecciano prospettive ottimistiche e criticità metodologiche ancora irrisolte. Da un lato, numerosi studi segnalano livelli incoraggianti di concordanza tra i punteggi generati dal modello e quelli assegnati da valutatori umani, in particolare nelle dimensioni più formali del testo, come l'accuratezza grammaticale o la struttura sintattica. Dall'altro, emergono costantemente limiti strutturali relativi alla capacità del sistema di interpretare il contenuto, valutare l'organizzazione logica e fornire un feedback autenticamente formativo. La tendenza dei modelli generativi a stabilizzare i punteggi attorno a valori medi, la scarsa personalizzazione dei commenti, l'assenza di consapevolezza del contesto educativo e la difficoltà nel cogliere l'intenzionalità comunicativa dell'autore della produzione scritta costituiscono elementi ricorrenti di attenzione. Pur mostrando buoni livelli di affidabilità interna e riproducibilità, i sistemi AES basati su modelli generativi sembrano tuttora richiedere forme di supervisione e calibrazione, sia nella valutazione sommativa, sia nella produzione di feedback.

4. Metodi di ricerca

Questo capitolo illustra l'impianto metodologico adottato per la realizzazione della presente ricerca. Dopo aver definito l'obiettivo specifico che guida l'intero lavoro empirico (§4.1), verrà descritto in dettaglio il corpus utilizzato per l'analisi, ovvero il Corpus CELI (§4.2), delineandone caratteristiche, composizione e criteri di selezione. Si passerà poi alla presentazione del disegno dello studio (§4.3), con un'attenzione particolare alla sua articolazione in due fasi distinte ma complementari: lo studio pilota (§4.3.1), funzionale alla messa a punto degli strumenti e alla verifica preliminare dell'impostazione metodologica, e lo studio principale (§4.3.2), che rappresenta il cuore dell'indagine.

L'approccio metodologico è stato scelto con l'intento di garantire la massima coerenza tra le ipotesi formulate e i dati raccolti, nonché per assicurare la validità e l'affidabilità dei risultati. Le scelte metodologiche sono motivate tanto da considerazioni teoriche quanto dalla natura dei dati linguistici analizzati, con particolare attenzione agli strumenti e alle tecniche di analisi adottati. In questa prospettiva, il presente capitolo fornisce le basi per comprendere la logica sottesa all'organizzazione dello studio e alla successiva interpretazione dei risultati, che verranno discussi nei capitoli successivi.

4.1 Obiettivo della ricerca

Il presente studio si propone di confrontare i punteggi assegnati da un sistema di valutazione automatica, rappresentato da un modello ChatGPT, istruito esplicitamente per il compito, con quelli attribuiti da valutatori umani esperti a produzioni scritte di apprendenti di italiano L2/LS. Il confronto si colloca in un contesto certificativo autentico, facendo riferimento alle prove CELI.

Le produzioni analizzate provengono dal Corpus CELI (Spina et al., 2022) e sono suddivise per livello secondo il Quadro Comune Europeo di Riferimento per le Lingue (QCER): B1 (CELI 2), B2 (CELI 3), C1 (CELI 4) e C2 (CELI 5). A ciascuna delle produzioni è stato applicato lo stesso criterio valutativo da parte dei due soggetti valutanti — il modello ChatGPT e i esaminatori umani — mediante la medesima scala di livelli e punteggi prevista dalle griglie CVCL, articolata secondo quattro dimensioni fondamentali della competenza scritta: lessicale, grammaticale, sociolinguistica e coerenza e coesione testuale.

L'analisi si basa su un confronto diretto tra le due modalità di valutazione, volto a misurare la variazione tra punteggio umano e punteggio automatico, sia a livello globale che in ciascuna delle dimensioni analitiche. L'obiettivo è verificare se e come si manifestano scarti sistematici, identificando eventuali tendenze lungo i diversi livelli di competenza.

Alla luce di questa premessa, la domanda di ricerca generale che orienta l'indagine è la seguente:

In che misura la valutazione automatica della produzione scritta di apprendenti di italiano L2/LS, effettuata tramite un modello ChatGPT opportunamente configurato, coincide con quella dei valutatori umani in relazione alle dimensioni previste dalla griglia di valutazione CVCL?

Sulla base della letteratura scientifica recente, si formulano le seguenti ipotesi di ricerca. Per quanto riguarda la dimensione lessicale, si ipotizza che ChatGPT tenda ad assegnare punteggi leggermente superiori rispetto ai valutatori umani, soprattutto nei testi di livello intermedio. Tale previsione trova riscontro

nello studio di Geçkin, Kızıltaş e Çınar (2023), i quali rilevano una tendenza del sistema a stabilizzarsi su punteggi centrali, evitando gli estremi.

Per la dimensione grammaticale, a livello generale, si prevede una concordanza elevata tra valutazione automatica e umana. I risultati di Mizumoto et al. (2024) mostrano infatti che ChatGPT, nella rilevazione di errori grammaticali, raggiunge livelli di correlazione molto alti con i giudizi umani. La dimensione sociolinguistica, al contrario, rappresenta un ambito in cui si attende una bassa concordanza.

Studi condotti da Steiss et al. (2024) e da Poole e Polio (2024) evidenziano come i modelli di IA, pur formalmente efficaci, tendano a standardizzare il registro e non riescano a cogliere pienamente l'intenzionalità comunicativa e l'adeguatezza rispetto al contesto. Infine, per la dimensione della coerenza e coesione testuale, si ipotizza una capacità solo parziale da parte di ChatGPT. Sebbene il sistema potrebbe essere in grado di identificare correttamente i legami coesivi superficiali, potrebbe risulta meno efficace nella valutazione della coerenza argomentativa complessiva, come segnalato da Yoon, Miszoglad e Pierce (2023).

4.2 Il Corpus CELI

Il corpus utilizzato per lo svolgimento del presente studio è costituito da un sottoinsieme del Corpus CELI, risorsa realizzata dal gruppo di ricerca, guidato dalla prof.ssa Stefania Spina, dell'Università per Stranieri di Perugia nell'ambito del progetto PRIN 2017 PHRAME – Misure di complessità fraseologica in italiano L2. Il Corpus è stato concepito per caratterizzare in modo sistematico l'interlingua di apprendenti adulti di livello intermedio e avanzato, offrendo dati affidabili sui diversi livelli di competenza linguistica raggiunti secondo il QCER.

La risorsa comprende esclusivamente testi scritti prodotti nell'ambito degli esami di certificazione linguistica CELI, gestiti dal CVCL dell'Università per Stranieri di Perugia. In particolare, nel corpus sono state selezionate le prove scritte tratte dagli esami CELI 2, CELI 3, CELI 4 e CELI 5, corrispondenti

rispettivamente ai livelli B1, B2, C1 e C2 del QCER. La scelta di concentrare la raccolta su questi quattro livelli ha risposto all'obiettivo di ottenere un campione rappresentativo dell'italiano scritto da apprendenti intermedi e avanzati.

La selezione dei testi inclusi nel Corpus CELI si è concentrata esclusivamente sulla Prova di Produzione di testi scritti, ovvero la Parte B degli esami di certificazione CELI, che si articola in compiti specificamente progettati per valutare la capacità del candidato di esprimersi in forma scritta in maniera coerente, appropriata e adeguata ai diversi registri comunicativi.

Come riportato in Spina et al. (2022), i compiti di produzione scritta variano in modo graduale in termini di complessità e finalità comunicativa in funzione del livello del QCER a cui l'esame fa riferimento. A partire dalla struttura generale dei compiti previsti per la Prova di Produzione scritta, così come descritti in dettaglio da Grego Bolli e Pelliccia (2005), sono stati selezionati e inclusi nel corpus i testi prodotti dai candidati dei quattro livelli di certificazione per rispondere ai seguenti compiti: scrivere una breve lettera/e-mail per il CELI 2 (task B.3); una breve composizione su esperienze personali/interessi generali per il CELI 3 (task B.1); una composizione su aspetti della società/racconto di esperienze personali per il CELI 4 (task B.2); e una relazione di un saggio/racconto di fantasia/descrizione di un'esperienza personale per il CELI 5 (task B.1).

Questa selezione ha garantito l'inclusione nel corpus di testi non troppo difformi tra loro per ogni livello, tenendo conto del numero minimo e massimo di parole richiesto nei rispettivi compiti, e privilegiando produzioni maggiormente articolate e più estese in termini di lunghezza rispetto ad altri compiti presenti nella Prova di Produzione scritta. (Spina et al., 2022).

In secondo luogo, sono stati inclusi nel corpus gli elaborati di candidati che avevano superato l'esame e che, per il singolo compito di produzione scritta selezionato, avevano ottenuto almeno la sufficienza in sede di valutazione. Tali accorgimenti hanno assicurato che il corpus rappresentasse produzioni di apprendenti competenti per il livello dichiarato, evitando di includere testi potenzialmente non rappresentativi dello standard del livello.

Per quanto concerne le dimensioni del Corpus, esso contiene complessivamente 3041 testi suddivisi tra i livelli QCER selezionati, e ciascun testo è associato a metadati relativi al candidato, al testo, alla traccia e ai dati del corpus. In particolare, per ciascun candidato sono registrati: il genere, la data di nascita, il numero di matricola e la nazionalità. Quest'ultima informazione è l'unico indizio relativo alla provenienza linguistica dello scrivente, dato che non si dispone della L1 autodichiarata, pertanto non è stato ritenuto un dato utile ai fini della presente ricerca. Sul versante dei dati d'esame, ogni testo riporta il livello QCER della certificazione conseguita, la sede del centro d'esame in cui la prova è stata sostenuta e vari punteggi associati alla performance. In particolare, sono inclusi: il punteggio totale ottenuto nell'intero esame dato dalla somma di prova scritta + prova orale, con indicazione della relativa fascia di valutazione e il punteggio conseguito nella prova scritta (sottoparte d'esame) in quella sessione. Inoltre, ogni elaborato scritto è associato al punteggio specifico attribuito a quella produzione. Questo punteggio specifico è ulteriormente suddiviso nei quattro criteri valutativi utilizzati dai certificatori CELI: competenza lessicale, competenza grammaticale, competenza sociolinguistica e coerenza e coesione testuale. Tali criteri riflettono la scala di livelli e punteggi del CVCL.

Infine, riguardo alla traccia, che rappresenta il compito assegnato, per ogni testo si registra un identificativo univoco della traccia di produzione scritta svolta dal candidato. Nell'insieme, dunque, il Corpus offre metadati dettagliati per caratterizzare chi ha scritto, che cosa ha scritto (genere e tipo di testo) e in che modo è stato valutato quel testo specifico.

Uno degli aspetti distintivi del Corpus CELI è rappresentato dalla sua elevata accessibilità, resa possibile grazie alla pubblicazione su CQPweb (Corpus Query Processor web), un'interfaccia descritta da Hardie (2012), che consente interrogazioni testuali e linguistiche complesse su corpora annotati. Spina evidenzia come rendere liberamente fruibile online questo tipo di corpus rappresenti un elemento di grande rilevanza, in quanto fornisce agli studiosi dati di alta qualità su cui testare ipotesi e agli enti certificatori uno strumento di benchmark per le proprie prove.

Dal punto di vista della didattica e valutazione, in particolare, il Corpus offre un solido riferimento empirico: i dati provenienti da contesti d'esame reali possono supportare lo sviluppo di test di lingua basati su corpora, contribuendo a rendere le prove più tarate sull'effettiva produzione dei candidati in termini di difficoltà delle forme linguistiche realmente usate a ogni livello. Inoltre, il corpus può essere utilizzato per affinare la formazione dei valutatori e la taratura delle griglie di valutazione: disponendo di migliaia di testi già valutati con punteggi dettagliati, è possibile estrarre esempi di produzioni tipiche da presentare nei corsi per esaminatori, o analizzare in che modo alcuni tratti linguistici possano incidere sui punteggi assegnati.

4.3 Disegno dello studio e metodologia adottata

Il presente lavoro si basa su un'indagine sperimentale articolata in due fasi, distinte ma interconnesse, concepite per esplorare e valutare l'affidabilità di un modello AES basato su ChatGPT-4 e per confrontarlo con l'assegnazione dei punteggi da parte dei valutatori esperti nell'ambito delle prove esaminate per le diverse tipologie di certificazioni CELI. La metodologia adottata si colloca all'interno di un impianto sperimentale di tipo quantitativo, basato sull'elaborazione di dati linguistici autentici estratti dal Corpus CELI, in riferimento alle prove di produzione scritta degli esami CELI 2, CELI 2, CELI 3, CELI 4 e CELI 5, e valutate secondo i criteri forniti dalle scale di punteggio elaborate e adottate dal CVCL dell'Università per Stranieri di Perugia.

La scelta di articolare la ricerca in due fasi risponde a un'impostazione metodologica progressiva, finalizzata a verificare l'affidabilità del protocollo sperimentale, prima di estenderne l'applicazione su un campione più ampio.

Un passaggio chiave nell'impostazione dello studio ha riguardato la definizione del protocollo valutativo automatizzato, finalizzato a sottoporre testi autentici a ChatGPT, nella versione basata sull'architettura GPT-4 sviluppata da OpenAI, in condizioni valutative strutturate secondo i criteri certificativi CELI. La scelta di adottare GPT-4 è maturata in seguito all'analisi della documentazione

tecnica rilasciata da OpenAI (2023), che descrive un'evoluzione rispetto ai modelli precedenti in termini di capacità di interpretare istruzioni complesse, mantenere coerenza nella produzione e rispondere in modo articolato a compiti linguistici strutturati.

Nel contesto della ricerca, si è dunque deciso di sperimentare l'impiego di questo modello, istruito tramite *prompt* specifici, per la valutazione automatica delle produzioni scritte in L2/LS, in virtù di caratteristiche che ne indicano una possibile idoneità nel gestire compiti valutativi guidati e a rispondere in modo coerente a istruzioni formali.

Nel disegno complessivo dell'indagine, la fase sperimentale è stata avviata da uno studio pilota finalizzato a testare, in un contesto controllato e su un numero limitato di testi, la coerenza interna delle valutazioni prodotte da ChatGPT. Questa prima fase è stata pensata per verificare la reattività del modello a *prompt* specificamente progettati per la valutazione della produzione scritta e di osservare la stabilità dei punteggi generati in cicli valutativi indipendenti. I risultati ottenuti da questa fase hanno offerto indicazioni preliminari utili a validare il protocollo sperimentale e a predisporre in modo più mirato l'estensione sistematica dell'indagine nella fase successiva. Per la descrizione dello studio pilota, si rimanda a § 4.4.

A questa prima fase ha fatto seguito lo studio principale, strutturato secondo un disegno sperimentale sistematico volto a rendere possibile l'osservazione di eventuali corrispondenze o discrepanze tra il sistema valutativo umano e automatizzato. In questo contesto, il modello GPT è stato impiegato per valutare un campione di 800 testi autentici estratti dal Corpus CELI, distribuiti in modo bilanciato sui livelli B1, B2, C1 e C2 del QCER e selezionati tramite un criterio randomizzato. I testi, come sopra menzionato, sono stati prodotti da candidati nel contesto delle prove di certificazione CELI 2, CELI 3, CELI 4 e Celi 5, e rispondono a compiti comunicativi differenziati per livello, progettati per valutare la capacità di esprimersi in forma scritta in situazioni coerenti con la vita quotidiana, accademica o professionale. La selezione del campione ha incluso

produzioni associate a tracce diverse, rappresentative della varietà testuale proposta dagli esami CELI, al fine di garantire un confronto attendibile tra fasce di competenza.

Il disegno sperimentale è stato strutturato per consentire il confronto tra i punteggi assegnati dal modello e quelli forniti da valutatori esperti, con riferimento a quattro dimensioni linguistiche specifiche: competenza lessicale, grammaticale, sociolinguistica e coerenza/coesione testuale. Per una trattazione approfondita dello studio principale si rimanda al § 4.5.

4.4 Studio Pilota

La presente sezione descrive lo studio pilota, condotto sul Corpus CELI mediante l'impiego del modello GPT-4, con l'obiettivo di testare la coerenza e la stabilità della valutazione automatica delle produzioni scritte in italiano L2/LS.

In questa prima fase sperimentale, è stato selezionato un campione ridotto di dodici testi autentici, rappresentativi di diversi livelli di competenza, al fine di osservare in modo controllato il comportamento del modello prima dell'estensione della ricerca.

Tale fase ha rappresentato un passaggio esplorativo fondamentale nell'ambito dello studio, in quanto ha consentito di verificare in maniera preliminare l'affidabilità del modello nell'assegnazione di punteggi alle produzioni scritte dei candidati, in vista di un successivo confronto sistematico con i punteggi assegnati dal valutatore umano.

Per la realizzazione dello studio, sono stati selezionati dodici testi rappresentativi di quattro diversi livelli di competenza linguistica per i livelli QCER: B1, B2, C1 e C2. La selezione ha previsto un campionamento randomizzato di tre produzioni scritte per ciascun livello, estratte dal corpus CELI e indicative delle tipologie testuali e della complessità linguistica specifica di ogni fascia di competenza. Ogni testo prodotto dal candidato è stato associato a una tipologia di compito della Prova di Produzione di testi scritti degli esami CELI.

Il criterio di selezione ha mirato a garantire la varietà interna del campione,

nonché la possibilità di osservare l'eventuale incidenza del livello di competenza sull'andamento della valutazione automatica.

Ogni testo è stato sottoposto al modello ChatGPT, realizzato mediante l'uso di un *prompt* appositamente costruito e pensato per indirizzare la valutazione verso la considerazione di un parametro rappresentativo dalla scala CELI, cioè la dimensione lessicale.

Il *prompt* è stato strutturato con particolare attenzione alla definizione chiara del ruolo del modello, alla presentazione esplicita del compito e alla specificità dei criteri attesi di risposta. Si riporta di seguito il *prompt* impiegato per l'interazione con il modello.

Sei un valutatore professionista esperto e devi valutare il testo prodotto dall'apprendente di italiano lingua non materna che trovi nella sezione TESTO DA VALUTARE. Il testo è stato scritto per rispondere al compito che trovi nella sezione COMPITO. Per la valutazione usa i criteri che trovi nella sezione CRITERI DI VALUTAZIONE. Come risposta devi fornirmi due sezioni:

- una prima chiamata SPIEGAZIONE in cui spiegherai passo dopo passo come hai applicato i criteri
- una seconda chiamata PUNTEGGIO contenente il punteggio della valutazione

#COMPITO
<COMPITO>
TESTO DA VALUTARE
<TESTO ELABORATO DAL CANDIDATO>
#CRITERI DI VALUTAZIONE
<LIVELLO QCER E SCALA DI COMPETENZA E PUNTEGGI>

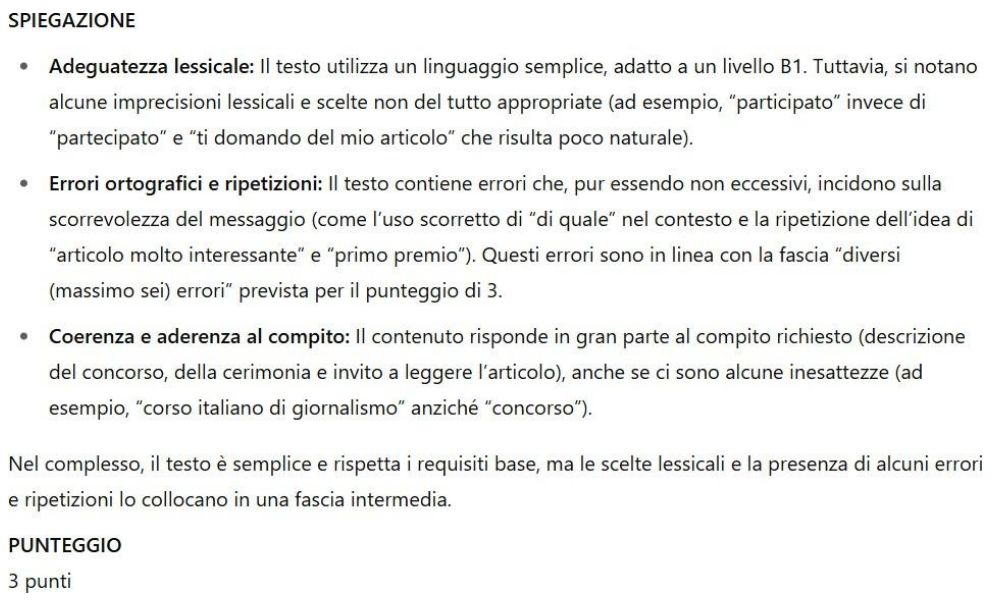
Nella prima sezione, viene assegnato al modello il ruolo di un valutatore professionista esperto, la fine di simulare una situazione valutativa autentica. Questa strategia è stata adottata per orientare il modo in cui il GPT struttura e presenta le proprie valutazioni, con l'obiettivo di renderle il più possibile simili a quelle fornite da un valutatore umano, sia in termini di contenuto che di argomentazione.

Segue la sezione "TESTO DA VALUTARE", che costituisce il punto di

partenza dell'attività valutativa e di assegnazione del punteggio e contiene la produzione scritta dell'apprendente. Ogni testo è contestualizzato dal relativo "COMPITO", che corrisponde alla traccia d'esame della Prova di Produzione scritta. L'abbinamento fra testo e compito consente al modello di ancorare la propria valutazione al tipo di performance attesa.

Nella sezione "CRITERI DI VALUTAZIONE", il *prompt* contiene l'indicazione esplicita alla scala CELI di riferimento, con particolare attenzione, nello studio pilota, alla competenza lessicale e ai relativi punteggi, per consentire al modello di orientarsi in base a criteri chiari e vincolanti.

Il *prompt* richiede poi di articolare la risposta in due sezioni distinte: "SPIEGAZIONE" e "PUNTEGGIO". La sezione "SPIEGAZIONE" è stata inserita al fine di rendere trasparente il processo valutativo di GPT, simulando il ragionamento di un valutatore esperto umano nella giustificazione dei criteri applicati. A titolo esemplificativo, si riporta di seguito la figura 7, che presenta un estratto della sezione "SPIEGAZIONE" così come generata dal modello. L'esempio evidenzia il tipo di argomentazione adottata dal sistema nell'applicazione dei criteri valutativi, offrendo un'evidenza del processo inferenziale che sottende l'assegnazione del punteggio.



SPIEGAZIONE

- **Adeguatezza lessicale:** Il testo utilizza un linguaggio semplice, adatto a un livello B1. Tuttavia, si notano alcune imprecisioni lessicali e scelte non del tutto appropriate (ad esempio, "participato" invece di "partecipato" e "ti domando del mio articolo" che risulta poco naturale).
- **Errori ortografici e ripetizioni:** Il testo contiene errori che, pur essendo non eccessivi, incidono sulla scorrevolezza del messaggio (come l'uso scorretto di "di quale" nel contesto e la ripetizione dell'idea di "articolo molto interessante" e "primo premio"). Questi errori sono in linea con la fascia "diversi (massimo sei) errori" prevista per il punteggio di 3.
- **Coerenza e aderenza al compito:** Il contenuto risponde in gran parte al compito richiesto (descrizione del concorso, della cerimonia e invito a leggere l'articolo), anche se ci sono alcune inesattezze (ad esempio, "corso italiano di giornalismo" anziché "concorso").

Nel complesso, il testo è semplice e rispetta i requisiti base, ma le scelte lessicali e la presenza di alcuni errori e ripetizioni lo collocano in una fascia intermedia.

PUNTEGGIO

3 punti

Figura 7 - Esempio di output generato dal modello nella sezione SPIEGAZIONE.

L'inserimento della sezione PUNTEGGIO, a sua volta, è finalizzato a indirizzare GPT a produrre il punteggio numerico compatibile con la scala CELI, rendendo possibile il confronto diretto con i punteggi assegnati dai valutatori umani.

I segni grafici di apertura e chiusura angolare sono usati come marcatori per racchiudere etichette identificative di specifiche sezioni, come "<COMPITO>", "<TESTO ELABORATO DAL CANDIDATO>" e "<LIVELLO QCER E SCALA DI COMPETENZA E PUNTEGGI>", al fine di segnalare la posizione e la funzione di contenuti effettivamente presenti in fase di esecuzione dello stesso *prompt*, e ne rappresentano la struttura in modo generalizzato ma conforme all'originale.

Dal punto di vista strutturale e funzionale, il *prompt* segue una configurazione gerarchica e leggibile tramite sintassi Markdown, uno standard di formattazione facilmente decodificabile da LLM come GPT, i quali ne riconoscono e ne sfruttano le regolarità per ottimizzare la segmentazione e l'interpretazione delle istruzioni.

Al fine di testare la stabilità interna del modello e di verificare la presenza di eventuali oscillazioni nei giudizi, la procedura di valutazione è stata ripetuta per dieci cicli consecutivi per ciascun testo. In ogni ciclo, il modello ha prodotto un punteggio indipendente, senza memoria delle valutazioni precedenti, in modo da assicurare l'indipendenza statistica delle misurazioni.

4.5 Studio principale

La seconda fase della sperimentazione rappresenta il nucleo centrale di questa ricerca, in quanto ha permesso di indagare in modo sistematico la relazione tra la valutazione umana e quella automatica applicata alla produzione scritta di apprendenti di italiano L2/LS. Mentre lo studio pilota ha avuto la funzione di esplorare la coerenza e l'affidabilità del modello, questa fase ha esteso il protocollo sperimentale su un ampio campione di produzioni autentiche, al fine di predisporre un confronto quantitativo attendibile tra le modalità di attribuzione

del punteggio automatiche e umane. Il corpus analizzato è composto da 800 testi scritti selezionati dal Corpus CELI, suddivisi in modo bilanciato tra i livelli di competenza linguistica B1, B2, C1 e C2 del QCER, con 200 testi per ciascun livello.

Le produzioni prese in esame sono state redatte da candidati durante sessioni ufficiali degli esami per le certificazioni CELI, e sono associate alle tracce presenti nella sezione di Produzione scritta. I compiti proposti, differenziati per livello, sono progettati per verificare la capacità degli apprendenti di interagire in forma scritta in situazioni comunicative realistiche attraverso testi di varia natura: descrittivi, narrativi, argomentativi o epistolari, in forma formale o informale.

La selezione delle produzioni ha tenuto conto della varietà testuale per cercare di rappresentare in maniera bilanciata la gamma di compiti previsti all'interno dell'esame, attenendosi comunque alla tipologia di prova prevista. Più nello specifico, per il livello B1 (CELI 2), è stata presa in esame la produzione di una breve lettera o e-mail a partire da una traccia fornita, corrispondente al task B.3; per il livello B2 (CELI 3), le produzioni analizzate rispondono a compiti di composizione su esperienze personali o temi di attualità, selezionati tra due alternative proposte nel task B.1; per il livello C1 (CELI 4), sono stati considerati testi redatti che prevedevano la redazione di una relazione, un racconto o una lettera formale, come previsto dal task B.2; per il livello C2 (CELI 5), sono stati selezionati i testi prodotti in risposta a uno dei tre input proposti, tra cui figurano saggi, racconti di fantasia o descrizioni di esperienze personali, corrispondenti al task B.1.

Per ciascun testo incluso nel campione, era disponibile nel Corpus CELI il punteggio attribuito da un valutatore umano esperto, assegnato secondo i criteri forniti dalle scale di punteggi e di livello del CVCL e riferito a ciascuna delle quattro dimensioni previste dalla griglia (lessicale, grammaticale, sociolinguistica e coerenza e coesione testuale).

A partire da tali produzioni, è stata condotta una nuova procedura di valutazione, affidata al modello GPT, al quale è stato richiesto di applicare i medesimi criteri tramite il *prompt* strutturato e validato nello studio pilota, per simulare il contesto operativo e le istruzioni fornite ai valutatori umani.

Il successivo confronto tra la valutazione umana e quella generata dal modello è stato condotto con l'obiettivo di verificare il grado di variazione tra le due attribuzioni di punteggio.

Per ogni osservazione, sono stati raccolti i seguenti dati: la dimensione valutata, il livello QCER per il quale la candidata o il candidato ha sostenuto la prova, l'identificatore univoco della produzione, la traccia assegnata, il punteggio attribuito dal valutatore umano, il punteggio assegnato dal modello, il punteggio massimo previsto per la dimensione e lo scarto tra i due punteggi, calcolato in forma di errore normalizzato. Quest'ultimo valore ha consentito di misurare la distanza tra le due valutazioni, rapportandola alla scala adottata, e di rendere comparabili i risultati.

5. Risultati

Questo capitolo è dedicato alla presentazione dei risultati emersi dalle due fasi del percorso sperimentale descritte nelle precedenti sezioni. L'obiettivo è illustrare in modo sistematico i dati raccolti in ciascuna fase e introdurre gli elementi quantitativi su cui si fonda la successiva discussione critica.

Coerentemente con l'impostazione metodologica adottata, i risultati vengono presentati in modo distinto per ciascuna fase della sperimentazione.

In § 5.1 vengono illustrati gli esiti dello studio pilota, volto di testare la coerenza interna del modello e la stabilità nell'assegnazione dei punteggi. In § 5.2, invece sarà presentata l'analisi dei dati relativi allo studio principale, finalizzato al confronto sistematico tra l'attribuzione automatica e umana dei punteggi alle produzioni scritte.

L'analisi si articola attraverso una lettura descrittiva e comparativa dei punteggi, supportata da indicatori statistici per osservare le variazioni tra le due modalità valutative.

5.1 Risultati dello studio pilota

A seguito della raccolta dei dati raccolti durante lo studio pilota, è stata effettuata un'analisi quantitativa finalizzata a verificare la coerenza e la stabilità delle valutazioni generate da ChatGPT.

In particolare, l'indagine si è concentrata sull'osservazione della deviazione standard dei punteggi assegnati a testi appartenenti a diversi livelli di competenza linguistica, al fine di individuare eventuali differenze nella capacità del modello di mantenere un giudizio costante al variare della complessità linguistica delle produzioni.

La Tabella 2, riportata nella pagina successiva, illustra i risultati ottenuti, offrendo una panoramica della distribuzione dei punteggi e delle relative variazioni.

Liv. QCER	Combinazione	Voto1	Voto2	Voto3	Voto4	Voto5	Voto6	Voto7	Voto8	Voto9	Voto 10	Media	Dev. Std.
B1	ID 28 traccia 1	4	4	4	4	4	4	4	4	4	4	4	0
B1	ID 976 traccia 9	3	3	3	3	3	3	3	3	3	3	3	0
B1	ID 1828 traccia 17	2	2	2	2	3	2	2	2	2	3	2,2	0,4
B2	ID 2001 traccia 18	3	3	3	3	3	3	3	3	3	3	3	0
B2	ID 2109 traccia 26	4	4	4	4	4	4	4	4	4	4	4	0
B2	ID 2111 traccia 27	4	4	4	4	4	4	4	4	4	4	4	0
C1	ID 1 traccia 4	7	7	7	7	7	7	8	7	7	7	7,1	0,3
C1	ID 1192 traccia 13	7	7	7	7	7	7	7	6	7	7	6,9	0,3
C1	ID 3003 traccia 62	5	5	5	4	6	6	6	5	6	6	5,4	0,66
C2	ID 1793 traccia 24	8	7	7	7	7	8	8	8	7	7	7,4	0,48
C2	ID 2891 traccia 32	8	8	9	9	9	9	8	9	9	8	8,6	0,48
C2	ID 2511 traccia 38	7	8	8	8	7	7	7	7	7	7	7,3	0,45

Tabella 2 - Distribuzione dei punteggi assegnati dal modello realizzato.

L'analisi dei dati evidenzia una leggera oscillazione nella stabilità valutativa tra i livelli inferiori (B1 e B2) e quelli superiori (C1 e C2).

In particolare, per i livelli B1 e B2, la deviazione standard risulta nulla in quasi tutte le combinazioni e attesta, in tal modo, una capacità del modello di produrre valutazioni estremamente stabili e ripetibili su testi caratterizzati da una complessità linguistica contenuta. La totale assenza di oscillazione nella maggior parte dei casi suggerisce che il modello riconosce con precisione e coerenza il livello di competenza espresso dai testi più semplici, producendo un giudizio replicabile indipendentemente dal ciclo di valutazione. L'unico caso di leggera variabilità nel livello B1 è rappresentato dalla combinazione ID 1828 traccia 17, che registra una deviazione standard pari a 0,4. Tale valore, sebbene più elevato rispetto agli altri della medesima fascia, risulta comunque contenuto e indicativo di una stabilità valutativa alta.

Per quanto riguarda i livelli superiori, C1 e C2, la stabilità della valutazione si mantiene comunque buona, sebbene si registri un incremento della deviazione standard. I valori oscillano infatti tra 0,3 e 0,66, a conferma di come l'aumentare della complessità linguistica e sintattica delle produzioni avanzate introduca un margine di variabilità nella risposta del modello. Tale fenomeno è

particolarmente evidente nella combinazione ID 3003 traccia 62 nel livello C1, che presenta la deviazione standard più alta, corrispondente a 0,66. Questo dato può essere interpretato alla luce delle difficoltà che il modello può incontrare nella valutazione di testi caratterizzati da una maggiore densità informativa e da una varietà lessicale e strutturale più marcata.

Nonostante ciò, la variabilità riscontrata si mantiene entro limiti accettabili e non pregiudica la complessiva affidabilità della valutazione automatica. La buona coerenza osservata anche ai livelli C1 e C2 suggerisce che il modello, pur esposto a testi linguisticamente più articolati, è riuscito comunque a mantenere un buon grado di uniformità nel giudizio.

L'analisi quantitativa condotta sui dati del campione pilota ha dunque evidenziato una sostanziale stabilità nelle valutazioni fornite dal modello, fornendo una base metodologicamente affidabile per l'avvio della fase successiva della ricerca.

5.2 Risultati dello studio principale

Dopo aver verificato la stabilità interna del modello attraverso lo studio pilota, questa seconda fase dell'analisi si concentra sul confronto tra i punteggi assegnati da ChatGPT e quelli attribuiti da valutatori umani esperti, relativi alle prove CELI di produzione scritta.

L'obiettivo è osservare il grado di variazione tra valutazione automatica e valutazione umana, in riferimento alle quattro dimensioni linguistiche delle scale di competenza e punteggio CVCL, e valutare l'incidenza di eventuali scostamenti in rapporto al livello della prova sostenuta.

5.2.1 Panoramica complessiva

La tabella 3 presenta la media di errore[7], calcolata per ciascuna delle quattro dimensioni linguistiche previste dalla griglia CELI (coerenza e coesione, grammatica, lessico, sociolinguistica), suddivisa per livello QCER della prova sostenuta dai diversi candidati.

Livello QCER	Coerenza e coesione	Competenza grammaticale	Competenza lessicale	Competenza sociolinguistica	Totale complessivo
B1	12,4	14,0	14,8	14,0	13,8
B2	18,4	15,6	19,2	16,4	17,4
C1	13,0	14,3	12,0	13,3	13,1
C2	19,5	12,3	12,5	13,0	14,3
Totale medio	15,8	14,0	14,6	14,2	14,7

Tabella 3. Media di errore (%) tra valutazione umana e automatica per ciascuna dimensione linguistica e livello QCER della prova sostenuta.

L'analisi dei dati evidenzia una certa variabilità sia tra le dimensioni valutative, sia tra i livelli delle prove. A livello complessivo, la coerenza e coesione testuale presenta la media di errore più elevata (15,8%), seguita dalla competenza lessicale (14,6%), da quella sociolinguistica (14,2%) e infine dalla grammatica (14,0%). La figura 8 fornisce una panoramica generale dei risultati emersi.

[7] Per media di errore, si intende lo scarto assoluto tra il punteggio assegnato dal modello GPT e quello attribuito dal valutatore umano rapportato al punteggio massimo previsto per la dimensione. Il valore risultante, espresso in percentuale, consente di misurare in modo proporzionale la distanza tra le due valutazioni, indipendentemente dall'ampiezza della scala utilizzata.

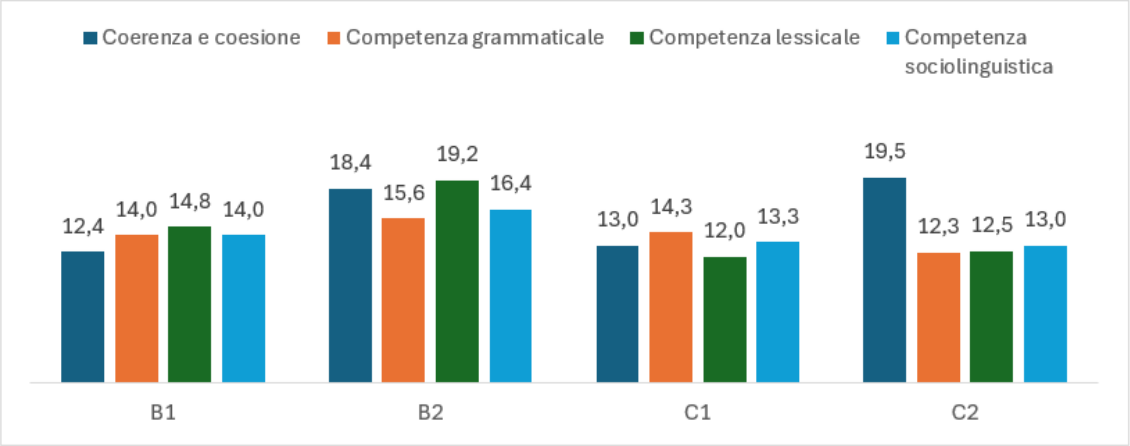


Figura 8. Media di errore (%) tra valutazione umana e automatica per ciascuna dimensione linguistica e livello QCER della prova sostenuta.

Con riferimento ai livelli delle prove, il quadro risulta differenziato. Il livello B1 presenta valori contenuti e distribuiti in modo relativamente omogeneo tra le dimensioni, con un range compreso tra il 12,4% (coerenza e coesione) e il 14,8% (lessico). Ciò può riflettere una maggiore prevedibilità e semplicità strutturale dei testi prodotti, che facilitano la convergenza tra le due modalità di valutazione.

I risultati relativi al livello B2 evidenziano sorprendentemente le variazioni più consistenti tra valutazione automatica e valutazione umana, con scarti significativamente superiori rispetto a quelli registrati per gli altri livelli. In quasi tutte le dimensioni analizzate si osservano una media errore percentuale superiore al 15, con un picco del 19,2 nella competenza lessicale e una percentuale del 18,4 nella coerenza e coesione testuale. Anche la dimensione grammaticale e la dimensione della coerenza e coesione mostrano valori elevati di discrepanza, rispettivamente pari al 17,4 e al 15,4.

La distribuzione di tali dati mostra che le variazioni si concentrano in modo più netto nelle dimensioni linguistiche che richiedono un'elaborazione autonoma e consapevole da parte del parlante, ovvero nella selezione, combinazione e articolazione delle risorse lessicali, sintattiche e testuali.

Queste evidenze lasciano ipotizzare che le produzioni ascrivibili a questa soglia possano dar luogo, più di altre, a giudizi divergenti, nonostante l'impiego di un unico strumento valutativo, ovvero la griglia ufficiale CELI 3. Pur

condividendo la medesima base descrittiva, infatti, il modello automatico e il valutatore umano sembrano adottare criteri di applicazione sensibilmente differenti, in particolare nei confronti di scelte linguistiche che si collocano in una zona di ambiguità tra l'adeguatezza funzionale e la complessità espressiva.

Il livello C1 si caratterizza per una maggiore regolarità: i valori oscillano tra il 12,0%, risultato relativo alla dimensione lessicale e il 14,3% della dimensione grammaticale, con una media complessiva tra le più basse dell'intero corpus analizzato (13,1%).

Nel livello C2 i valori di errore risultano contenuti nella dimensione grammaticale (12,3%), nel lessico (12,5%) e nella dimensione sociolinguistica (13,0%), ma registrano un picco del 19,5% nella dimensione della coerenza e coesione testuale. Questo scarto potrebbe indicare che, nei testi più liberi e articolati tipici delle prove di livello C2, il modello incontra maggiori difficoltà nel valutare l'organizzazione discorsiva complessiva, dimensione che richiede inferenze globali e un'interpretazione più sottile dell'intento comunicativo.

Nel loro insieme, questi dati forniscono una prima lettura delle tendenze valutative di ChatGPT in rapporto al punteggio assegnato dal valutatore umano, suggerendo l'esistenza di aree di maggiore divergenza tra modello e valutatori umani, sia sul piano delle dimensioni, sia in relazione al livello QCER della prova sostenuta. Le sezioni successive si propongono di approfondire questi risultati attraverso un'analisi incrociata che metta in relazione ciascuna dimensione linguistica con i singoli livelli QCER.

5.2.2 Risultati relativi alla dimensione lessicale

La lettura dei dati mostra una distribuzione non lineare dell'errore tra i diversi livelli, con una media complessiva pari al 14,6%. Il valore più elevato si registra al livello B2 (19,2%), seguito dal B1 (14,8%), mentre i valori più bassi riguardano il livello C2 (12,5%) e, in particolare, il C1 (12,0%), che presenta lo scarto minimo, come evidenza la figura 9.

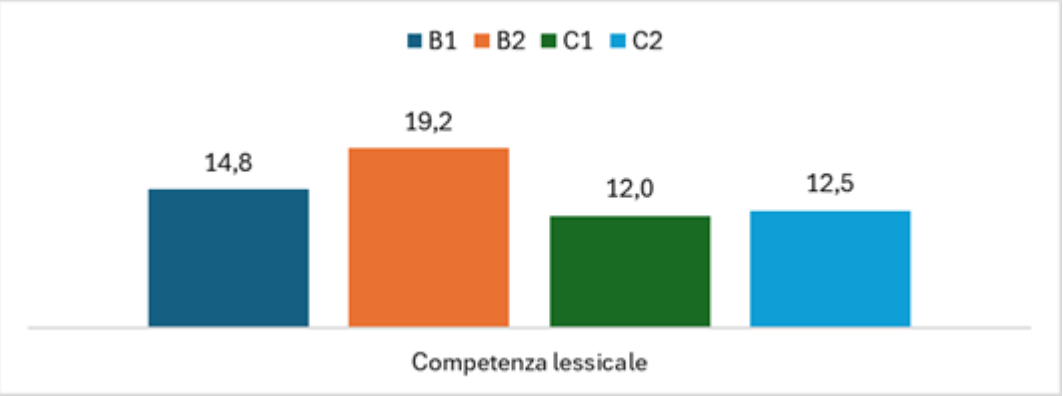


Figura 9. Media di errore (%) tra la valutazione automatica e quella umana nella competenza lessicale, in base al livello QCER della prova sostenuta.

I dati mostrano una distribuzione differenziata dello scarto medio tra la valutazione automatica e quella umana in funzione del livello di dimensione linguistica, con valori che oscillano tra il 12,0% per il C1 e il 19,2% per il B2. La variazione osservata suggerisce una relazione tra la natura delle produzioni scritte e il grado di variazione tra i due sistemi valutativi.

Il valore massimo, come osservato in precedenza, si registra al livello B2, dove lo media della variazione raggiunge il 19,2. I testi prodotti in questa fascia rispondevano al task B.1, che richiede la composizione di un testo a scelta tra due proposte, su esperienze personali o su temi di attualità. Si tratta di compiti che, pur essendo guidati, lasciano una certa libertà argomentativa e stimolano l'uso di risorse lessicali più articolate, soprattutto nei candidati che si collocano nella parte alta del livello. In questa fase dello sviluppo interlinguistico, come evidenziato da Housen, Kuiken e Vedder (2012), l'aumento della complessità testuale non è sempre accompagnato da un controllo pienamente consolidato. Il lessico impiegato tende a espandersi verso un repertorio più ampio, ma può includere scelte marcate da interferenze. L'ampiezza dello scarto rilevata in corrispondenza di questo livello potrebbe quindi riflettersi, da un lato, nella maggiore tolleranza del valutatore umano nei confronti di produzioni lessicali che, pur presentando imprecisioni, evidenziano uno sforzo di espansione comunicativa coerente con il profilo del livello B2, e dall'altro, nella minore flessibilità del modello automatico nel riconoscere e valorizzare tali tentativi, in quanto ancorato a regolarità

linguistiche apprese e a configurazioni di uso più standardizzate.

La tendenza rilevata nei livelli C1 e C2 mostra invece una maggiore stabilità: in questo caso lo scarto tra valutazioni si attesta su valori inferiori, rispettivamente di 12 per il livello C1 e 12,5 per il livello C2.

A questi livelli, il lessico appare generalmente più controllato, pertinente e adeguato al compito comunicativo. La maggiore padronanza formale e il consolidamento delle competenze pragmatiche da parte degli apprendenti sembrano facilitare l'allineamento nella valutazione tra i due sistemi anche in presenza di produzioni soggettivamente marcate o articolate. Il modello è dunque in grado di riconoscere strutture lessicali più complesse senza penalizzarle, purché coerenti e riconducibili a pattern linguistici sufficientemente frequenti nei dati di addestramento.

Il livello B1 presenta uno scarto intermedio (14,8%) In queste produzioni, la semplicità espressiva rispetto ai livelli superiori potrebbe favorire una maggiore prevedibilità del testo e quindi una maggiore coerenza tra le due modalità di assegnazione punteggio.

Nel complesso, i dati suggeriscono per che la dimensione lessicale è abbastanza sensibile alle differenze di interpretazione tra valutazione umana e automatica. Tale sensibilità può essere ricondotta alla complessità intrinseca del costrutto lessicale, che coinvolge non solo varietà e accuratezza, ma anche adeguatezza al registro, coerenza contestuale e appropriatezza comunicativa (Knoch, 2009; Fulcher & Davidson, 2007).

5.2.3 Risultati relativi alla dimensione grammaticale

La lettura dei dati relativi alla competenza grammaticale mostra uno scostamento contenuto tra la valutazione automatica e quella umana, con una media complessiva pari al 13,2%.

Rispetto alle altre dimensioni, la variabilità nei giudizi risulta quindi inferiore, suggerendo una maggiore convergenza tra i due sistemi valutativi su

questo aspetto della competenza linguistica.

Analizzando la distribuzione per livello, si osserva che il valore più elevato si registra, ancora una volta, al livello B2, con uno scarto medio del 15,6%; seguono i livelli C1 (14,3%), B1 (14%), mentre il valore minimo si riscontra al livello C2, dove lo scarto medio si attesta su 12,3%., come evidenziato nella figura 10.

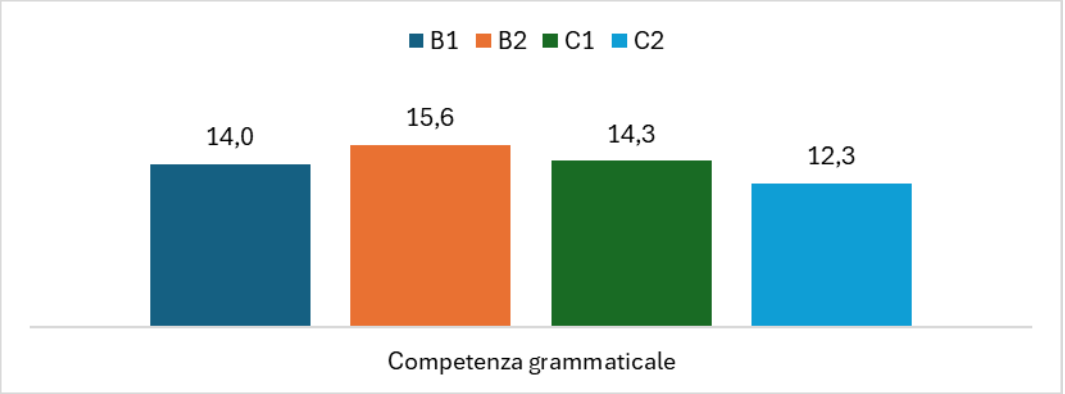


Figura 10. Media di errore (%) tra la valutazione automatica e quella umana nella competenza grammaticale, in base al livello QCER della prova sostenuta.

Tuttavia, la differenza tra i livelli risulta meno marcata rispetto ad altre dimensioni, a conferma del fatto che la grammatica rappresenta un aspetto più stabile della competenza linguistica.

La relativa omogeneità osservata nella dimensione grammaticale può essere ricondotta a una combinazione di fattori. In primo luogo, la natura intrinsecamente più oggettiva delle strutture grammaticali contribuisce a limitarne la suscettibilità a interpretazioni soggettive. Inoltre, l'errore grammaticale tende a essere facilmente identificabile e classificabile, favorendo una maggiore convergenza tra valutazione automatica e umana. Come riportato da Mizumoto et al. (2023), i modelli AES, addestrati su corpora annotati tendono a riconoscere con maggiore affidabilità errori grammaticali, confermando la maggiore stabilità di questa dimensione nel confronto tra valutazione automatica e umana.

Il valore minimo di scarto è stato riscontrato al livello C2, suggerendo che, a questo livello, la padronanza grammaticale raggiunge una soglia di stabilità tale da garantire un'elevata coerenza tra valutazione automatica e umana. I livelli B1

e C1 presentano scarti di entità contenuta, a conferma di una relativa stabilità nella rilevazione e interpretazione degli elementi grammaticali, anche in contesti in cui la competenza non è ancora pienamente sviluppata o si colloca già su un piano più avanzato. Il livello B2, invece, si distingue per un valore di scarto superiore alla media complessiva, evidenziando una fase di transizione in cui il controllo grammaticale, seppur generalmente solido, può risultare non sempre sufficientemente sistematico da produrre valutazioni perfettamente allineate tra i due metodi. Secondo il QCER, la competenza grammaticale a questo livello è caratterizzata da un buon controllo delle strutture, ed eventuali errori grammaticali a questo livello tendono a essere sporadici, non sistematici, spesso su aspetti complessi o nell'uso creativo della lingua (QCER VC, 2020). È plausibile quindi ipotizzare che quest'ultima caratteristica abbia influenzato l'interpretazione da parte del sistema automatico, generando una maggiore discrepanza valutativa rispetto al giudizio umano, in analogia a quanto già osservato per la dimensione lessicale.

La minore variabilità nella valutazione e assegnazione punteggi relativa alla dimensione grammaticale è stata più volte sottolineata in letteratura. Gli studi di Weigle (2002) evidenziano come la grammatica, rispetto ad altre dimensioni del testo scritto, generi un maggiore accordo tra valutatori, grazie alla natura più discreta, regolare e sistematica degli errori grammaticali, che li rende più facilmente rilevabili e classificabili. Anche Knoch (2009), conferma che le dimensioni grammaticali sono meno ambigue e quindi meno suscettibili a divergenze di interpretazione rispetto, ad esempio, alla dimensione della coerenza e coesione testuale; a loro volta, Attali & Burstein (2006) mostrano che i sistemi di valutazione automatica come e-rater ottengono le prestazioni più affidabili proprio nella rilevazione degli errori grammaticali, grazie alla possibilità di addestrare modelli su strutture sintattiche ricorrenti e regolari. Anche Fulcher & Davidson (2007) sottolineano che le valutazioni basate su criteri grammaticali tendono a presentare una maggiore riproducibilità, in quanto fondate su norme linguistiche stabilite, mentre altre dimensioni, come l'organizzazione del discorso o il lessico, richiedono inferenze interpretative più soggettive.

I risultati emersi, dunque, si pongono in linea con quanto riportato in letteratura, suggerendo che questa dimensione sia più facilmente modellizzabile rispetto ad altre, e che presenti una minore sensibilità alle differenze interpretative tra valutazione automatica e umana.

5.2.4 Risultati relativi alla dimensione sociolinguistica

Per quanto riguarda la competenza sociolinguistica, i dati mostrano una media complessiva di errore pari al 1

4,2%, un valore che si colloca in posizione intermedia rispetto alle altre dimensioni analizzate. L'errore medio percentuale risulta inferiore a quello registrato per il lessico (14,6%), ma superiore alla grammatica (13,2%). La Figura 10 presenta la media dell'errore percentuale calcolata mettendo a confronto i punteggi relativi a questa dimensione assegnati da ChatGPT e quelli forniti da valutatori umani alle produzioni scritte selezionate per ciascun livello QCER.

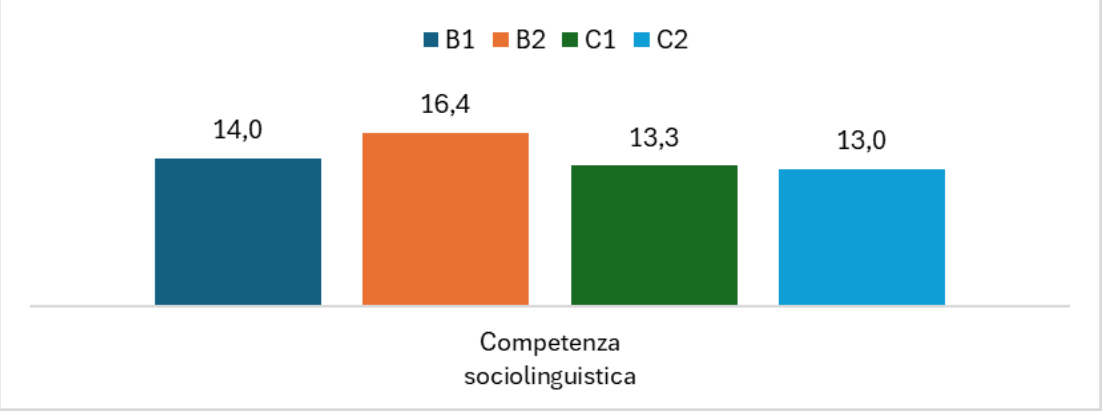


Figura 10. Media di errore (%) tra la valutazione automatica e quella umana nella competenza sociolinguistica, in base al livello QCER della prova sostenuta.

I risultati emersi evidenziano uno scarto più marcato tra valutazione automatica e valutazione umana nella competenza sociolinguistica per i livelli B1 e B2, dove la media dell'errore percentuale raggiunge rispettivamente il 14,0% e il 16,4%. I livelli C1 e C2, invece, mostrano scarti più contenuti (13,3% e 13,0%). Come già osservato nelle altre dimensioni, anche per la competenza

sociolinguistica il livello B2 si conferma il più problematico in termini di coerenza tra valutazione automatica e giudizio umano.

È importante sottolineare che tali valutazioni sono state generate da ChatGPT in risposta a un *prompt* strutturato ad hoc, in cui il modello è stato istruito a simulare il comportamento di un valutatore esperto umano, con accesso esplicito al compito, al testo prodotto dall'apprendente, e alla scala analitica di riferimento. Il modello non ha avuto accesso a esempi predefiniti o a dati di addestramento specificamente tarati sulla valutazione certificativa, ma ha operato a partire da istruzioni esplicite, organizzando la propria risposta in due sezioni distinte: una spiegazione del ragionamento valutativo e l'attribuzione di un punteggio numerico.

Alla luce di questi aspetti, lo scarto maggiore riscontrato nel livello B2 può essere interpretato come indice di una difficoltà specifica del modello nel tradurre i criteri sociolinguistici in decisioni valutative coerenti, in particolare in produzioni che presentano maggiore variabilità stilistica e scelte linguistiche meno canoniche. A questo livello, l'apprendente è in grado di identificare e interpretare codici socioculturali e sociolinguistici e di modificare consapevolmente il proprio modo di esprimersi affinché risulti adeguato alla situazione a gestire il registro, l'intenzionalità comunicativa e la pertinenza formale; tuttavia, il testo prodotto può presentare margini di instabilità, con occasionali imprecisioni nell'uso di formule, toni o convenzioni comunicative. Questo fattore potrebbe rendere più complesso per il modello GPT il riconoscimento del grado di adeguatezza sociolinguistica rispetto al contesto del compito.

Nei livelli più avanzati, al contrario, è plausibile che l'uso linguistico degli apprendenti si avvicini a modelli di riferimento più formalizzati, fattore che potrebbe aver contribuito a una valutazione automatica più allineata a quella umana.

5.2.5 Risultati relativi alla dimensione della coerenza e coesione testuale

La Figura 11. mostra la media dell'errore percentuale tra la valutazione automatica e quella umana per la dimensione della coerenza e coesione testuale. Il dato complessivo per tutti i livelli è pari al 15,8%, ma l'analisi disaggregata per livello QCER evidenzia importanti variazioni per i livelli B2 e C2. In particolare, per il livello B1 si evidenzia una media errore percentuale di 12,4, per il B2 di 18,4, per il C1 di 13 per arrivare al 19,5 registrato per il livello C2, come riportato nella figura 11.

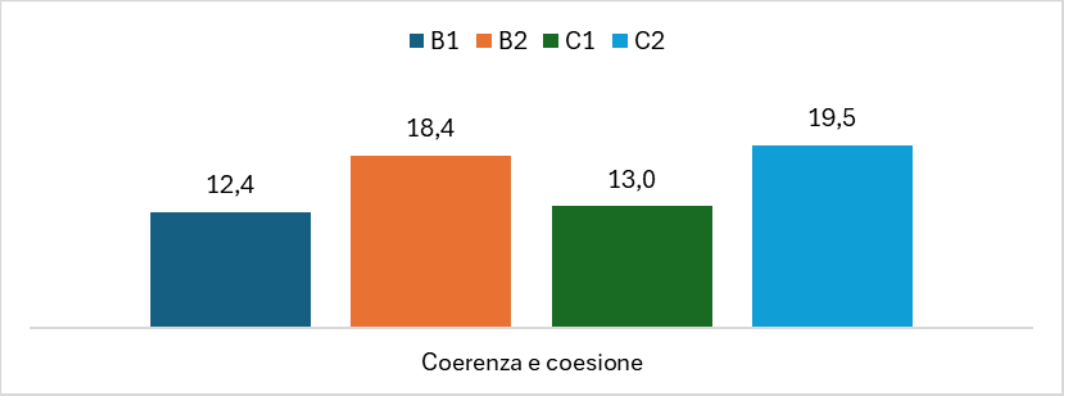


Figura 11. Media di errore (%) tra la valutazione automatica e quella umana nella competenza della coerenza e coesione, in relazione al livello QCER della prova sostenuta.

I valori estremi si registrano ai livelli B2 e C2, entrambi al di sopra della media generale, suggerendo una maggiore difficoltà del modello nel valutare correttamente questa dimensione. Al livello B1, la relativamente bassa discrepanza può essere attribuita alla tendenziale linearità e trasparenza delle strutture testuali. I testi prodotti a questo livello possono presentare coerenza e coesione prevalentemente realizzate attraverso strutture sintattiche semplici, che l'AES è in grado di intercettare con una certa affidabilità. Tale regolarità potrebbe aver ridotto l'ambiguità interpretativa da parte della valutazione automatica, allineandola maggiormente a quella del valutatore umano.

Con il passaggio al livello B2, si registra un incremento significativo della

variazione. In questa fase dell'apprendimento, gli studenti iniziano ad applicare strutture testuali più complesse, cercando di articolare il testo con connettivi logici e strategie argomentative. Tuttavia, l'efficacia di tali strategie non è sempre garantita, e il modello potrebbe penalizzare eccessivamente la mancata linearità o l'uso irregolare di elementi coesivi, non riuscendo a riconoscere tentativi validi ma non perfettamente formalizzati, mentre il valutatore umano potrebbe essere più incline a riconoscere e valorizzare lo sforzo comunicativo dell'utente, anche in presenza di occasionali deviazioni nella linearità coesiva o nella coerenza semantica.

Il livello C2 mostra la media di errore percentuale più elevata in assoluto. Questo dato può essere interpretato come indicativo del fatto che ChatGPT incontra maggiori difficoltà nel valutare produzioni linguistiche più articolate, dove coerenza e coesione sono spesso costruite attraverso meccanismi impliciti piuttosto che mediante segnali testuali espliciti. Al livello C2, infatti, i testi possono presentare un'organizzazione del discorso articolata, progressioni tematiche non lineari, digressioni strategiche e variazioni stilistiche consapevoli che si distaccano da un modello compositivo "prototipico". Il modello automatico tende generalmente a fare affidamento su indizi più superficiali e standardizzati – come la presenza di connettivi testuali, la ripetizione lessicale o la struttura sintattica – per inferire la coerenza e la coesione di un testo. Questo approccio, tuttavia, si rivela inadeguato quando il testo si sviluppa secondo logiche non convenzionali, ma pienamente coerenti dal punto di vista semantico e pragmatico. Tali considerazioni trovano riscontro diretto nello studio di Yoon, Miszoglad e Pierce (2023), che hanno analizzato il feedback fornito da ChatGPT su testi prodotti da apprendenti di L2. I risultati dello studio evidenziano che il modello tende a riconoscere solo la coerenza esplicita e la coesione di superficie, trascurando la coerenza globale del testo, la qualità della progressione logica e l'efficacia del collegamento tra le idee sul piano discorsivo. In particolare, ChatGPT ha mostrato difficoltà nell'identificare l'intento comunicativo dell'autore e nell'interpretare scelte organizzative più complesse, come le digressioni strategiche, le anafore concettuali o i riferimenti tematici non esplicitamente marcati.

Lo scarto significativo rispetto al giudizio umano a questo livello potrebbe essere motivato quindi da una lettura degli aspetti superficiali del testo da parte della valutazione automatica. Al contrario, il valutatore umano è generalmente in grado di riconoscere la coerenza globale, anche quando questa è costruita in modo implicito o attraverso strutture testuali non lineari.

6. Conclusioni

L'analisi condotta ha permesso di osservare in modo sistematico il comportamento valutativo di un sistema automatico basato su ChatGPT-4 e istruito tramite *prompt*, messo a confronto con i punteggi attribuiti da valutatori umani nell'ambito della Prova Scritta delle certificazioni CELI relative ai livelli B1, B2, C1 e C2 QCER.

Lo studio pilota, condotto su un numero limitato di testi, ha avuto una duplice valenza. Da un lato, ha permesso di testare la stabilità interna delle valutazioni generate dal modello GPT su produzioni linguistiche differenziate per livello di competenza. Dall'altro, ha fornito un primo riscontro sull'efficacia del protocollo di *prompting* elaborato per orientare la valutazione automatica secondo i criteri delle scale CELI. I risultati emersi da questa fase hanno mostrato una buona coerenza tra le valutazioni automatiche, ripetute in più cicli.

Un elemento metodologicamente rilevante riguarda proprio la struttura del *prompt* utilizzato per istruire il modello, che è stato concepito per guidare il modello a operare secondo il ruolo di un valutatore esperto, esplicitando i criteri di riferimento e richiedendo una giustificazione dei punteggi assegnati. Tale impostazione ha permesso di simulare un contesto valutativo realistico e formalizzato. Tuttavia, il *prompt* non includeva esempi concreti di valutazioni né testi già corredati di punteggi umani, affidandosi interamente alla capacità del modello di interpretare e applicare istruzioni astratte. Se da un lato ciò ha garantito l'indipendenza valutativa del sistema, dall'altro potrebbe aver rappresentato un limite, in quanto l'assenza di esempi contestuali potrebbe aver ridotto la capacità del modello di calibrare con precisione la soglia tra livelli

adiacenti, specialmente in presenza di produzioni meno prototipiche o linguisticamente marcate.

Lo studio principale, sviluppato su un campione ampio e articolato di testi autentici, ha consentito di approfondire e ampliare le osservazioni iniziali. I risultati hanno confermato la maggiore coerenza tra valutazione automatica e umana nelle dimensioni più strutturate – come la grammatica – e una minore corrispondenza nelle dimensioni più interpretative, come la coerenza e coesione testuale e la competenza lessicale. Le discrepanze non sono risultate uniformi, ma si sono concentrate soprattutto nel livello B2, tradizionalmente considerato una soglia di passaggio. Anche al livello C2 sono state rilevate alcune difficoltà nella valutazione dell'organizzazione del testo, probabilmente dovute all'elevata complessità espressiva tipica delle produzioni avanzate. Nel loro insieme, i risultati delle due fasi della ricerca mostrano che, pur in presenza di limiti, un sistema come il modello realizzato può offrire un contributo significativo alla riflessione sulla valutazione linguistica, soprattutto come strumento di osservazione, confronto e supporto al valutatore umano, il quale con occhio critico e consapevole, può avvalersi delle risposte generate per affinare il proprio giudizio, verificare la coerenza interna dei criteri applicati o individuare eventuali aree di ambiguità interpretativa.

6.1 Complementarità tra valutazione automatica e umana: verso un approccio ibrido?

I risultati emersi da questo studio suggeriscono che, più che costituire un'alternativa alla valutazione umana, i sistemi di valutazione automatica come quello testato potrebbero trovare una collocazione significativa all'interno di modelli valutativi ibridi, in cui componente umana e automatica interagiscono in modo complementare. In questo quadro, l'IA non viene intesa come sostitutiva del giudizio esperto, ma come uno strumento potenzialmente utile per affiancare, verificare, rendere più trasparente o più coerente l'assegnazione dei punteggi.

Un primo ambito di applicazione potrebbe riguardare la fase preliminare

del processo valutativo, in cui il modello, se opportunamente guidato, potrebbe fornire una stima iniziale, utile a orientare il valutatore umano o ad attivare un confronto critico. In contesti con grandi volumi di elaborati, tale utilizzo potrebbe contribuire a una gestione più efficiente delle risorse, pur mantenendo il controllo finale nelle mani del valutatore. In una prospettiva più ampia, la collaborazione tra valutazione automatica e umana potrebbe essere concepita come un processo iterativo, in cui i due agenti – umano e artificiale – contribuiscono reciprocamente al miglioramento dell'accuratezza e della coerenza del giudizio. Il modello, infatti, potrebbe essere impiegato non solo come filtro preliminare, ma anche come sistema di confronto in itinere, capace di segnalare discrepanze tra valutatori o incoerenze interne nelle attribuzioni di punteggio. In tal senso, l'IA potrebbe agire da "specchio metodologico", favorendo una riflessione metacognitiva sul processo di valutazione stesso e promuovendo un atteggiamento critico sulle decisioni interpretative assunte dall'esaminatore. Tale dinamica dialogica potrebbe inoltre stimolare pratiche di calibrazione continua, in cui le valutazioni umane e automatiche si confrontano periodicamente per affinare la coerenza del protocollo e migliorare la trasparenza complessiva del sistema valutativo.

Il contributo dell'IA potrebbe dunque assumere un ruolo di filtro o di pre-valutatore, capace di identificare tendenze, anomalie o cluster di prestazioni omogenei, facilitando così una prima organizzazione dei dati valutativi. L'uso di un sistema automatico in questa funzione non avrebbe carattere prescrittivo, ma analitico e diagnostico: il modello potrebbe fornire indicatori preliminari che aiuterebbero il valutatore a orientarsi tra grandi quantità di testi e a concentrare l'attenzione sulle produzioni che richiedono un esame più approfondito. Tale impiego potrebbe risultare particolarmente vantaggioso in contesti di certificazione su larga scala, dove la gestione del carico valutativo rappresenta un elemento critico e dove la possibilità di un'analisi preselettiva, condotta con criteri trasparenti e replicabili, potrebbe ottimizzare l'efficienza complessiva del processo, senza tuttavia compromettere l'autonomia interpretativa del giudizio umano. Inoltre, la fase preliminare di interazione con il sistema potrebbe contribuire a rendere più coerente la successiva valutazione manuale, favorendo

un approccio più sistematico e consapevole alla lettura degli elaborati.

Un secondo ambito, di natura più qualitativa, riguarda la formazione dei valutatori. I punteggi assegnati dal sistema, accompagnati da spiegazioni dettagliate del ragionamento valutativo, potrebbero essere impiegati come stimolo per riflettere sui criteri applicati, confrontare approcci valutativi e individuare incoerenze nell'attribuzione dei punteggi. In questo senso, il modello potrebbe contribuire a rendere espliciti i processi inferenziali che, nella pratica valutativa umana, tendono spesso a rimanere impliciti. Questa funzione metariflessiva potrebbe costituire un valore aggiunto non soltanto in termini di trasparenza e coerenza inter-valutatore, ma anche come occasione di sviluppo professionale per i valutatori stessi, che, confrontandosi con la logica valutativa del sistema, sarebbero indotti a esplicitare e a problematizzare i propri criteri di giudizio. In prospettiva, un'interazione di questo tipo potrebbe promuovere una cultura valutativa più consapevole, in cui la componente umana e quella automatica dialoghino in modo complementare, integrando intuizione esperta e rigore analitico.

Proprio in quest'ottica di dialogo si colloca una riflessione sul ruolo del valutatore all'interno di un approccio ibrido. Sebbene il valutatore esperto disponga di strumenti critici consolidati per esercitare un giudizio autonomo, l'introduzione di un supporto automatico richiede un'attenta definizione della sua funzione e del suo grado di influenza sul processo valutativo. Anche in contesti professionali, l'esposizione a un output generato dal modello potrebbe infatti orientare, in modo più o meno consapevole, la percezione iniziale della qualità testuale o la lettura di aspetti marginali ma significativi. Per questa ragione, l'integrazione dell'IA nei contesti valutativi dovrebbe essere accompagnata da una riflessione metodologica condivisa, volta a chiarire la ripartizione delle competenze tra componente umana e componente automatica, e a salvaguardare il ruolo del valutatore come responsabile ultimo del giudizio, capace di interagire con l'*output* del sistema in modo critico, argomentato e consapevole.

In tale prospettiva, l'adozione di modelli ibridi di valutazione potrebbe implicare una ridefinizione della natura stessa del giudizio valutativo, che non

verrebbe più concepito come atto individuale e autonomo, ma come esito di un processo dialogico e multilivellare. L'assegnazione del punteggio potrebbe dunque derivare dall'interazione tra diversi piani inferenziali: da un lato quello umano, fondato su competenze interpretative, contestuali e disciplinari; dall'altro quello algoritmico, basato su regolarità statistiche e su una rappresentazione formalizzata del linguaggio e dei criteri di qualità testuale. Una tale convergenza tra inferenze di natura differente porrebbe, in ogni caso, questioni di ordine epistemologico riguardanti i processi attraverso cui l'evidenza valutativa viene costruita e condivisa, nonché le forme di legittimazione che si riconoscono alle decisioni ibride, risultanti dalla cooperazione tra umani e sistemi intelligenti.

Dal punto di vista operativo, l'integrazione dell'IA nei processi di valutazione linguistica dovrebbe fondarsi su protocolli di *governance* chiari, in grado di delimitare il campo d'azione del sistema e di garantire la tracciabilità del processo decisionale. In questo quadro, un approccio *human-in-the-loop* — ovvero un modello di interazione in cui l'essere umano rimane costantemente coinvolto nel ciclo valutativo, supervisionando, confermando o rivedendo le proposte generate dai sistemi automatici — potrebbe rappresentare una soluzione equilibrata. Tale configurazione non solo preserverebbe la centralità del giudizio umano, ma consentirebbe di valorizzare il contributo del sistema come strumento di analisi e di confronto, lasciando la piena responsabilità del processo decisionale al valutatore.

L'integrazione di sistemi automatici richiede, tuttavia, anche una riflessione di natura etica e sociologica. L'introduzione dell'IA nei contesti valutativi potrebbe infatti incidere sulla percezione di equità, imparzialità e legittimità del giudizio, generando tanto aspettative di oggettività quanto potenziali forme di diffidenza. Se, da un lato, l'IA potrebbe contribuire a ridurre la variabilità soggettiva e a garantire una maggiore coerenza tra valutatori, dall'altro lato essa potrebbe introdurre nuove forme di opacità algoritmica o fenomeni di eccessiva deferenza nei confronti dell'automatismo (*automation bias*). In quest'ottica, sarebbe auspicabile promuovere un equilibrio dinamico tra l'autorevolezza del giudizio esperto e la capacità standardizzante della macchina, accompagnato da percorsi

formativi orientati alla consapevolezza critica e alla comprensione dei limiti e delle potenzialità dei sistemi impiegati.

In prospettiva, la complementarità tra valutazione automatica e umana potrebbe aprire la strada a modelli realmente collaborativi, in cui l'IA non si configurerebbe come un sostituto del valutatore, ma come un interlocutore tecnico, capace di supportarne il lavoro attraverso un'analisi più sistematica, trasparente e replicabile. Questo approccio, tuttavia, richiede ancora molte riflessioni di natura metodologica, etica e pedagogica, soprattutto in ambito certificativo, dove l'affidabilità, l'equità e la trasparenza restano requisiti imprescindibili.

Il presente studio, pur nei suoi limiti, ha voluto offrire un primo contributo in questa direzione, sottolineando la necessità di procedere con cautela ma anche con apertura nei confronti di strumenti che, se integrati in modo critico e consapevole, potrebbero arricchire le pratiche di valutazione linguistica.

6.2 Sviluppi futuri

I risultati emersi da questo primo studio esplorativo indicano che il modello testato, pur nella sua attuale configurazione, è in grado di produrre valutazioni plausibili e talvolta coerenti con quelle umane, ma anche che persistono aree di criticità. Queste evidenze, pur parziali, suggeriscono la necessità di un approfondimento sistematico delle potenzialità e dei limiti dell'IA applicata alla valutazione linguistica, al fine di comprendere in che misura tali strumenti possano essere integrati in modo eticamente e metodologicamente sostenibile nei processi di certificazione e di formazione linguistica.

Uno sviluppo potenzialmente rilevante potrebbe consistere nell'ampliamento del campione di testi analizzati, sia in termini di quantità che di distribuzione tra diverse sessioni d'esame, tracce e profili di candidati. Un campione più esteso consentirebbe di consolidare la significatività dei risultati, aumentando la robustezza statistica delle osservazioni e permettendo un'analisi più articolata del comportamento del sistema AES l'intero spettro dei livelli QCER.

Tale approfondimento permetterebbe di comprendere meglio quali aspetti delle scale analitiche risultano più agevolmente modellizzabili e quali, invece, pongano ancora sfide significative. In parallelo, si potrebbe approfondire in modo più mirato il comportamento del modello all'interno delle singole dimensioni delle scale CELI, indagando, ad esempio, se e come il sistema riesca a distinguere tra i diversi livelli di accuratezza, varietà o pertinenza richiesti in ciascun intervallo di punteggio. Una lettura più granulare degli output, condotta su sottogruppi definiti, potrebbe contribuire a chiarire quali aspetti dei criteri valutativi vengano recepiti in modo più stabile dal modello e quali, invece, risultino più esposti a variabilità interpretativa.

Un secondo ambito di sviluppo riguarda la progettazione e calibrazione dei *prompt*. Il presente studio ha fatto ricorso a un'impostazione descrittiva, priva di esempi, proprio per osservare la risposta del modello alle istruzioni fornite. In studi successivi, si potrebbe sperimentare l'effetto di *prompt* più strutturati o arricchiti da esempi guida, mantenendo sempre l'aderenza ai criteri forniti dalle scale di livello e punteggio, per verificare se un simile approccio possa aiutare a ridurre gli scarti sistematici senza compromettere la generalizzabilità del protocollo. Un'ulteriore prospettiva di ricerca potrebbe dunque prevedere la sperimentazione di strategie di *prompt engineering* basate su tecniche di *few-shot* o *chain-of-thought prompting*, che consentirebbero di osservare in che misura l'esplicitazione graduale del ragionamento valutativo possa incrementare la trasparenza e la coerenza del giudizio generato. Parallelamente, sarebbe opportuno indagare il potenziale uso di feedback iterativi, in cui il modello, opportunamente istruito, possa riformulare la propria valutazione a seguito di un confronto con punteggi umani, simulando così una forma di apprendimento regolato o di allineamento valutativo. Queste sperimentazioni, se condotte in modo controllato, potrebbero aprire nuove strade per l'addestramento mirato di sistemi dedicati alla valutazione linguistica.

Inoltre, nonostante il presente lavoro abbia adottato GPT-4 come modello di riferimento, è importante ricordare che il panorama degli LLM è in continua evoluzione. Modelli successivi o anche alternativi, come ad esempio Claude,

Gemini o LLaMA, che si distinguono per architetture e modalità di apprendimento differenti, potrebbero essere oggetto di studi comparativi, con l'obiettivo di valutare quale modello offra la maggiore aderenza ai criteri delle scale di livello e punteggi adottate per la prove scritte CELI. L'analisi delle divergenze intermodello, ad esempio, potrebbe fornire indicazioni preziose sul grado di convergenza semantica dei diversi sistemi e contribuire alla definizione di standard di affidabilità intermodello, concettualmente analoghi all'affidabilità inter-valutatore nella valutazione umana. In una prospettiva più sperimentale, si potrebbe anche ipotizzare l'uso combinato di più modelli, per esplorare l'eventuale complementarità tra diverse IA nell'assegnazione dei punteggi.

Infine, potrebbe essere utile affiancare agli aspetti tecnici e quantitativi indagini di tipo qualitativo rivolte a candidati e valutatori, che potrebbero contribuire a esplorare come questi attori interpretano e valutano la presenza di un sistema automatico all'interno del processo di attribuzione del punteggio. Comprendere i diversi atteggiamenti nei confronti dell'AES, tra aspettative e riserve, potrebbe rappresentare un passaggio fondamentale per valutare la percezione di questi strumenti nei contesti istituzionali e per interrogarsi sulla loro accettabilità, utilità percepita e potenziale impatto sulla trasparenza del processo valutativo. In questa direzione, un'indagine qualitativa sistematica potrebbe includere strumenti di raccolta dati come interviste semi-strutturate, focus group o questionari riflessivi, con l'obiettivo di far emergere le dimensioni emotive, etiche e professionali che accompagnano l'introduzione dell'IA nella valutazione linguistica. L'analisi delle rappresentazioni e delle resistenze dei valutatori, così come delle percezioni dei candidati, costituirebbe un elemento imprescindibile per comprendere la sostenibilità sociale e pedagogica di un eventuale modello ibrido di valutazione.

Nel complesso, gli sviluppi delineati si muovono verso la costruzione di un paradigma valutativo dialogico, in cui l'IA non sostituisce il giudizio umano, ma lo affianca in un processo di reciproco affinamento. La sfida principale consiste dunque nel definire un equilibrio tra automazione e interpretazione, capace di preservare il valore del giudizio del valutatore esperto e, al tempo stesso, di

valorizzare il potenziale analitico dei sistemi di IA.

In tale quadro, la ricerca futura potrebbe dunque orientarsi non solo alla misurazione delle prestazioni del modello, ma anche alla comprensione del suo impatto epistemologico, pedagogico e istituzionale sul concetto stesso di valutazione linguistica.

Bibliografia e sitografia

Alderson, J. C., Clapham, C., & Wall, D. (1995). Language test construction and evaluation. Cambridge University Press.

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 715–725). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1068.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater®. Version 2.0. Educational Testing Service.

Australian Education Union (2017). *NAPLAN Online: Rage against the machine*. https://aeunt.org.au/naplan-online-rage-against-the-machine.

Bachman, L. F. (1990). Fundamental considerations in language testing. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford: Oxford University Press.

Balboni, P. E. (2015). Le sfide di Babele. Insegnare le lingue nelle società complesse (3ª ed.). Torino: UTET Università.

Barkaoui, K. (2007). Participants, texts, and processes in second language writing assessment: A narrative review of the literature. The Canadian Modern Language Review 64, 97-132.

Barni, M. (2023). Valutare le competenze nelle L2: Teorie, metodi, strumenti, politiche linguistiche. Roma: Carocci.

Binns, R. (2018). Algorithmic accountability and public reason. Philosophy and Technology, 31(4), 543-556. https://doi.org/10.1007/s13347-017-0263-5.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Bennett, R. E., & Zhang, M. (2015). Validity and automated scoring. In F. Drasgow (Ed.), Technology and testing: Improving educational and psychological measurement (pp. 142–173). Routledge. https://doi.org/10.4324/9781315871493-8.

Bereiter, C. (2003). Automated essay scoring and the problem of writing. In M. D. Shermis & J. Burstein (Eds.), Automated essay scoring: A cross- disciplinary perspective (pp. 137–147). Lawrence Erlbaum Associates.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), Advances in Neural Information Processing Systems (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Burstein, J. (2003). The development and use of automated essay scoring systems. In M. D. Shermis & J. C. Burstein (Eds.), Automated essay scoring: A cross- disciplinary perspective (pp. 3-24). Lawrence Erlbaum Associates.

Burstein, J., & Chodorow, M. (1999). Automated essay scoring for learning and assessment. In Proceedings of the 16th International Conference on Computational Linguistics (COLING'96) (pp. 254-258). Association for Computational Linguistics.

Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterion: Online essay evaluation: An application for automated evaluation of student essays. In Proceedings of the 15th Annual Conference on Innovative Applications of Artificial Intelligence. Acapulco, Mexico.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998). Computer analysis of essays. Paper presented at the NCME Symposium on Automated Scoring, Montreal, Canada.

Burstein, J., & Marcu, D. (2000). Benefits of modularity in an automated essay scoring system. In R. Zajac (Ed.), Proceedings of the COLING-2000 Workshop on Using Toolsets and Architectures To Build NLP Systems (pp. 44–50). International Committee on Computational Linguistics.

Canale, M. (1983). On some dimensions of language proficiency. In J. W. Oller Jr. (Ed.), Issues in Language Testing Research (pp. 333–342). Rowley, MA: Newbury House.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics, 1(1), 1–47.

Carroll, J. B. (1961). Fundamental considerations in testing English proficiency of foreign students. In Testing the English proficiency of foreign students (pp. 30–40). Washington, DC: Center for Applied Linguistics.

Casani, E. (2020). Valutare la competenza morfosintattica in italiano L2. In Valutare la competenza linguistica in italiano L2: Un'analisi corpus-based dei livelli del QCER (pp. 15–46). Firenze: Cesati.

Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. Issues in Applied Linguistics, 6(2), 5–35.

Chapelle, C. A., & Brindley, G. (2002). Assessment. In N. Schmitt (Ed.), An Introduction to Applied Linguistics (pp. 267-288). Oxford University Press.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. Language Testing, 2(2), 155–163. https://doi.org/10.1177/026553228500200204.

Chung, G. K. W. K., & O'Neil, H. F. (1997). Methodological approaches to online scoring of essays: An investigation of reliability, validity, and gender bias. Journal of Educational Computing Research, 17(3).

Cinganotto, L., & Montanucci, G. (2024a). L'intelligenza artificiale per l'apprendimento dell'italiano L2/LS. Risultati preliminari di una sperimentazione. Status Quaestionis, 26, 617–635. https://doi.org/10.13133/2239-1983/18792.

Cinganotto, L., & Montanucci, G. (2024b). Exploring the integration of artificial intelligence in online language learning: A case example on Italian as a foreign language. In S. Greco & L. Cinganotto (Eds.), Innovation in education for deeper learning (pp. 37–54). INDIRE-IUL Press.

Cinganotto, L., Sbardella, T., & Montanucci, G. (2024). Dagli algoritmi alle competenze linguistiche: Il ruolo dell'intelligenza artificiale nell'educazione linguistica online. The Journal of Language and Teaching Technology, 6, 63–74.

Coonan, C. M., Bier, A., & Ballarin, E. (a cura di). (2018). La didattica delle lingue nel nuovo millennio. Le sfide dell'internazionalizzazione. Venezia: Edizioni Ca' Foscari.

Corder, S. P. (1981). Error analysis and interlanguage. Oxford University Press.

Cotos, E. (2018). Automated writing evaluation. In J. I. Liontas (Ed.), The TESOL Encyclopedia of English Language Teaching (pp. 1-6). Wiley. https://doi.org/10.1002/9781118784235.eelt0391.

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press.

Council of Europe. (2020). Common European Framework of Reference for Languages: Companion Volume with New Descriptors. Strasbourg: Council of Europe.

Council of Europe. (2011). Manual for language test development and examining: For use with the CEFR. Strasbourg: Council of Europe Publishing.

Cumming, A. (2001). Learning to write in a second language: Two decades of research. International Journal of English Studies, 1(2), 1–23.

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided Error Analysis. System, 26(2), 163–174.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). Dictionary of Language Testing. Cambridge: Cambridge University Press.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics. DOI:10.18653/v1/N19- 1423.

Dikli, S. (2006). *An overview of automated scoring of essays. The Journal of Technology, Learning, and Assessment, 5*(1). Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College. https://ejournals.bc.edu/index.php/jtla/article/view/1640/1489.

Eckes, T. (2012). Operational rater training and rating quality: The impact of rating experience and rating script. Language Testing, 29(2), 165–184.

Elliot, S. (2003). Intellimetric™: From here to validity. In M.D. Shermis & J.C. Burstein (Eds.), Automatic essay scoring: A cross-disciplinary perspective (pp. 71-86). Mahwah, NJ: Lawrence Erlbaum Associates.

Ellis, R. (2008). The Study of Second Language Acquisition (2ª ed.). Oxford: Oxford University Press.

Ellis, R. (2009). Implicit and explicit learning, knowledge and instruction. In R. Ellis, Implicit and explicit knowledge in second language learning, testing and teaching (pp. 3–26). Bristol, UK: Multilingual Matters.

Faseeh, M., Jaleel, A., Iqbal, N., Ghani, A., Abdusalomov, A., Mehmood, A., & Cho, Y.-I. (2024). Hybrid approach to automated essay scoring: Integrating deep learning embeddings with handcrafted linguistic features for improved accuracy. Mathematics, 12(21), 3416. https://doi.org/10.3390/math12213416.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1(2).

Fulcher, G. (s.d.). What is language testing? Consultato il 18 giugno 2023, URL: https://languagetesting.info/whatis/lt.html.

Fulcher, G., & Davidson, F. (2007). Language Testing and Assessment: An Advanced Resource Book. Routledge.

Gallina, F. (2022). Osservare e valutare la competenza lessicale in italiano L2. Milano: Franco Angeli.

Geçkin, V., & Aydın, E. (2023). Assessing second-language academic writing: AI vs. human raters. Journal of Educational Technology & Online Learning, 6(4), 1096–1108. https://doi.org/10.31681/jetol.1336599.

Granger, S. (1998). The Computer Learner Corpus: A Versatile New Source of Data for SLA Research. In S. Granger (Ed.), Learner English on Computer (pp. 3–18). Longman.

Grego Bolli, G., & Pelliccia, F. (2005). CELI. Breve guida ai certificati di italiano L2. Perugia: Guerra Edizioni.

Gross, M. (1997). The construction of local grammars. In E. Roche & Y. Schabes (Eds.), Finite-State Language Processing (pp. 329–354). MIT Press.

Hamp-Lyons, L. (2001). Ethics, fairness and responsibility in standardized testing. System, 29(4).

Hamp-Lyons, L. (2001). Fourth generation writing assessment. In T. Silva and P. K. Matsuda (Eds.), On Second Language Writing (pp. 117–125). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hamp-Lyons, L. (1991). Assessing Second Language Writing in Academic Contexts. Norwood, NJ: Ablex.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162. https://doi.org/10.1080/00437956.1954.11659520.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), Dimensions of L2 performance and proficiency: Investigating complexity, accuracy and fluency in SLA (pp. 1–20). Amsterdam: John Benjamins Publishing Company.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. The Modern Language Journal, 91(4), 663–667.

Hyland, K. (2003). Second Language Writing. Cambridge: Cambridge University Press.

Jafrancesco, E., & La Grassa, M. (Eds.). (2021). Competenza lessicale e apprendimento dell'italiano L2. Roma: Aracne Editrice.

Imperial, J. M., Forey, G., & Madabushi, H. T. (2024). STANDARDIZE: Aligning language models with expert-defined standards for content generation. arXiv. https://arxiv.org/abs/2402.12593.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274.

Kim, H., Baghestani, S., Yin, S., Karatay, Y., Kurt, S., Beck, J., & Karatay, L. (2024). ChatGPT for writing evaluation: Examining the accuracy and reliability of AI-generated scores compared to human raters. In C. A. Chappelle, G. H. Beckett, & J. Ranalli (Eds.), Exploring artificial intelligence in applied linguistics (pp. 73–95). Iowa State University Digital Press. https://doi.org/10.31274/isudp.2024.154.06.

Lado, R. (1957). Linguistics across cultures: Applied linguistics for language teachers. Ann Arbor, MI: University of Michigan Press.

Lado, R. (1961). Language Testing: The Construction and Use of Foreign Language Tests. London: Longman.

Landauer, T. K., Laham, D. D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Lawrence Erlbaum Associates.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310.

Lewis, M. (1993). The Lexical Approach: The State of ELT and a Way Forward. Hove: Language Teaching Publications.

Li, M. (2024). Leveraging ChatGPT for second language writing feedback and assessment. International Journal of Computer-Assisted Language Learning and Teaching, 14(1), 1–17. https://doi.org/10.4018/IJCALLT.360382.

Little, D. (2006). The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. Strasbourg: Council of Europe.

Lo Cascio, V. (2012). Lingua e comunicazione. Torino: UTET.

Long, M. H. (2015). Second language acquisition and task-based language teaching. Malden, MA: Wiley-Blackwell.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? Language Testing, 19(3), 246–276. https://doi.org/10.1191/0265532202lt230oa.

Maggini, L. (2021). Lessico, uso e apprendimento nell'approccio lessicale. In E. Jafrancesco & M. La Grassa (Eds.), *Competenza lessicale e apprendimento dell'italiano L2* (pp. 61–83). Aracne Editrice.

Malik, A., Mayhew, S., Piech, C., & Bicknell, K. (2024). From Tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 15670–15693). Association for Computational Linguistics.

Mansour, W. A., Albatarni, S., Eltanbouly, S., & Elsayed, T. (2024). Can large language models automatically score proficiency of written essays? In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC- COLING 2024) (pp. 2777–2786). Torino, Italia: ELRA & ICCL.

McNamara, T. (2000). Language Testing. Oxford: Oxford University Press.

Myers, M. (2003). What can computers and AES contribute to a K–12 writing program? In M. Shermis & J. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective (pp. 3–20). Mahwah, NJ: Lawrence Erlbaum Associates.

National Council of Teachers of English. (2013). *Machine scoring fails the test. NCTE Position Statement on Machine Scoring.* National Council of Teachers of English. URL: https://ncte.org/statement/machine_scoring/.

North, B. (2000). The development of a common framework scale of language proficiency. New York: Peter Lang.

Norris, J. M. (2009). Task-based teaching and testing. In M. H. Long & C.J. Doughty (Eds.), The Handbook of Language Teaching (pp. 578–594). Oxford: Wiley-Blackwell.

OpenAI. (2020). Introducing the OpenAI API. Consultato il 18 aprile 2022. URL: https://openai.com/blog/openai-api.

OpenAI. (2023). GPT-4 Technical Report. Consultato il 30 dicembre 2023. URL: https://openai.com/research/gpt-4.

OpenAI. (2024). OpenAI API documentation: Overview. Consultato il 12 marzo 2024. URL: https://platform.openai.com/docs/overview.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems, 35, 27730–27744.

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*(1), 100234.

Page, E. B. (1966). The imminence of grading essays by computer. Phi Delta Kappan, 47(5), 238-243.

Page, E.B. (1968). The use of computer in analyzing student essays. International Review in Education, 14, 210-225.

Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective (pp. 43-54) Mahwah, NJ: Lawrence Erlbaum Associates.

Perelman, L. C. (2013). Critique of Mark D. Shermis & Ben Hamner, "Contrasting state-of-the-art automated scoring of essays: Analysis." Journal of Writing Assessment, 6(1).

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. OpenAI Technical Report.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI.

Rudner, L. M., & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. Practical Assessment, Research & Evaluation, 7(26).

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. Journal of Technology, Learning, and Assessment, 1(2).

Selinker, L. (1972). *Interlanguage. International Review of Applied Linguistics in Language Teaching, 10*(1–4), 209–241. https://doi.org/10.1515/iral.1972.10.1-4.209.

Serragiotto, G. (2016). La valutazione degli apprendimenti linguistici. Roma: Bonacci Editore.

Shermis, M. & Barrera, F. (2002). Exit assessments: Evaluating writing ability through Automated Essay Scoring (ERIC document reproduction service no ED 464 950).

Shermis, M. D., & Burstein, J. (Eds.). (2003). Automated essay scoring: A cross-disciplinary perspective. Lawrence Erlbaum Associates.

Shermis, M. D., Raymat, M. V., & Barrera, F. (2003). Assessing writing through the curriculum with Automated Essay Scoring (ERIC document reproduction service no ED 477 929).

Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? Language Testing, 14(3), 340–349.

Shohamy, E. (2001a). The power of tests: A critical perspective on the uses of language tests. Harlow, UK: Longman/Pearson Education.

Shohamy, E. (2001b). Democratic assessment as an alternative. Language Testing, 18(4), 373–391.

Sireci, S. G., & Rizavi, S. (1999). Comparing computerized and human scoring of WritePlacer essays (Laboratory of Psychometric and Evaluative Research Report No. 354). School of Education, University of Massachusetts Amherst. https://files.eric.ed.gov/fulltext/ED463324.pdf

Scarino, A., & Liddicoat, A. J. (2009). Teaching and Learning Languages: A Guide. Curriculum Corporation.

Selinker, L. (1972). *Interlanguage. International Review of Applied Linguistics in Language Teaching, 10*(1–4), 209–241. https://doi.org/10.1515/iral.1972.10.1-4.209.

Shermis, M. & Barrera, F. (2002). Exit assessments: Evaluating writing ability through Automated Essay Scoring (ERIC document reproduction service no ED 464 950). URL: ERIC - ED464950 - Exit Assessments: Evaluating Writing Ability through Automated Essay Scoring. Poláková

Shermis, M. D. & Burstein, J. (2003). Automated Essay Scoring: A cross disciplinary perspective. Mahwah, NJ: Lawrence Erlbaum Associates.

Shermis, M. D., Raymat, M. V., & Barrera, F. (2003). Assessing writing through the curriculum with Automated Essay Scoring (ERIC document reproduction service no ED 477 929). URL: https://files.eric.ed.gov/fulltext/ED477929.pdf

Shohamy, E. (2001). The Power of Tests: A Critical Perspective on the Uses of Language Tests. Harlow: Pearson Education.

Spina, S., Fioravanti, I., Forti, L., Santucci, V., Scerra, A., & Zanda, F. (2022). Il Corpus CELI: una nuova risorsa per studiare l'acquisizione dell'italiano L2. "Italiano a stranieri. Rivista online di didattica dell'italiano a stranieri", 2(2). https://doi.org/10.54103/iam-2022-18161.

Spina, S., Fioravanti, I., Forti, L., & Zanda, F. (2024). The CELI corpus: Design and linguistic annotation of a new online learner corpus. Second Language Research, 40(2), 457-477. https://doi.org/10.1177/02676583231176370.

Spinelli B. e Parizzi F. (2010) Profilo della lingua italiana. Livelli di riferimento del QCER A1, A2, B1, B2, Firenze, La Nuova Italia/ RCS Libri.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. Learning and Instruction, 91, 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 1882–1891). Association for Computational Linguistics.
https://doi.org/10.18653/v1/D16-1193.

Taylor, L., & Galaczi, E. D. (2011). Scoring validity. In L. Taylor (Ed.), Examining speaking: Research and practice in assessing second language speaking (pp. 171–233). Cambridge: Cambridge University Press.

Turner, C. E., & Purpura, J. E. (2016). Learning-oriented assessment in language classrooms. In D. Tsagari & J. Banerjee (Eds.), Handbook of second language assessment (pp. 255–274). Berlin: De Gruyter Mouton.

Uto, M. (2021). A review of deep-neural automated essay scoring models. Behaviormetrika, 48, 459-484. https://doi.org/10.1007/s41237-021-00142-y.

Vantage Learning. (2000). A true score study of IntelliMetric accuracy for holistic and dimensional scoring of college entry-level writing program (RB-407). Newtown, PA: Vantage Learning.

Vantage Learning. (2001). About IntelliMetric (PB-540). Newtown, PA: Vantage Learning.

Vantage Learning. (2003). How does IntelliMetric score essay responses? (RB-929). Newtown, PA: Vantage Learning.

Vedovelli, M. (2002). Guida all'italiano per stranieri: La prospettiva del Quadro comune europeo per le lingue. Roma: Carocci.

Vedovelli, M. (2011). Storia linguistica dell'emigrazione italiana nel mondo (2ª ed.). Roma: Carocci.

Vedovelli, M., & Casini, S. (2016). Che cos'è la linguistica educativa. Carocci.

Villarini, A. (2021). Lo sviluppo della competenza lessicale: incursioni tra le ipotesi teoriche e le applicazioni glottodidattiche. In E. Jafrancesco & M. La Grassa (a cura di), Competenza lessicale e apprendimento dell'italiano L2 (pp. 51–63). Firenze University Press. https://doi.org/10.36253/978-88-5518-403-8.05.

Wang, J., & Brown, M. S. (2007). Automated Essay Scoring Versus Human Scoring: A Comparative Study. The Journal of Technology, Learning and Assessment, 6(2).

Weigle, S. C. (2002). Assessing Writing. Cambridge University Press.

Wei, P., Wang, X., & Dong, H. (2023). The impact of automated writing evaluation on second language writing skills of Chinese EFL learners: A randomized controlled trial. Frontiers in Psychology, 14, 1-12. https://doi.org/10.3389/fpsyg.2023.1249991.

Weir, C. J. (2005). Language Testing and Validation: An Evidence-Based Approach. Basingstoke: Palgrave Macmillan.

Woodworth, M., & Barkaoui, K. (2020). Perspectives on using automated writing evaluation systems to provide written corrective feedback in the ESL classroom. TESL Canada Journal, 37(2), 234–247. https://doi.org/10.18806/tesl.v37i2.1340.

Yancey, K. P., LaFlair, G. T., Verardi, A. R., & Burstein, J. (2023). Rating short L2 essays on the CEFR scale with GPT-4. In Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics.

Yoon, S.-Y., Miszoglad, E., & Pierce, L. R. (2023). Evaluation of ChatGPT feedback on ELL writers' coherence and cohesion. arXiv preprint arXiv:2310.06505. https://arxiv.org/abs/2310.06505.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. Communications of the ACM, 64(3), 107–115. https://doi.org/10.1145/3446776.

Appendice A - Tabelle riassuntive dei dati raccolti

Legenda

- **Dimensione**: categoria analitica utilizzata nella valutazione delle produzioni scritte. Le dimensioni considerate includono:
    - Competenza lessicale
    - Competenza grammaticale
    - Competenza sociolinguistica
    - Coerenza e coesione
- **QCER**: il livello linguistico QCER per il quale il candidato ha sostenuto l'esame.
- **ID**: numero identificativo assegnato a ciascun testo prodotto.
- **Traccia**: numero identificativo della traccia.
- **Punt. VU** (Punteggio Valutatore Umano): punteggio assegnato da valutatori esperti secondo le scale di competenza e punteggi delle prove CELI.
- **Punt. GPT**: punteggio generato automaticamente dal sistema di valutazione automatica.
- **Punt. Max**: punteggio massimo previsto.
- **Err. Norm.** (Errore Normalizzato): deviazione normalizzata tra punteggio umano e automatico, rispetto al punteggio massimo.

Dimensione	QCER	ID	Traccia	Punt. VU	Punt. GPT	Punt. Max	Err. Norm.
Competenza lessicale	B1	285	1	2	2	5	0
Competenza lessicale	B1	42	1	2	2	5	0
Competenza lessicale	B1	724	1	3	3	5	0
Competenza lessicale	B1	853	1	3	3	5	0
Competenza lessicale	B1	54	1	4	3	5	20
Competenza lessicale	B1	138	1	4	4	5	0
Competenza lessicale	B1	137	1	5	3	5	40

Competenza lessicale	B1	224	1	5	3	5	40
Competenza lessicale	B1	547	1	5	4	5	20
Competenza lessicale	B1	803	1	5	3	5	40
Competenza lessicale	B1	1332	9	2	4	5	40
Competenza lessicale	B1	1527	9	2	3	5	20
Competenza lessicale	B1	622	9	3	3	5	0
Competenza lessicale	B1	1456	9	3	3	5	0
Competenza lessicale	B1	494	9	4	4	5	0
Competenza lessicale	B1	1459	9	4	3	5	20
Competenza lessicale	B1	637	9	5	3	5	40
Competenza lessicale	B1	1315	9	5	4	5	20
Competenza lessicale	B1	1571	9	5	4	5	20
Competenza lessicale	B1	1627	9	5	4	5	20
Competenza lessicale	B1	1777	17	2	3	5	20
Competenza lessicale	B1	1836	17	2	3	5	20
Competenza lessicale	B1	1825	17	3	3	5	0
Competenza lessicale	B1	1837	17	3	4	5	20
Competenza lessicale	B1	1981	17	4	4	5	0

Competenza lessicale	B1	1983	17	4	4	5	0
Competenza lessicale	B1	1989	17	5	4	5	20
Competenza lessicale	B1	2019	17	5	5	5	0
Competenza lessicale	B1	2021	17	5	4	5	20
Competenza lessicale	B1	2037	17	5	4	5	20
Competenza lessicale	B1	2170	25	1	3	5	40
Competenza lessicale	B1	2535	25	2	3	5	20
Competenza lessicale	B1	2546	25	2	3	5	20
Competenza lessicale	B1	2127	25	3	3	5	0
Competenza lessicale	B1	2143	25	3	3	5	0
Competenza lessicale	B1	2550	25	4	4	5	0
Competenza lessicale	B1	2555	25	4	4	5	0
Competenza lessicale	B1	2131	25	5	4	5	20
Competenza lessicale	B1	2137	25	5	3	5	40
Competenza lessicale	B1	2138	25	5	4	5	20
Competenza lessicale	B1	2571	33	1	2	5	20
Competenza lessicale	B1	2455	33	2	3	5	20
Competenza lessicale	B1	2463	33	2	3	5	20

Competenza lessicale	B1	2581	33	3	3	5	0
Competenza lessicale	B1	2589	33	3	3	5	0
Competenza lessicale	B1	2603	33	4	3	5	20
Competenza	B1	2606	33	4	4	5	0
lessicale							
Competenza lessicale	B1	2585	33	5	4	5	20
Competenza lessicale	B1	2593	33	5	4	5	20
Competenza lessicale	B1	2598	33	5	5	5	0

Dimensione	**QCER**	**ID**	**Traccia**	**Punt. VU**	**Punt. GPT**	**Punt. Max**	**Err. Norm.**
Competenza grammaticale	B1	2742	39	1	2	5	20
Competenza grammaticale	B1	2688	39	2	3	5	20
Competenza grammaticale	B1	2709	39	2	3	5	20
Competenza grammaticale	B1	2750	39	3	4	5	20
Competenza grammaticale	B1	2639	39	3	3	5	0
Competenza grammaticale	B1	2722	39	4	4	5	0
Competenza grammaticale	B1	2741	39	4	4	5	0
Competenza grammaticale	B1	2724	39	5	4	5	20
Competenza grammaticale	B1	2739	39	5	3	5	40

Competenza grammaticale	B1	2749	39	5	3	5	40
Competenza grammaticale	B1	2771	48	1	2	5	20
Competenza grammaticale	B1	2778	48	2	3	5	20
Competenza grammaticale	B1	2761	48	3	3	5	0
Competenza grammaticale	B1	2803	48	3	3	5	0
Competenza grammaticale	B1	2786	48	4	3	5	20
Competenza grammaticale	B1	2792	48	4	3	5	20
Competenza grammaticale	B1	2757	48	5	3	5	40
Competenza grammaticale	B1	2776	48	5	4	5	20
Competenza grammaticale	B1	2785	48	5	4	5	20
Competenza grammaticale	B1	2787	48	5	4	5	20
Competenza grammaticale	B1	126	1	2	3	5	20
Competenza grammaticale	B1	280	1	2	2	5	0
Competenza grammaticale	B1	819	1	3	3	5	0
Competenza grammaticale	B1	894	1	3	3	5	0
Competenza grammaticale	B1	29	1	4	3	5	20
Competenza grammaticale	B1	46	1	4	3	5	20
Competenza grammaticale	B1	803	1	5	3	5	40

Competenza grammaticale	B1	817	1	5	5	5	0
Competenza grammaticale	B1	818	1	5	3	5	40
Competenza grammaticale	B1	852	1	5	5	5	0
Competenza grammaticale	B1	1508	9	2	2	5	0
Competenza grammaticale	B1	1527	9	2	3	5	20
Competenza grammaticale	B1	505	9	3	3	5	0
Competenza grammaticale	B1	572	9	3	3	5	0
Competenza grammaticale	B1	1656	9	4	3	5	20
Competenza grammaticale	B1	1493	9	4	4	5	0
Competenza grammaticale	B1	1143	9	5	4	5	20
Competenza grammaticale	B1	1206	9	5	5	5	0
Competenza grammaticale	B1	1210	9	5	4	5	20
Competenza grammaticale	B1	1258	9	5	5	5	0
Competenza grammaticale	B1	2595	33	2	3	5	20
Competenza grammaticale	B1	2677	33	2	2	5	0
Competenza grammaticale	B1	2453	33	2	3	5	20
Competenza grammaticale	B1	2580	33	3	3	5	0
Competenza grammaticale	B1	2586	33	3	3	5	0

Competenza grammaticale	B1	2454	33	4	3	5	20
Competenza grammaticale	B1	2459	33	4	3	5	20
Competenza grammaticale	B1	2591	33	4	3	5	20
Competenza grammaticale	B1	2583	33	5	5	5	0
Competenza grammaticale	B1	2598	33	5	4	5	20
Competenza sociolinguistica	B1	1858	17	2	4	5	40
Competenza sociolinguistica	B1	1866	17	2	3	5	20
Competenza sociolinguistica	B1	1830	17	3	4	5	20
Competenza sociolinguistica	B1	1862	17	3	3	5	0
Competenza sociolinguistica	B1	1778	17	4	3	5	20
Competenza sociolinguistic a	B1	1815	17	4	4	5	0
Competenza sociolinguistica	B1	1901	17	5	4	5	20
Competenza sociolinguistica	B1	2195	17	5	4	5	20

Competenza sociolinguistica	B1	1780	17	5	4	5	20
Competenza sociolinguistica	B1	1825	17	5	4	5	20
Competenza sociolinguistica	B1	2535	25	2	3	5	20
Competenza sociolinguistica	B1	2324	25	3	3	5	0
Competenza sociolinguistica	B1	2373	25	3	3	5	0
Competenza sociolinguistica	B1	2382	25	3	3	5	0
Competenza sociolinguistic a	B1	2145	25	4	3	5	20
Competenza sociolinguistica	B1	2381	25	4	4	5	0
Competenza sociolinguistic a	B1	2412	25	4	4	5	0
Competenza sociolinguistic a	B1	2224	25	5	4	5	20
Competenza sociolinguistica	B1	2274	25	5	3	5	40
Competenza sociolinguistica	B1	2311	25	5	4	5	20

Competenza sociolinguistic a	B1	2628	39	3	3	5	0
Competenza sociolinguistic a	B1	2645	39	3	3	5	0
Competenza sociolinguistic a	B1	2660	39	3	4	5	20
Competenza sociolinguistic a	B1	2675	39	4	4	5	0
Competenza sociolinguistic a	B1	2689	39	4	4	5	0
Competenza sociolinguistic a	B1	2702	39	4	3	5	20
Competenza sociolinguistic a	B1	2743	39	5	4	5	20
Competenza sociolinguistic a	B1	2744	39	5	4	5	20
Competenza sociolinguistic a	B1	2746	39	5	3	5	40
Competenza sociolinguistic a	B1	2750	39	5	4	5	20
Competenza sociolinguistic a	B1	2775	48	3	3	5	0
Competenza sociolinguistic a	B1	2768	48	3	4	5	20

Competenza sociolinguistica	B1	2783	48	4	4	5	0
Competenza sociolinguistica	B1	2784	48	4	4	5	0
Competenza sociolinguistica	B1	2757	48	4	4	5	0
Competenza sociolinguistica	B1	2791	48	4	4	5	0
Competenza	B1	277	48	5	3	5	40
sociolinguistica		0					
Competenza sociolinguistica	B1	2788	48	5	4	5	20
Competenza sociolinguistica	B1	2790	48	5	4	5	20
Competenza sociolinguistica	B1	2794	48	5	4	5	20
Competenza sociolinguistica	B1	2578	33	3	3	5	0
Competenza sociolinguistica	B1	2586	33	3	3	5	0
Competenza sociolinguistica	B1	2601	33	3	3	5	0
Competenza sociolinguistica	B1	2463	33	4	4	5	0

Competenza sociolinguistica	B1	2581	33	4	3	5	20
Competenza sociolinguistica	B1	2587	33	4	4	5	0
Competenza sociolinguistica	B1	2585	33	5	3	5	40
Competenza sociolinguistica	B1	2605	33	5	4	5	20
Competenza sociolinguistica	B1	2583	33	5	4	5	20
Competenza sociolinguistica	B1	2598	33	5	3	5	40

Dimensione	**QCER**	**ID**	**Traccia**	**Punt. VU**	**Punt. GPT**	**Punt. Max**	**Err. Norm.**
Coerenza e coesione	B1	466	1	3	4	5	20
Coerenza e coesione	B1	113	1	3	3	5	0
Coerenza e coesione	B1	548	1	3	4	5	20
Coerenza e coesione	B1	53	1	4	4	5	0
Coerenza e coesione	B1	163	1	4	4	5	0
Coerenza e coesione	B1	322	1	4	4	5	0
Coerenza e coesione	B1	119	1	5	3	5	40

Coerenza e coesione	B1	733	1	5	4	5	20
Coerenza e coesione	B1	61	1	5	4	5	20
Coerenza e coesione	B1	87	1	5	4	5	20
Coerenza e coesione	B1	843	9	3	3	5	0
Coerenza e coesione	B1	505	9	3	4	5	20
Coerenza e coesione	B1	572	9	3	3	5	0
Coerenza e coesione	B1	955	9	3	3	5	0
Coerenza e	B1	1030	9	4	3	5	20
coesione							
Coerenza e coesione	B1	1117	9	4	5	5	20
Coerenza e coesione	B1	901	9	4	3	5	20
Coerenza e coesione	B1	973	9	4	3	5	20
Coerenza e coesione	B1	842	9	4	3	5	20
Coerenza e coesione	B1	914	9	5	4	5	20
Coerenza e coesione	B1	1848	17	2	2	5	0
Coerenza e coesione	B1	1887	17	2	2	5	0
Coerenza e coesione	B1	1989	17	3	3	5	0
Coerenza e coesione	B1	1829	17	3	3	5	0
Coerenza e coesione	B1	1816	17	4	2	5	40

Coerenza e coesione	B1	1835	17	4	3	5	20
Coerenza e coesione	B1	1824	17	5	4	5	20
Coerenza e coesione	B1	1845	17	5	4	5	20
Coerenza e coesione	B1	1847	17	5	5	5	0
Coerenza e coesione	B1	1884	17	5	5	5	0
Coerenza e coesione	B1	2612	25	2	3	5	20
Coerenza e coesione	B1	2145	25	2	2	5	0
Coerenza e coesione	B1	2311	25	3	3	5	0
Coerenza e coesione	B1	2153	25	3	3	5	0
Coerenza e coesione	B1	2379	25	4	4	5	0
Coerenza e coesione	B1	2401	25	4	4	5	0
Coerenza e coesione	B1	2432	25	5	4	5	20
Coerenza e coesione	B1	2493	25	5	4	5	20
Coerenza e coesione	B1	2529	25	5	5	5	0
Coerenza e coesione	B1	2530	25	5	4	5	20
Coerenza e coesione	B1	2589	33	3	3	5	0
Coerenza e coesione	B1	2590	33	3	3	5	0
Coerenza e coesione	B1	2600	33	3	3	5	0

Coerenza e coesione	B1	2469	33	4	3	5	20
Coerenza e coesione	B1	2573	33	4	3	5	20
Coerenza e coesione	B1	2606	33	4	4	5	0
Coerenza e coesione	B1	2594	33	5	3	5	40
Coerenza e coesione	B1	2677	33	5	3	5	40
Coerenza e coesione	B1	2465	33	5	3	5	40
Coerenza e coesione	B1	2559	33	5	5	5	0

Dimensione	**QCER**	**ID**	**Traccia**	**Punt. VU**	**Punt. GPT**	**Punt. Max**	**Err. Norm.**
Competenza lessicale	B2	435	2	2	3	5	20
Competenza lessicale	B2	785	2	2	3	5	20
Competenza lessicale	B2	453	2	3	3	5	0
Competenza lessicale	B2	3	2	3	3	5	0
Competenza lessicale	B2	214	2	4	4	5	0
Competenza lessicale	B2	230	2	4	4	5	0
Competenza lessicale	B2	570	2	5	4	5	20
Competenza lessicale	B2	695	2	5	3	5	40
Competenza lessicale	B2	858	2	5	3	5	40

Competenza lessicale	B2	883	2	5	4	5	20
Competenza lessicale	B2	279	3	2	3	5	20
Competenza lessicale	B2	951	3	3	3	5	0
Competenza lessicale	B2	62	3	3	3	5	0
Competenza lessicale	B2	100	3	4	3	5	20
Competenza lessicale	B2	155	3	4	3	5	20
Competenza lessicale	B2	485	3	4	3	5	20
Competenza lessicale	B2	13	3	5	3	5	40
Competenza lessicale	B2	14	3	5	3	5	40
Competenza lessicale	B2	16	3	5	3	5	40
Competenza lessicale	B2	45	3	5	4	5	20
Competenza lessicale	B2	1065	10	2	3	5	20
Competenza lessicale	B2	1070	10	2	3	5	20
Competenza lessicale	B2	1455	10	3	3	5	0
Competenza lessicale	B2	1510	10	3	3	5	0
Competenza lessicale	B2	1565	10	4	3	5	20
Competenza lessicale	B2	1591	10	4	3	5	20
Competenza lessicale	B2	1295	10	5	4	5	20

Competenza lessicale	B2	473	10	5	3	5	40
Competenza lessicale	B2	482	10	5	3	5	40
Competenza lessicale	B2	539	10	5	3	5	40
Competenza lessicale	B2	1477	11	2	3	5	20
Competenza lessicale	B2	808	11	2	2	5	0
Competenza lessicale	B2	603	11	3	4	5	20
Competenza lessicale	B2	1353	11	3	3	5	0
Competenza lessicale	B2	1244	11	4	3	5	20
Competenza lessicale	B2	1367	11	4	3	5	20
Competenza	B2	483	11	5	3	5	40
lessicale							
Competenza lessicale	B2	1019	11	5	3	5	40
Competenza lessicale	B2	1229	11	5	2	5	60
Competenza lessicale	B2	633	11	5	4	5	20
Competenza lessicale	B2	1664	18	2	3	5	20
Competenza lessicale	B2	1707	18	3	3	5	0
Competenza lessicale	B2	1801	18	3	3	5	0
Competenza lessicale	B2	1661	18	3	3	5	0
Competenza lessicale	B2	1997	18	4	4	5	0

Competenza lessicale	B2	1998	18	4	3	5	20
Competenza lessicale	B2	2118	18	4	3	5	20
Competenza lessicale	B2	1795	18	5	4	5	20
Competenza lessicale	B2	1800	18	5	4	5	20
Competenza lessicale	B2	1938	18	5	4	5	20

Dimensione	**QCER**	**ID**	**Traccia**	**Punt. VU**	**Punt. GPT**	**Punt. Max**	**Err. Norm.**
Competenza grammaticale	B2	1987	26	2	2	5	0
Competenza grammaticale	B2	2011	26	2	2	5	0
Competenza grammaticale	B2	2086	26	3	3	5	0
Competenza grammaticale	B2	2088	26	3	3	5	0
Competenza grammaticale	B2	2083	26	4	4	5	0
Competenza grammaticale	B2	2085	26	5	4	5	20
Competenza grammaticale	B2	1970	26	5	4	5	20
Competenza grammaticale	B2	1947	26	5	4	5	20
Competenza grammaticale	B2	2040	26	5	4	5	20
Competenza grammaticale	B2	2108	26	5	4	5	20
Competenza grammaticale	B2	1963	27	2	3	5	20

Competenza grammaticale	B2	2120	27	2	3	5	20
Competenza grammaticale	B2	2112	27	3	3	5	0
Competenza grammaticale	B2	1973	27	3	3	5	0
Competenza grammaticale	B2	1957	27	4	4	5	0
Competenza grammaticale	B2	1992	27	4	3	5	20
Competenza grammaticale	B2	2070	27	5	3	5	40
Competenza grammaticale	B2	1946	27	5	4	5	20
Competenza grammaticale	B2	1949	27	5	4	5	20
Competenza grammaticale	B2	1960	27	5	4	5	20
Competenza grammaticale	B2	183	2	2	3	5	20
Competenza	B2	799	2	3	3	5	0
grammaticale							
Competenza grammaticale	B2	858	2	3	3	5	0
Competenza grammaticale	B2	772	2	3	3	5	0
Competenza grammaticale	B2	455	2	4	4	5	0
Competenza grammaticale	B2	507	2	4	4	5	0
Competenza grammaticale	B2	8	2	5	4	5	20
Competenza grammaticale	B2	9	2	5	4	5	20
Competenza grammaticale	B2	12	2	5	3	5	40

Competenza grammaticale	B2	17	2	5	4	5	20
Competenza grammaticale	B2	916	11	3	3	5	0
Competenza grammaticale	B2	483	11	4	3	5	20
Competenza grammaticale	B2	650	11	4	3	5	20
Competenza grammaticale	B2	1049	11	4	3	5	20
Competenza grammaticale	B2	3101	11	4	3	5	20
Competenza grammaticale	B2	1229	11	5	3	5	40
Competenza grammaticale	B2	633	11	5	4	5	20
Competenza grammaticale	B2	1150	11	5	4	5	20
Competenza grammaticale	B2	1521	11	5	4	5	20
Competenza grammaticale	B2	3098	11	5	4	5	20
Competenza grammaticale	B2	1708	18	3	3	5	0
Competenza grammaticale	B2	1721	18	4	3	5	20
Competenza grammaticale	B2	2064	18	4	3	5	20
Competenza grammaticale	B2	2078	18	4	2	5	40
Competenza grammaticale	B2	1717	18	4	4	5	0
Competenza grammaticale	B2	1794	18	5	4	5	20
Competenza grammaticale	B2	1795	18	5	3	5	40

Competenza grammaticale	B2	1800	18	5	4	5	20
Competenza grammaticale	B2	1995	18	5	4	5	20
Competenza grammaticale	B2	2003	18	5	4	5	20

Competenza sociolinguistica	B2	329	3	3	3	5	0
Competenza sociolinguistica	B2	358	3	3	3	5	0
Competenza sociolinguistica	B2	381	3	3	3	5	0
Competenza sociolinguistica	B2	485	3	4	4	5	0
Competenza sociolinguistica	B2	516	3	4	4	5	0
Competenza sociolinguistica	B2	568	3	4	3	5	20
Competenza sociolinguistica	B2	800	3	5	4	5	20
Competenza sociolinguistica	B2	802	3	5	3	5	40
Competenza sociolinguistica	B2	859	3	5	3	5	40

Competenza sociolinguistica	B2	860	3	5	3	5	40
Competenza sociolinguistica	B2	1331	10	3	2	5	20
Competenza sociolinguistica	B2	1435	10	3	3	5	0
Competenza sociolinguistica	B2	1436	10	3	3	5	0
Competenza sociolinguistica	B2	473	10	4	4	5	0
Competenza sociolinguistica	B2	479	10	4	3	5	20
Competenza sociolinguistica	B2	481	10	4	4	5	0
Competenza sociolinguistica	B2	1553	10	5	4	5	20
Competenza	B2	1557	10	5	3	5	40
sociolinguistica							
Competenza sociolinguistica	B2	1565	10	5	4	5	20
Competenza sociolinguistica	B2	1590	10	5	3	5	40
Competenza sociolinguistica	B2	1448	11	3	2	5	20

Competenza sociolinguistica	B2	1477	11	3	3	5	0
Competenza sociolinguistica	B2	1520	11	3	3	5	0
Competenza sociolinguistica	B2	1589	11	4	4	5	0
Competenza sociolinguistica	B2	1600	11	4	3	5	20
Competenza sociolinguistica	B2	1609	11	4	4	5	0
Competenza sociolinguistica	B2	1232	11	5	4	5	20
Competenza sociolinguistica	B2	1240	11	5	3	5	40
Competenza sociolinguistica	B2	1330	11	5	4	5	20
Competenza sociolinguistica	B2	1334	11	5	3	5	40
Competenza sociolinguistica	B2	1802	18	4	4	5	0
Competenza sociolinguistica	B2	2119	18	4	4	5	0
Competenza sociolinguistica	B2	1689	18	5	4	5	20

Competenza sociolinguistica	B2	1693	18	5	3	5	40
Competenza sociolinguistica	B2	1697	18	5	4	5	20
Competenza sociolinguistica	B2	1699	18	5	4	5	20
Competenza sociolinguistica	B2	1701	18	5	5	5	0
Competenza sociolinguistica	B2	1794	18	5	4	5	20
Competenza sociolinguistica	B2	1795	18	5	4	5	20
Competenza sociolinguistica	B2	1800	18	5	4	5	20
Competenza sociolinguistica	B2	1971	27	4	3	5	20
Competenza sociolinguistica	B2	1988	27	4	4	5	0

Competenza sociolinguistica	B2	1991	27	4	4	5	0
Competenza sociolinguistica	B2	2070	27	4	3	5	20
Competenza sociolinguistica	B2	1946	27	5	4	5	20

Competenza sociolinguistica	B2	1949	27	5	4	5	20
Competenza sociolinguistica	B2	1960	27	5	5	5	0
Competenza sociolinguistica	B2	2005	27	5	3	5	40
Competenza sociolinguistica	B2	2071	27	5	3	5	40
Competenza sociolinguistica	B2	2111	27	5	4	5	20

Dimensione	**QCER**	**ID**	**Traccia**	**Punt. VU**	**Punt. GPT**	**Punt. Max**	**Err. Norm.**
Coerenza e coesione	B2	207	2	2	2	5	0
Coerenza e coesione	B2	441	2	2	2	5	0
Coerenza e coesione	B2	96	2	3	3	5	0
Coerenza e coesione	B2	169	2	3	3	5	0
Coerenza e	B2	205	2	4	3	5	20
coesione							
Coerenza e coesione	B2	71	2	4	4	5	0
Coerenza e coesione	B2	519	2	5	4	5	20
Coerenza e coesione	B2	524	2	5	4	5	20

Coerenza e coesione	B2	528	2	5	4	5	20
Coerenza e coesione	B2	664	2	5	4	5	20
Coerenza e coesione	B2	90	3	2	3	5	20
Coerenza e coesione	B2	303	3	2	3	5	20
Coerenza e coesione	B2	179	3	3	4	5	20
Coerenza e coesione	B2	231	3	3	3	5	0
Coerenza e coesione	B2	425	3	4	4	5	0
Coerenza e coesione	B2	428	3	4	3	5	20
Coerenza e coesione	B2	100	3	5	4	5	20
Coerenza e coesione	B2	155	3	5	3	5	40
Coerenza e coesione	B2	180	3	5	4	5	20
Coerenza e coesione	B2	234	3	5	3	5	40
Coerenza e coesione	B2	1451	10	2	2	5	0
Coerenza e coesione	B2	827	10	2	2	5	0
Coerenza e coesione	B2	778	10	3	4	5	20
Coerenza e coesione	B2	1331	10	3	2	5	20
Coerenza e coesione	B2	694	10	4	4	5	0
Coerenza e coesione	B2	707	10	4	3	5	20

Coerenza e coesione	B2	1437	10	5	3	5	40
Coerenza e coesione	B2	539	10	5	3	5	40
Coerenza e coesione	B2	550	10	5	3	5	40
Coerenza e coesione	B2	627	10	5	4	5	20
Coerenza e coesione	B2	1353	11	2	2	5	0
Coerenza e coesione	B2	1006	11	2	2	5	0
Coerenza e coesione	B2	1149	11	3	3	5	0
Coerenza e coesione	B2	1330	11	3	3	5	0
Coerenza e coesione	B2	1477	11	4	3	5	20
Coerenza e coesione	B2	483	11	4	4	5	0
Coerenza e coesione	B2	1019	11	5	4	5	20
Coerenza e coesione	B2	1071	11	5	3	5	40
Coerenza e coesione	B2	1244	11	5	3	5	40
Coerenza e coesione	B2	1367	11	5	3	5	40
Coerenza e coesione	B2	1974	26	2	4	5	40
Coerenza e coesione	B2	2103	26	2	3	5	20
Coerenza e coesione	B2	2086	26	3	4	5	20
Coerenza e coesione	B2	2020	26	3	4	5	20

Coerenza e coesione	B2	2057	26	4	4	5	0
Coerenza e coesione	B2	2083	26	4	3	5	20
Coerenza e coesione	B2	2036	26	5	2	5	60
Coerenza e coesione	B2	1947	26	5	4	5	20
Coerenza e coesione	B2	2076	26	5	4	5	20
Coerenza e coesione	B2	2098	26	5	3	5	40

Dimensione	**QCER**	**ID**	**Traccia**	**Punt. VU**	**Punt. GPT**	**Punt. Max**	**Err. Norm.**
Competenza lessicale	C1	1164	5	3	4	8	13
Competenza lessicale	C1	259	5	4	4	8	0
Competenza lessicale	C1	1163	5	5	4	8	13
Competenza lessicale	C1	185	5	5	5	8	0
Competenza lessicale	C1	160	5	6	5	8	13
Competenza lessicale	C1	212	5	6	5	8	13
Competenza lessicale	C1	370	5	7	5	8	25
Competenza lessicale	C1	424	5	7	7	8	0
Competenza lessicale	C1	35	5	8	8	8	0
Competenza lessicale	C1	170	5	8	6	8	25

Competenza lessicale	C1	2935	29	4	4	8	0
Competenza lessicale	C1	2113	29	4	5	8	13
Competenza lessicale	C1	2936	29	5	4	8	13
Competenza lessicale	C1	2338	29	5	5	8	0
Competenza lessicale	C1	2878	29	6	6	8	0
Competenza lessicale	C1	2926	29	6	6	8	0
Competenza lessicale	C1	2190	29	7	4	8	38
Competenza lessicale	C1	2312	29	7	5	8	25
Competenza lessicale	C1	2267	29	8	7	8	13
Competenza lessicale	C1	2490	29	8	7	8	13
Competenza lessicale	C1	2830	40	3	5	8	25
Competenza lessicale	C1	2895	40	3	4	8	13
Competenza lessicale	C1	2847	40	4	5	8	13
Competenza lessicale	C1	2825	40	4	5	8	13
Competenza lessicale	C1	2822	40	5	5	8	0
Competenza lessicale	C1	2818	40	5	6	8	13
Competenza lessicale	C1	2819	40	5	7	8	25
Competenza	C1	2828	40	5	5	8	0
lessicale							

Competenza lessicale	C1	2821	40	6	5	8	13
Competenza lessicale	C1	2815	40	7	7	8	0
Competenza lessicale	C1	2814	41	3	4	8	13
Competenza lessicale	C1	2843	41	3	5	8	25
Competenza lessicale	C1	2827	41	4	4	8	0
Competenza lessicale	C1	2844	41	4	4	8	0
Competenza lessicale	C1	2846	41	5	4	8	13
Competenza lessicale	C1	2893	41	5	4	8	13
Competenza lessicale	C1	2824	41	6	5	8	13
Competenza lessicale	C1	2823	41	6	5	8	13
Competenza lessicale	C1	2896	41	6	5	8	13
Competenza lessicale	C1	2845	41	7	7	8	0
Competenza lessicale	C1	2976	54	3	6	8	38
Competenza lessicale	C1	2985	54	3	7	8	50
Competenza lessicale	C1	2979	54	4	7	8	38
Competenza lessicale	C1	2990	54	5	5	8	0
Competenza lessicale	C1	2967	54	5	5	8	0
Competenza lessicale	C1	2982	54	6	7	8	13

Competenza lessicale	C1	2989	54	6	7	8	13
Competenza lessicale	C1	2975	54	7	5	8	25
Competenza lessicale	C1	2978	54	7	7	8	0
Competenza lessicale	C1	2983	54	7	7	8	0

Dimensione	**QCER**	**ID**	**Traccia**	**Punt. VU**	**Punt. GPT**	**Punt. Max**	**Err. Norm.**
Competenza grammaticale	C1	314	4	3	6	8	38
Competenza grammaticale	C1	72	4	4	6	8	25
Competenza grammaticale	C1	147	4	5	5	8	0
Competenza grammaticale	C1	362	4	5	6	8	13
Competenza grammaticale	C1	1183	4	6	5	8	13
Competenza grammaticale	C1	68	4	6	7	8	13
Competenza grammaticale	C1	81	4	7	7	8	0
Competenza grammaticale	C1	187	4	7	7	8	0
Competenza grammaticale	C1	459	4	8	7	8	13
Competenza grammaticale	C1	532	4	8	6	8	25
Competenza grammaticale	C1	2908	20	4	5	8	13
Competenza grammaticale	C1	2944	20	4	6	8	25

Competenza	C1	2954	20	5	7	8	25
grammaticale							
Competenza grammaticale	C1	1804	20	5	5	8	0
Competenza grammaticale	C1	2951	20	6	6	8	0
Competenza grammaticale	C1	1709	20	6	6	8	0
Competenza grammaticale	C1	1694	20	7	6	8	13
Competenza grammaticale	C1	1787	20	7	7	8	0
Competenza grammaticale	C1	2947	20	8	7	8	13
Competenza grammaticale	C1	2950	20	8	6	8	25
Competenza grammaticale	C1	2473	28	3	6	8	38
Competenza grammaticale	C1	2228	28	4	6	8	25
Competenza grammaticale	C1	2240	28	5	6	8	13
Competenza grammaticale	C1	2399	28	5	5	8	0
Competenza grammaticale	C1	2331	28	6	6	8	0
Competenza grammaticale	C1	2446	28	6	6	8	0
Competenza grammaticale	C1	2927	28	7	7	8	0
Competenza grammaticale	C1	2931	28	7	7	8	0
Competenza grammaticale	C1	2929	28	8	6	8	25
Competenza grammaticale	C1	2939	28	8	6	8	25

Competenza grammaticale	C1	2916	34	2	6	8	50
Competenza grammaticale	C1	2919	34	3	6	8	38
Competenza grammaticale	C1	2921	34	4	6	8	25
Competenza grammaticale	C1	2912	34	5	7	8	25
Competenza grammaticale	C1	2514	34	5	6	8	13
Competenza grammaticale	C1	2918	34	6	5	8	13
Competenza grammaticale	C1	2516	34	6	6	8	0
Competenza grammaticale	C1	2504	34	7	7	8	0
Competenza grammaticale	C1	2483	34	7	7	8	0
Competenza grammaticale	C1	2502	34	8	7	8	13
Competenza grammaticale	C1	2920	35	3	6	8	38
Competenza grammaticale	C1	2913	35	4	6	8	25
Competenza grammaticale	C1	2922	35	4	6	8	25
Competenza grammaticale	C1	2915	35	5	6	8	13
Competenza grammaticale	C1	2476	35	5	6	8	13
Competenza grammaticale	C1	2865	35	5	6	8	13
Competenza grammaticale	C1	2480	35	5	6	8	13
Competenza grammaticale	C1	2866	35	6	7	8	13

Competenza grammaticale	C1	2517	35	6	7	8	13
Competenza	C1	2506	35	7	7	8	0
grammaticale							

Dimensione	**QCER**	**ID**	**Traccia**	**Punt. VU**	**Punt. GPT**	**Punt. Max**	**Err. Norm.**
Competenza sociolinguistica	C1	811	12	4	4	6	0
Competenza sociolinguistica	C1	812	12	4	4	6	0
Competenza sociolinguistica	C1	1106	12	4	4	6	0
Competenza sociolinguistica	C1	710	12	5	5	6	0
Competenza sociolinguistica	C1	711	12	5	4	6	17
Competenza sociolinguistica	C1	714	12	5	4	6	17
Competenza sociolinguistica	C1	1419	12	6	4	6	33
Competenza sociolinguistica	C1	1427	12	6	4	6	33
Competenza sociolinguistica	C1	1466	12	6	5	6	17
Competenza sociolinguistica	C1	1523	12	6	4	6	33
Competenza sociolinguistica	C1	1189	13	4	4	6	0
Competenza sociolinguistica	C1	1276	13	4	5	6	17
Competenza sociolinguistica	C1	654	13	4	4	6	0
Competenza sociolinguistica	C1	759	13	5	5	6	0

Competenza sociolinguistica	C1	820	13	5	5	6	0
Competenza sociolinguistica	C1	844	13	5	5	6	0
Competenza sociolinguistica	C1	1389	13	6	5	6	17
Competenza sociolinguistica	C1	1418	13	6	4	6	33
Competenza sociolinguistica	C1	1421	13	6	5	6	17
Competenza sociolinguistica	C1	1429	13	6	5	6	17
Competenza sociolinguistica	C1	2913	35	3	4	6	17
Competenza sociolinguistica	C1	2920	35	4	4	6	0
Competenza sociolinguistica	C1	2915	35	4	3	6	17
Competenza sociolinguistica	C1	2922	35	5	4	6	17
Competenza sociolinguistica	C1	2476	35	5	4	6	17
Competenza sociolinguistica	C1	2865	35	5	4	6	17
Competenza sociolinguistica	C1	2866	35	5	5	6	0
Competenza sociolinguistica	C1	2480	35	6	4	6	33
Competenza sociolinguistica	C1	2517	35	6	4	6	33
Competenza sociolinguistica	C1	2506	35	6	5	6	17
Competenza sociolinguistica	C1	2845	41	4	5	6	17
Competenza sociolinguistica	C1	2816	41	5	4	6	17

Competenza sociolinguistica	C1	2826	41	5	4	6	17
Competenza sociolinguistica	C1	2827	41	5	4	6	17
Competenza	C1	2844	41	5	5	6	0
sociolinguistica							
Competenza sociolinguistica	C1	2843	41	6	4	6	33
Competenza sociolinguistica	C1	2892	41	6	4	6	33
Competenza sociolinguistica	C1	2894	41	6	5	6	17
Competenza sociolinguistica	C1	2823	41	6	5	6	17
Competenza sociolinguistica	C1	2896	41	6	5	6	17
Competenza sociolinguistica	C1	2988	54	3	4	6	17
Competenza sociolinguistica	C1	2981	54	4	4	6	0
Competenza sociolinguistica	C1	2990	54	4	4	6	0
Competenza sociolinguistica	C1	2984	54	4	5	6	17
Competenza sociolinguistica	C1	2989	54	5	5	6	0
Competenza sociolinguistica	C1	2963	54	5	5	6	0
Competenza sociolinguistica	C1	2964	54	5	5	6	0
Competenza sociolinguistica	C1	2969	54	5	5	6	0
Competenza sociolinguistica	C1	2975	54	5	5	6	0
Competenza sociolinguistica	C1	2985	54	6	4	6	33

Dimensione	QCER	ID	Traccia	Punt. VU	Punt. GPT	Punt. Max	Err. Norm.
Coerenza e coesione	C1	1361	12	4	6	8	25
Coerenza e coesione	C1	1434	12	4	6	8	25
Coerenza e coesione	C1	1410	12	5	7	8	25
Coerenza e coesione	C1	1365	12	5	7	8	25
Coerenza e coesione	C1	511	12	6	6	8	0
Coerenza e coesione	C1	655	12	6	7	8	13
Coerenza e coesione	C1	811	12	7	7	8	0
Coerenza e coesione	C1	1106	12	7	6	8	13
Coerenza e coesione	C1	1523	12	8	7	8	13
Coerenza e coesione	C1	1559	12	8	8	8	0
Coerenza e coesione	C1	2949	20	4	6	8	25
Coerenza e coesione	C1	2903	20	4	6	8	25
Coerenza e coesione	C1	1941	20	5	6	8	13
Coerenza e coesione	C1	1709	20	6	6	8	0
Coerenza e coesione	C1	2945	20	6	6	8	0
Coerenza e coesione	C1	1943	20	7	7	8	0

Coerenza e coesione	C1	2953	20	7	6	8	13
Coerenza e coesione	C1	2905	20	8	7	8	13
Coerenza e coesione	C1	1804	20	8	6	8	25
Coerenza e coesione	C1	1785	20	8	8	8	0
Coerenza e coesione	C1	2192	28	4	6	8	25
Coerenza e coesione	C1	2142	28	4	6	8	25
Coerenza e coesione	C1	2446	28	5	5	8	0
Coerenza e coesione	C1	2449	28	5	6	8	13
Coerenza e coesione	C1	2334	28	6	6	8	0
Coerenza e coesione	C1	2022	28	6	6	8	0
Coerenza e coesione	C1	2228	28	7	6	8	13
Coerenza e coesione	C1	2268	28	7	7	8	0
Coerenza e coesione	C1	2393	28	8	7	8	13
Coerenza e coesione	C1	2394	28	8	7	8	13
Coerenza e coesione	C1	2476	35	6	7	8	13
Coerenza e coesione	C1	2865	35	6	6	8	0
Coerenza e coesione	C1	2913	35	7	7	8	0
Coerenza e coesione	C1	2915	35	7	8	8	13

Coerenza e coesione	C1	2866	35	7	7	8	0
Coerenza e coesione	C1	2920	35	8	8	8	0
Coerenza e coesione	C1	2922	35	8	6	8	25
Coerenza e coesione	C1	2480	35	8	6	8	25
Coerenza e coesione	C1	2517	35	8	7	8	13
Coerenza e coesione	C1	2506	35	8	6	8	25
Coerenza e coesione	C1	2845	41	4	5	8	13
Coerenza e coesione	C1	2844	41	6	6	8	0
Coerenza e coesione	C1	2846	41	6	6	8	0
Coerenza e coesione	C1	2843	41	7	5	8	25
Coerenza e coesione	C1	2816	41	8	5	8	38
Coerenza e coesione	C1	2826	41	8	8	8	0
Coerenza e coesione	C1	2827	41	8	6	8	25
Coerenza e coesione	C1	2893	41	8	6	8	25
Coerenza e coesione	C1	2824	41	8	6	8	25
Coerenza e coesione	C1	2814	41	8	5	8	38

Dimensione	**QCER**	**ID**	**Traccia**	**Punt. VU**	**Punt. GPT**	**Punt. Max**	**Err. Norm.**

Competenza lessicale	C2	1178	6	4	6	8	25
Competenza lessicale	C2	190	6	5	8	8	38
Competenza lessicale	C2	145	6	5	7	8	25
Competenza lessicale	C2	78	6	6	6	8	0
Competenza	C2	384	6	6	6	8	0
lessicale							
Competenza lessicale	C2	265	6	7	7	8	0
Competenza lessicale	C2	128	6	7	7	8	0
Competenza lessicale	C2	487	6	8	7	8	13
Competenza lessicale	C2	450	6	9	8	8	13
Competenza lessicale	C2	805	6	9	8	8	13
Competenza lessicale	C2	129	7	4	5	8	13
Competenza lessicale	C2	1179	7	4	6	8	25
Competenza lessicale	C2	356	7	5	6	8	13
Competenza lessicale	C2	1182	7	6	6	8	0
Competenza lessicale	C2	1160	7	7	6	8	13
Competenza lessicale	C2	1161	7	7	7	8	0
Competenza lessicale	C2	365	7	8	8	8	0
Competenza lessicale	C2	619	7	8	7	8	13

Competenza lessicale	C2	366	7	9	8	8	13
Competenza lessicale	C2	1175	7	9	8	8	13
Competenza lessicale	C2	614	14	5	7	8	25
Competenza lessicale	C2	1014	14	5	5	8	0
Competenza lessicale	C2	1016	14	6	6	8	0
Competenza lessicale	C2	1305	14	6	8	8	25
Competenza lessicale	C2	1383	14	7	7	8	0
Competenza lessicale	C2	1310	14	7	5	8	25
Competenza lessicale	C2	828	14	8	8	8	0
Competenza lessicale	C2	1018	14	8	8	8	0
Competenza lessicale	C2	1578	14	9	8	8	13
Competenza lessicale	C2	1340	14	9	8	8	13
Competenza lessicale	C2	2957	24	3	6	8	38
Competenza lessicale	C2	1965	24	5	7	8	25
Competenza lessicale	C2	1793	24	6	7	8	13
Competenza lessicale	C2	1810	24	6	8	8	25
Competenza lessicale	C2	1840	24	7	7	8	0
Competenza lessicale	C2	1905	24	7	7	8	0

Competenza lessicale	C2	1841	24	8	8	8	0
Competenza lessicale	C2	1634	24	8	8	8	0
Competenza lessicale	C2	1753	24	9	8	8	13
Competenza lessicale	C2	1851	24	9	8	8	13
Competenza lessicale	C2	3040	57	3	6	8	38
Competenza	C2	2959	57	5	7	8	25
lessicale							
Competenza lessicale	C2	3024	57	6	6	8	0
Competenza lessicale	C2	3025	57	6	7	8	13
Competenza lessicale	C2	3028	57	6	1	8	63
Competenza lessicale	C2	2980	57	7	7	8	0
Competenza lessicale	C2	3034	57	7	8	8	13
Competenza lessicale	C2	2961	57	7	7	8	0
Competenza lessicale	C2	2987	57	8	7	8	13
Competenza lessicale	C2	2977	57	8	7	8	13

Dimensione	**QCER**	**ID**	**Traccia**	**Punt. VU**	**Punt. GPT**	**Punt. Max**	**Err. Norm.**
Competenza grammaticale	C2	363	8	4	6	6	33

Competenza grammaticale	C2	440	8	4	5	6	17
Competenza grammaticale	C2	191	8	5	5	6	0
Competenza grammaticale	C2	69	8	6	6	6	0
Competenza grammaticale	C2	115	8	7	7	6	0
Competenza grammaticale	C2	25	8	7	6	6	17
Competenza grammaticale	C2	15	8	8	7	6	17
Competenza grammaticale	C2	297	8	8	6	6	33
Competenza grammaticale	C2	32	8	8	7	6	17
Competenza grammaticale	C2	405	8	8	7	6	17
Competenza grammaticale	C2	712	15	5	6	6	17
Competenza grammaticale	C2	992	15	5	6	6	17
Competenza grammaticale	C2	881	15	6	6	6	0
Competenza grammaticale	C2	1157	15	6	6	6	0
Competenza grammaticale	C2	685	15	7	7	6	0
Competenza grammaticale	C2	1424	15	7	7	6	0
Competenza grammaticale	C2	1308	15	8	5	6	50
Competenza grammaticale	C2	1309	15	8	6	6	33
Competenza grammaticale	C2	1339	15	8	6	6	33

Competenza grammaticale	C2	1416	15	8	7	6	17
Competenza grammaticale	C2	2435	30	4	6	6	33
Competenza grammaticale	C2	2438	30	5	6	6	17
Competenza grammaticale	C2	2899	30	5	4	6	17
Competenza grammaticale	C2	2942	30	5	6	6	17
Competenza grammaticale	C2	2129	30	5	7	6	33
Competenza grammaticale	C2	2091	30	6	5	6	17
Competenza	C2	2900	30	6	6	6	0
grammaticale							
Competenza grammaticale	C2	2437	30	7	7	6	0
Competenza grammaticale	C2	2439	30	7	7	6	0
Competenza grammaticale	C2	2130	30	7	7	6	0
Competenza grammaticale	C2	2608	42	4	5	6	17
Competenza grammaticale	C2	2700	42	4	6	6	33
Competenza grammaticale	C2	2726	42	5	5	6	0
Competenza grammaticale	C2	2729	42	5	6	6	17
Competenza grammaticale	C2	2620	42	6	6	6	0
Competenza grammaticale	C2	2841	42	6	6	6	0
Competenza grammaticale	C2	2832	42	6	6	6	0

Competenza grammaticale	C2	2626	42	7	6	6	17
Competenza grammaticale	C2	2727	42	7	6	6	17
Competenza grammaticale	C2	2714	42	7	7	6	0
Competenza grammaticale	C2	3083	63	5	5	6	0
Competenza grammaticale	C2	3011	63	5	5	6	0
Competenza grammaticale	C2	3018	63	6	6	6	0
Competenza grammaticale	C2	3012	63	6	6	6	0
Competenza grammaticale	C2	3010	63	6	6	6	0
Competenza grammaticale	C2	3089	63	6	6	6	0
Competenza grammaticale	C2	2998	63	7	6	6	17
Competenza grammaticale	C2	3007	63	7	6	6	17
Competenza grammaticale	C2	3016	63	7	6	6	17
Competenza grammaticale	C2	3017	63	7	6	6	17

Dimensione	**QCER**	**ID**	**Traccia**	**Punt. VU**	**Punt. GPT**	**Punt. Max**	**Err. Norm.**
Competenza sociolinguistica	C2	579	16	5	6	8	13
Competenza sociolinguistica	C2	1544	16	6	7	8	13
Competenza sociolinguistica	C2	1384	16	6	7	8	13

Competenza sociolinguistica	C2	1319	16	7	7	8	0
Competenza sociolinguistica	C2	1012	16	8	7	8	13
Competenza sociolinguistica	C2	1098	16	8	7	8	13
Competenza sociolinguistica	C2	1196	16	9	7	8	25
Competenza sociolinguistica	C2	1273	16	9	7	8	25
Competenza sociolinguistica	C2	1420	16	9	8	8	13
Competenza sociolinguistica	C2	1433	16	9	7	8	25
Competenza sociolinguistica	C2	1791	22	5	6	8	13
Competenza	C2	1909	22	5	8	8	38
sociolinguistica							
Competenza sociolinguistica	C2	1767	22	6	7	8	13
Competenza sociolinguistica	C2	1772	22	6	7	8	13
Competenza sociolinguistica	C2	1936	22	7	7	8	0
Competenza sociolinguistica	C2	2956	22	7	7	8	0
Competenza sociolinguistica	C2	1688	22	8	8	8	0
Competenza sociolinguistica	C2	1690	22	8	7	8	13
Competenza sociolinguistica	C2	1908	22	9	7	8	25
Competenza sociolinguistica	C2	1966	22	9	8	8	13
Competenza sociolinguistica	C2	2199	31	5	7	8	25

Competenza sociolinguistica	C2	2418	31	5	7	8	25
Competenza sociolinguistica	C2	2264	31	6	7	8	13
Competenza sociolinguistica	C2	2303	31	6	7	8	13
Competenza sociolinguistica	C2	2489	31	7	7	8	0
Competenza sociolinguistica	C2	2607	31	7	7	8	0
Competenza sociolinguistica	C2	2305	31	8	8	8	0
Competenza sociolinguistica	C2	2306	31	8	8	8	0
Competenza sociolinguistica	C2	2302	31	9	7	8	25
Competenza sociolinguistica	C2	2451	31	9	7	8	25
Competenza sociolinguistica	C2	3071	50	4	6	8	25
Competenza sociolinguistica	C2	3053	50	5	7	8	25
Competenza sociolinguistica	C2	3046	50	6	6	8	0
Competenza sociolinguistica	C2	3078	50	6	6	8	0
Competenza sociolinguistica	C2	3062	50	7	7	8	0
Competenza sociolinguistica	C2	3080	50	7	7	8	0
Competenza sociolinguistica	C2	3057	50	8	7	8	13
Competenza sociolinguistica	C2	3065	50	8	8	8	0
Competenza sociolinguistica	C2	3074	50	9	7	8	25

Competenza sociolinguistica	C2	3079	50	9	7	8	25
Competenza sociolinguistica	C2	3086	65	5	7	8	25
Competenza sociolinguistica	C2	3090	65	5	7	8	25
Competenza sociolinguistica	C2	3022	65	6	8	8	25
Competenza sociolinguistica	C2	3023	65	6	7	8	13
Competenza sociolinguistica	C2	3021	65	7	7	8	0
Competenza sociolinguistica	C2	3094	65	7	7	8	0
Competenza sociolinguistica	C2	3015	65	8	8	8	0
Competenza sociolinguistica	C2	3088	65	8	7	8	13
Competenza	C2	3013	65	9	8	8	13
sociolinguistica							
Competenza sociolinguistica	C2	3020	65	9	7	8	25

Coerenza e coesione	C2	3018	63	4	5	8	13
Coerenza e coesione	C2	3083	63	6	4	8	25
Coerenza e coesione	C2	3011	63	7	6	8	13
Coerenza e coesione	C2	3012	63	8	6	8	25
Coerenza e coesione	C2	3010	63	8	7	8	13
Coerenza e coesione	C2	3089	63	8	4	8	50

Coerenza e coesione	C2	2998	63	9	7	8	25
Coerenza e coesione	C2	3007	63	9	7	8	25
Coerenza e coesione	C2	3016	63	9	8	8	13
Coerenza e coesione	C2	3017	63	9	8	8	13
Coerenza e coesione	C2	1304	14	4	7	8	38
Coerenza e coesione	C2	1417	14	4	7	8	38
Coerenza e coesione	C2	828	14	5	6	8	13
Coerenza e coesione	C2	614	14	6	6	8	0
Coerenza e coesione	C2	1318	14	7	5	8	25
Coerenza e coesione	C2	1578	14	7	7	8	0
Coerenza e coesione	C2	1271	14	8	7	8	13
Coerenza e coesione	C2	1341	14	8	7	8	13
Coerenza e coesione	C2	1016	14	9	6	8	38
Coerenza e coesione	C2	1015	14	9	6	8	38
Coerenza e coesione	C2	3069	49	5	7	8	25
Coerenza e coesione	C2	3052	49	6	6	8	0
Coerenza e coesione	C2	3067	49	6	7	8	13
Coerenza e coesione	C2	2780	49	7	7	8	0

Coerenza e coesione	C2	3058	49	7	6	8	13
Coerenza e coesione	C2	3070	49	8	7	8	13
Coerenza e coesione	C2	3076	49	8	7	8	13
Coerenza e coesione	C2	3045	49	9	7	8	25
Coerenza e coesione	C2	2755	49	9	7	8	25
Coerenza e coesione	C2	3051	49	9	7	8	25
Coerenza e coesione	C2	3032	58	3	7	8	50
Coerenza e coesione	C2	2974	58	5	7	8	25
Coerenza e coesione	C2	3044	58	5	7	8	25
Coerenza e coesione	C2	3029	58	6	6	8	0
Coerenza e	C2	2962	58	6	5	8	13
coesione							
Coerenza e coesione	C2	2970	58	7	5	8	25
Coerenza e coesione	C2	3041	58	8	7	8	13
Coerenza e coesione	C2	2973	58	8	6	8	25
Coerenza e coesione	C2	3043	58	9	7	8	25
Coerenza e coesione	C2	3033	58	9	7	8	25
Coerenza e coesione	C2	1909	22	3	7	8	50
Coerenza e coesione	C2	1903	22	4	7	8	38

Coerenza e coesione	C2	1791	22	5	6	8	13
Coerenza e coesione	C2	1936	22	6	7	8	13
Coerenza e coesione	C2	1844	22	7	8	8	13
Coerenza e coesione	C2	1934	22	7	7	8	0
Coerenza e coesione	C2	2911	22	8	6	8	25
Coerenza e coesione	C2	1763	22	8	7	8	13
Coerenza e coesione	C2	2910	22	9	8	8	13
Coerenza e coesione	C2	1633	22	9	9	8	0

Appendice B - Scale di competenza e punteggi delle prove CELI

CELI 2 – CELI 2 a – CELI 2 *i* (livello B1) – PRODUZIONE SCRITTA - SCALE DI COMPETENZE E PUNTEGGI[1]

COMPETENZA LESSICALE

Si attribuisce il punteggio di		ad un compito che presenta un repertorio lessicale
3 punti	5 punti	sempre adeguato, con rari errori ortografici.
2 punti	4 punti	semplice, quasi sempre adeguato, con pochi errori ortografici. Talvolta ripetitivo.
1 punto	3 punti	semplice, ma a tratti inadeguato, con diversi (massimo sei) errori ortografici e ripetizioni.
0 -1 punti	2 punti	limitato, spesso inadeguato, con frequenti errori ortografici e ripetizioni che a volte rendono poco compiuto il senso del testo.
0 punti	0-1 punti	povero, inadeguato e con sistematici errori ortografici che rendono nel complesso non compiuto il senso del testo.
Prova B.2	*Prova B.3*	

COMPETENZA SOCIOLINGUISTICA

Si attribuisce il punteggio di		ad un compito che presenta
5 punti	5 punti	moduli espressivi sempre appropriati alla situazione. Il candidato è sempre pienamente comprensibile.
4 punti	4 punti	qualche errore che però non compromette mai l'appropriatezza dei moduli espressivi. Il candidato riesce sempre a far capire ciò che vuole esprimere.
3 punti	3 punti	moduli espressivi non sempre appropriati, ma riesce a far capire gran parte di ciò che vuole esprimere.
2 punti	2 punti	moduli espressivi spesso non appropriati. Non sempre riesce a far capire ciò che vuole esprimere.
0-1 punti	0-1 punti	errori che rendono non appropriati i moduli espressivi. Il candidato non riesce a far capire ciò che vuole esprimere.
Prova B.2	*Prova B.3*	

COMPETENZA GRAMMATICALE

Si attribuisce il punteggio di		ad un compito che presenta
3 punti	5 punti	strutture semplici, con rari errori e con un buon collegamento tra frasi.
2 punti	4 punti	pochi errori, anche se a volte le frasi non sono ben collegate.
1 punto	3 punti	diversi (massimo sei) errori che però non compromettono mai la comunicazione. Frasi scarsamente collegate, ma la correlazione dei tempi è rispettata.
0 -1 punti	2 punti	frequenti errori che rendono a volte poco comprensibile il testo. Scarsi collegamenti tra frasi. La correlazione dei tempi spesso non è rispettata.
0 punti	0-1 punti	sistematici errori che, per quantità e gravità, rendono poco comprensibile il testo.
Prova B.2	*Prova B.3*	

COERENZA E COESIONE

Si attribuisce il punteggio di		ad un compito che è svolto
4 punti	5 punti	totalmente (nel pieno rispetto delle istruzioni), in modo ben organizzato sul piano logico, coeso e lineare.
3 punti	4 punti	totalmente, con un discreto ordine logico anche se le varie parti del testo non hanno uno sviluppo equilibrato.
2 punti	3 punti	con alcune omissioni non rilevanti ai fini della sua realizzazione e con un ordine logico appena sufficiente.
1 punto	2 punti	solo in parte (uno o più punti delle istruzioni non sono svolti) e in modo non ben organizzato sul piano logico.
0 punti	0-1 punti	in minima parte e in cui si nota l'incapacità di comprendere quanto richiesto.
Prova B.2	*Prova B.3*	

[1] Per i descrittori relativi al lessico e alla grammatica si deve fare riferimento ai repertori linguistici del *Profilo* (B. Spinelli, F. Parizzi, *Profilo della lingua italiana*, La Nuova Italia, Firenze 2010).

CELI 3 – CELI 3 a (livello B2) - PRODUZIONE SCRITTA - SCALE DI COMPETENZE E PUNTEGGI[1]

COMPETENZA LESSICALE

Si attribuisce il punteggio di		ad un compito che presenta un repertorio lessicale
5 punti	5 punti	buono e sempre adeguato, con rari errori ortografici.
4 punti	4 punti	adeguato, con pochi errori ortografici.
3 punti	3punti	quasi sempre adeguato, con diversi (massimo quattro) errori ortografici. Talvolta ripetitivo.
2 punti	2 punti	semplice e spesso inadeguato, con frequenti errori ortografici e ripetizioni.
0-1 punti	0-1 punti	limitato e inadeguato, con sistematici errori ortografici che a volte rendono poco compiuto il senso del testo.
Prova B.1	*Prova B.2*	

COMPETENZA SOCIOLINGUISTICA

Si attribuisce il punteggio di		ad un compito che presenta
5 punti	6 punti	moduli espressivi sempre corretti e appropriati alla situazione.
4 punti	5 punti	imperfezioni nell'uso di moduli espressivi che risultano tuttavia appropriati.
3 punti	4 punti	errori che però non compromettono mai l'appropriatezza dei moduli espressivi.
2 punti	3 punti	moduli espressivi spesso non appropriati.
0-1 punti	0-2 punti	moduli espressivi quasi sempre non appropriati.
Prova B.1	*Prova B.2*	

COMPETENZA GRAMMATICALE

Si attribuisce il punteggio di		ad un compito che presenta
5 punti	4 punti	una certa varietà di strutture. Il testo prodotto è ben articolato. Rari gli errori, generalmente buona la padronanza grammaticale.
4 punti	3 punti	pochi errori, ma qualche difficoltà nel collegamento tra frasi.
3 punti	2 punti	strutture semplici, con diversi (massimo quattro) errori. Frasi non sempre ben collegate tra loro.
2 punti	1 punto	strutture limitate, con frequenti errori e con scarsi collegamenti tra frasi.
0-1 punti	0 punti	sistematici errori e difficoltà nella costruzione delle frasi.
Prova B.1	*Prova B.2*	

COERENZA E COESIONE

Si attribuisce il punteggio di		ad un compito che è svolto
5 punti	5 punti	totalmente (nel pieno rispetto delle istruzioni), con un buon ordine logico e una buona coesione.
4 punti	4 punti	totalmente, con un discreto ordine logico. Il testo è coeso, anche se le varie parti non hanno sempre uno sviluppo equilibrato.
3 punti	3 punti	con alcune omissioni non rilevanti ai fini della sua realizzazione e con un ordine logico e una coesione appena sufficienti.
2 punti	2 punti	solo in parte (uno o più punti delle istruzioni non sono svolti) e in modo non sempre ben organizzato sul piano logico.
0-1 punti	0-1 punti	in minima parte e in cui si nota spesso l'incapacità di comprendere a pieno quanto richiesto.
Prova B.1	*Prova B.2*	

[1] Per i descrittori relativi al lessico e alla grammatica si deve fare riferimento ai repertori linguistici del *Profilo* (B. Spinelli, F. Parizzi, *Profilo della lingua italiana*, La Nuova Italia, Firenze 2010).

CELI 4 (livello C1) – PRODUZIONE SCRITTA - SCALE DI COMPETENZE E PUNTEGGI

COMPETENZA LESSICALE

Si attribuisce il punteggio di		ad un compito che presenta un repertorio lessicale
3 punti	8 punti	vasto, vario e sempre adeguato, con occasionali errori ortografici e con una buona padronanza di espressioni idiomatiche.
2 punti	7 punti	vario e adeguato, con rari errori ortografici e una discreta padronanza di espressioni idiomatiche.
1 punto	6 punti	adeguato, anche se talvolta ripetitivo, con pochi (massimo tre) errori ortografici.
0 -1 punti	4-5 punti	limitato e a volte inadeguato, con diversi errori ortografici e ripetizioni.
0 punti	0-3 punti	molto limitato e generalmente inadeguato, con frequenti errori ortografici e ripetizioni.
Prova B.1	*Prova B.2*	

COMPETENZA SOCIOLINGUISTICA

Si attribuisce il punteggio di		ad un compito che presenta
6 punti	6 punti	moduli espressivi sempre corretti e appropriati alla situazione. La lingua è usata in modo flessibile, con una buona efficacia comunicativa.
5 punti	5 punti	piccole imperfezioni nell'uso di moduli espressivi che risultano tuttavia appropriati. La lingua è usata in modo flessibile, con una discreta efficacia comunicativa.
4 punti	4 punti	imperfezioni nell'uso dei moduli espressivi. Non sempre efficace sul piano comunicativo.
3 punti	3 punti	errori nell'uso dei moduli espressivi. Scarsamente efficace sul piano comunicativo.
0-2 punti	0-2 punti	moduli espressivi spesso non appropriati. Mancanza di efficacia comunicativa.
Prova B.1	*Prova B.2*	

COMPETENZA GRAMMATICALE

Si attribuisce il punteggio di		ad un compito che presenta
4 punti	8 punti	un'ampia gamma di strutture corrette, con una costante padronanza e un elevato controllo linguistico (salvo occasionali errori in costruzioni di uso poco frequente). Il testo prodotto è ben strutturato.
3 punti	7 punti	una certa varietà di strutture quasi sempre corrette. Il testo è ben articolato, con un buon uso di collegamenti fra le varie parti. Rari gli errori, con una padronanza grammaticale nel complesso buona.
2 punti	6 punti	strutture non sempre corrette, con pochi (massimo tre) errori. Qualche difficoltà nel collegare le varie parti del testo.
1 punto	4-5 punti	strutture semplici e non sempre corrette, con diversi errori e con la prevalenza di collegamenti di tipo coordinativo.
0 punti	0-3 punti	strutture limitate e spesso non corrette, con frequenti errori.
Prova B.1	*Prova B.2*	

COERENZA E COESIONE

Si attribuisce il punteggio di		ad un compito che è svolto
7 punti	8 punti	totalmente (nel pieno rispetto delle istruzioni). Il testo presenta un chiaro ordine logico, è ben coeso, scorrevole e con un uso controllato degli schemi organizzativi.
6 punti	7 punti	totalmente. Il testo presenta un buon ordine logico, è coeso e abbastanza scorrevole.
5 punti	6 punti	con omissioni non rilevanti ai fini della sua realizzazione. Il testo presenta un discreto ordine logico, è abbastanza coeso, anche se le varie parti non hanno sempre uno sviluppo equilibrato.
4 punti	4-5 punti	con alcune omissioni. Il testo non è sempre ben organizzato sul piano logico e risulta poco coeso.
0-3 punti	0-3 punti	solo in parte (uno o più punti delle istruzioni non sono svolti). Il testo non ben organizzato sul piano logico e risulta non coeso. Si nota spesso l'incapacità di trattare gli argomenti in modo approfondito e coerente.
Prova B.1	*Prova B.2*	

CELI 5 (livello C2) - PRODUZIONE SCRITTA - SCALE DI COMPETENZE E PUNTEGGI

COMPETENZA LESSICALE

Si attribuisce il punteggio di		ad un compito che presenta un repertorio lessicale
9 punti	6 punti	vasto, molto vario, sempre adeguato e preciso, con un'elevata padronanza di espressioni idiomatiche.
7-8 punti	4-5 punti	vasto, vario e adeguato e con una buona padronanza di espressioni idiomatiche. Occasionali imprecisioni non disturbano la lettura.
6 punti	3 punti	generalmente adeguato, con rari (massimo due) errori ortografici e con una discreta padronanza di espressioni idiomatiche.
4-5 punti	2 punti	quasi sempre adeguato, talvolta ripetitivo e con pochi errori che però abbassano il livello espressivo.
0-3 punti	0-1 punti	limitato e spesso inadeguato, con diversi errori ortografici e ripetizioni. Il livello espressivo è generalmente basso.
Prova B.1	*Prova B.2*	

COMPETENZA SOCIOLINGUISTICA

Si attribuisce il punteggio di		ad un compito che presenta
8-9 punti	6-7 punti	moduli espressivi sempre corretti e appropriati alla situazione. La lingua è usata in modo molto flessibile, con un'alta efficacia comunicativa. Nel testo vengono colte pienamente le implicazioni socioculturali, con buoni riferimenti ad elementi di civiltà italiana.
6-7 punti	5 punti	moduli espressivi appropriati. La lingua è usata in modo flessibile, con una buona efficacia comunicativa. Nel testo vengono colte le implicazioni socioculturali, anche con riferimenti ad elementi di civiltà italiana.
5 punti	4 punti	moduli espressivi appropriati, seppure con occasionali imperfezioni. Il testo, pur non cogliendo sempre le implicazioni socioculturali, è efficace sul piano comunicativo.
4 punti	3 punti	imperfezioni nell'uso dei moduli espressivi. Non sempre efficace sul piano comunicativo.
0-3 punti	0-2 punti	moduli espressivi talvolta non appropriati. Scarsamente efficace sul piano comunicativo.
Prova B.1	*Prova B.2*	

COMPETENZA GRAMMATICALE

Si attribuisce il punteggio di		ad un compito che presenta
8 punti	6 punti	un'ampia gamma di strutture corrette, con una costante padronanza e un quasi totale controllo linguistico, anche di costrutti complessi (salvo occasionali imperfezioni in costruzioni di uso raro). Il testo prodotto è ben strutturato in tutte le sue parti.
6-7 punti	5 punti	un'ampia gamma di strutture corrette, con una buona padronanza e un elevato controllo (salvo occasionali errori in costruzioni di uso poco frequente). Il testo prodotto è generalmente ben strutturato.
5 punti	4 punti	una certa varietà di strutture quasi sempre corrette. Il testo è ben articolato, con un buon uso di collegamenti fra le varie parti. Rari (massimo due) gli errori, con una padronanza grammaticale nel complesso buona che però denota la mancanza di un autentico controllo.
4 punti	3 punti	strutture non sempre corrette. Pochi errori, ma la padronanza non è sempre buona. Qualche difficoltà nel collegare le varie parti del testo.
0-3 punti	0-2 punti	strutture semplici e non sempre corrette, con diversi errori e con la prevalenza di collegamenti di tipo coordinativo.
Prova B.1	*Prova B.2*	

COERENZA E COESIONE

Si attribuisce il punteggio di		ad un compito che è svolto
8-9 punti	6 punti	totalmente (nel pieno rispetto delle istruzioni). Il testo presenta un ottimo ordine logico, è molto coeso e scorrevole, con una grande varietà di schemi organizzativi.
6-7 punti	5 punti	totalmente. Il testo presenta un chiaro ordine logico, è ben coeso, scorrevole e con un uso controllato degli schemi organizzativi.
5 punti	4 punti	totalmente. Il testo presenta un buon ordine logico, è coeso e abbastanza scorrevole.
4 punti	3 punti	con omissioni non rilevanti ai fini della sua realizzazione. Il testo presenta un discreto ordine logico, ma è poco coeso e le varie parti non hanno uno sviluppo equilibrato.
0-3 punti	0-2 punti	con alcune omissioni. Il testo non è ben organizzato sul piano logico e risulta generalmente non coeso.
Prova B.1	*Prova B.2*	