

Valutare la qualità dei testi generati in lingua italiana

Mirko Tavosanis (Università di Pisa)

mirko.tavosanis(at)unipi.it

Abstract

Il contributo presenta il problema della valutazione dei testi generati in lingua italiana da LLM collocandolo nel contesto delle pratiche di valutazione esistenti in diversi settori connessi. Nella prima parte viene mostrato il posizionamento della valutazione degli LLM in rapporto ad alcune pratiche ampiamente diffuse, dalla valutazione semiautomatica delle traduzioni automatiche alla valutazione umana di testi di vario genere. Nella seconda parte vengono presentate le condizioni della valutazione in rapporto alle capacità effettive degli LLM, che sono in grado di generare testi di tipo molto diverso. Viene poi presentato l'esame di un campione di testi generati da ChatGPT, eseguito appunto nella prospettiva dell'identificazione delle pratiche di valutazione più adatte. Le produzioni testuali oggetto dell'esame, ricavate da conversazioni condotte con ChatGPT-3.5 e ChatGPT-4, mostrano infatti un avvicinamento notevole alle produzioni professionali di esseri umani, ma anche alcuni punti di divergenza. Complessivamente, il contributo argomenta quindi in favore dell'importanza e dell'utilità della valutazione umana applicata ai testi di buon livello linguistico prodotti dai sistemi attuali basati sull'intelligenza artificiale, purché in questa applicazione venga evitata l'adozione passiva di criteri sviluppati per altri tipi di attività.

The paper addresses the issue of evaluating texts generated in Italian by Large Language Models (LLM), placing it within the context of existing evaluation practices in various related sectors. In the first part, the positioning of LLM evaluation is shown in relation to widely adopted practices, from the semi-automatic evaluation of machine translations to the human evaluation of texts pertaining to different textual genres. The second part describes the conditions of evaluation in relation to the actual capabilities of LLMs, which can generate a wide variety of texts. Subsequently, an examination of a sample of texts generated by ChatGPT is presented, conducted with the aim of identifying the most suitable evaluation practices. The textual productions examined, derived from conversations with ChatGPT-3.5 and ChatGPT-4, show a notable resemblance to professional human productions, while also revealing some points of divergence. Overall, the contribution argues in favour of the importance and utility of human evaluation applied to high-quality linguistic texts produced by current AI-based systems, provided that the passive adoption of criteria developed for other types of activities is avoided.

Keywords

lingua, intelligenza artificiale, testi generati da intelligenze artificiali, lingua italiana, valutazione, LLM, ChatGPT

Language, AI, AI generated texts, Italian language, evaluation, LLM, ChatGPT

1 Introduzione

L'apertura al grande pubblico di sistemi per la generazione di testo basati su Large Language Models (LLM) risale a tempi molto recenti. Un punto di svolta cruciale è stato infatti il lancio dei servizi online di ChatGPT il 30 novembre 2022 (Wu et al. 2023).

AILing

AI-Linguistica.

Linguistic Studies on AI-Generated Texts and Discourses

Tavosanis, Mirko
Valutare la qualità dei testi generati in lingua italiana

AI-Linguistica 2024. Vol. 1 No. 1

DOI: 10.62408/ai-ling.v1i1.14

ISSN: 2943-0070

La prima diffusione di questi sistemi ha prodotto in tutto il mondo reazioni molto differenziate, anche tra gli addetti ai lavori. Alcune valutazioni estreme (per esempio, quelle espresse in Lemoine 2023) sono arrivate fino ad attribuire agli LLM la qualifica di esseri senzienti. Per quanto idee del genere siano evidentemente contraddette dalla comune esperienza, il fatto che siano state presentate in pubblico ha un notevole pregio epistemico: rende evidente quanto sia difficile valutare anche solo le capacità di base di un prodotto innovativo.

All'inizio del 2024, comunque, sembra stia prendendo forma una comune consapevolezza almeno per quanto riguarda le capacità generali degli LLM. In sostanza, nonostante le prestazioni spesso sorprendenti, questi sistemi possono essere descritti molto meglio partendo dalla celebre etichetta di “pappagalli stocastici” (*stochastic parrots*: Bender et al. 2022) che da quella di “intelligenze artificiali”. Per esempio, l’idea, proposta da Wei et al. (2022), che gli LLM mostrino “proprietà emergenti”, con la comparsa di inattese capacità di ragionamento una volta superata una soglia dimensionale nel corpus su cui sono addestrati, sembra smentita dagli approfondimenti delle ricerche. Pare in effetti che le prestazioni dipendano in modo lineare dalle dimensioni dei corpora di addestramento, senza l’esistenza di soglie (Lu et al. 2023; Schaeffer, Brando e Sanmi 2023). In generale, è evidente che sul piano linguistico gli LLM hanno alcune capacità enormemente superiori a quelle dei sistemi disponibili in passato; ma altrettanto evidente è l’esistenza di forti limiti.

Inquadrare la situazione in modo esatto è quindi di notevole interesse. Sembra ormai diffusa, per fortuna, la consapevolezza della differenza tra le capacità linguistiche formali (rispettare le regole grammaticali) e quelle funzionali (“comprendere” il significato di un testo): gli LLM sono validi nella gestione delle prime, ma assai meno in quella delle seconde (Mahowald et al. 2023). Insomma, il fatto che esseri umani intelligenti e che abbiano pratica di scrittura siano in grado di scrivere testi complessi e articolati non vuol dire che un sistema che scrive testi complessi e articolati sia intelligente. Né il fatto che alcuni testi vadano bene dal punto di vista linguistico implica che vadano bene da tutti i punti di vista.

Per quanto riguarda le capacità generative in italiano, è comunque chiaro che gli LLM sono molto superiori ai grezzi sistemi del recente passato (alcuni commenti su questi ultimi sono in Tavosanis 2018: 16). Uno studioso di assoluta competenza come Claudio Marazzini, che si occupa con attenzione del rapporto tra lingua e tecnologia, ha rilasciato diverse interviste in cui si è espresso in termini entusiastici sulle capacità di questi modelli. Per esempio, ha evidenziato che ChatGPT “fa un uso corretto della lingua italiana”, è “ottimo sia nella comprensione sia nella scrittura” e “si comporta come un parlante nativo” (Marazzini 2023).

Non c’è dubbio che simili entusiasmi abbiano basi concrete. Per alcuni tipi di testo, l’italiano generato da questi modelli senz’altro contiene meno deviazioni dalla norma rispetto a quello prodotto dalla media degli studenti universitari del I anno di un corso di laurea dell’area umanistica (questo è un giudizio non documentato, ma fondato su un ventennio di esperienza come titolare di Laboratori di scrittura per studenti universitari del I anno presso l’Università di Pisa). Tuttavia,

ciò non vuol dire che l'italiano generato sia sempre perfetto. La valutazione dei limiti del prodotto può inoltre essere molto importante per capire le potenzialità dei sistemi, al di là di quanto possono dire le valutazioni superficiali – incluse quelle fornite dagli stessi sviluppatori, che (a differenza di ciò che avviene, per esempio, nella produzione di autoveicoli, o di sistemi informatici tradizionali) non sono in grado di prevedere in dettaglio ciò che i sistemi sono o non sono in grado di fare.

Andare oltre questo livello richiede però una riflessione piuttosto ampia e interdisciplinare. Nelle comunità che si occupano dello sviluppo informatico di questi sistemi esistono infatti indicatori e criteri standard (*benchmark*) per valutare la qualità degli strumenti (per l'italiano, il riferimento esemplare è dato dalla campagna di valutazioni EVALITA).¹ Tuttavia, per quanto riguarda gli LLM, il lavoro è ancora in corso: una promettente occasione di sviluppo dovrebbe essere data dall'evento CALAMITA – Challenge the Abilities of LAngeuage Models in ITAlian, previsto per il dicembre 2024 e finalizzato a sviluppare un *benchmark* dinamico e adattabile per valutare le capacità degli LLM in italiano.²

Tuttavia, al momento questo genere di valutazione ha un'ottica piuttosto diversa da quella più strettamente linguistica. Valutare la qualità dei testi generati è in effetti un problema complesso, che non sembra avere una soluzione singola e semplice. In questo genere di valutazioni possono invece confluire tradizioni di studio diverse e molto diversificate tra di loro: la valutazione dei sistemi informatici così come condotta nelle comunità di sviluppo, la valutazione dei testi tradizionali, il testing linguistico e così via. Ognuno di questi approcci presenta vantaggi; la particolare situazione degli LLM richiede però senz'altro una riflessione a monte, che ponga l'enfasi in particolare sui limiti. Il presente contributo vuole mostrare in effetti il posizionamento della valutazione degli LLM in rapporto ad alcune di queste tradizioni (Sezioni 2 e 3). Un simile confronto interdisciplinare è necessariamente imperfetto; sembra però utile presentarlo in modo esplicito per facilitare la discussione. A questo fa poi seguito un esame delle produzioni testuali di ChatGPT, eseguito nella prospettiva appunto di una possibile valutazione (Sezioni 4 e 5).

2 I limiti della valutazione

Molti tipi di valutazione sono relativamente semplici da applicare, oggettivi e fondati su solide basi. Tuttavia, in diversi ambiti, incluse aree di grande rilevanza pratica, la valutazione può essere difficile, arbitraria o priva di basi scientifiche. Un simile giudizio può sembrare strano, in un mondo in cui per esempio la valutazione degli esseri umani, a scuola, sul lavoro e in altri ambienti, è continua e spesso quantificata, con percentuali e punteggi precisi fino al centesimo e al millesimo. Non c'è in effetti dubbio sul fatto che alcune caratteristiche degli esseri umani possano appunto essere quantificate con notevole precisione: per esempio, l'altezza

¹ Sito di riferimento: <<https://www.evalita.it/>>.

² Sito di riferimento: <<https://clic2024.ilc.cnr.it/calamita/>>.

o il peso degli individui. Tuttavia, altre si prestano peggio a questo tipo di valutazioni, e spesso è sorprendentemente difficile distinguere tra i due casi.

Come punto di partenza per riflettere su questo quadro articolato può essere presa la capacità di trascrivere un testo letto ad alta voce. Tale capacità può essere valutata in modo oggettivo, sia per gli esseri umani sia per i sistemi informatici. In quest'ultimo caso, è pratica di riferimento il calcolo del numero di parole sbagliate o *Word Error Rate* (WER), ottenuto confrontando il testo letto con la trascrizione realizzata dal sistema. Il calcolo del WER è condotto da decenni facendo ricorso a programmi dedicati, che ricavano la cosiddetta distanza di Levenshtein tra l'obiettivo e il risultato effettivo calcolando in modo standard il peso delle parole cancellate, aggiunte o scambiate. Tra questi programmi ha un ruolo di riferimento in particolare SCLITE, contenuto all'interno del pacchetto *Speech Recognition Scoring Toolkit* realizzato negli Stati Uniti (NIST 2009). Per la capacità di trascrizione, il WER è quindi un indicatore oggettivo e ben noto di successo (Tavosanis 2018: 49-50).

Tuttavia, le cose cambiano quando si passa ad altri tipi di capacità. Per esempio, al di là di formule ancora utilissime ma molto semplici come GULPEASE, sviluppate negli anni Ottanta (Lucisano e Piemontese 1988), non esiste nessun indicatore contemporaneamente oggettivo e approfondito che permetta di valutare la leggibilità o la comprensibilità di un testo in lingua italiana. Al servizio degli studi linguistici sono stati certo sviluppati strumenti che offrono altre capacità molto utili, come quella di calcolare le percentuali delle parole e dei lemmi che in un testo rientrano nel Vocabolario di Base della lingua italiana (come descritto da De Mauro 1980 e 2016); sistemi come READ-IT (Dell'Orletta, Montemagni e Venturi 2011), hanno poi integrato tali funzioni con sofisticate capacità di analisi morfosintattica. Tuttavia, questo ampliamento di capacità non ha portato alla creazione di indicatori di riferimento, né criteri di semplificazione come quelli esemplarmente sintetizzati da Piemontese (1996) sono applicabili meccanicamente alla valutazione di un testo. Anche in rapporto ad altre lingue, per quanto riguarda lo sviluppo di sistemi informatici per la semplificazione di testi, i bilanci più recenti invitano quindi, in modo del tutto ragionevole, alla massima cautela nell'uso degli strumenti di valutazione disponibili (Alva-Manchego, Scarton e Specia 2021).

Su un altro piano, un esempio ben noto ai linguisti, e anche ai comuni cittadini, è la classificazione del QCER. Il Quadro Comune Europeo di Riferimento per la valutazione dell'apprendimento linguistico presenta infatti come riferimento solo una suddivisione in sei grandi livelli, da A1 a C2. All'interno di quegli amplessimi contenitori trova posto tutta la gamma di competenze degli apprendenti: dal conoscere poche ed elementari parole fino all'esprimersi con disinvoltura su qualunque argomento. Inoltre, e in modo correlato, le indicazioni su come valutare le competenze complessive, anche a semplici fini di certificazione linguistica, sono, e devono essere, notevolmente sfumate (QCER 2001: 177-196). Se un singolo esame di una singola competenza può avere una misurazione precisa e oggettiva, lo stesso non si può dire per le competenze linguistico-comunicative nel loro assieme.

Anche le ricerche nel settore del testing linguistico hanno permesso di approfondire e sistematizzare molti aspetti della valutazione (per un quadro d'insieme: Fulcher e Harding 2022). Sono di estremo interesse metodologico, per esempio, le proposte di scale di valutazione che tengano conto, oltre che della competenza linguistica dimostrata attraverso “complexity, accuracy, and fluency” (‘complessità, correttezza e fluidità’), della competenza funzionale, mostrata attraverso la “functional adequacy” (‘adeguatezza funzionale’) (Kuiken e Vedder 2022). Tuttavia, simili ricerche sono mirate essenzialmente alla valutazione di competenze L2 di esseri umani, e molti studi si rivolgono soprattutto ai livelli più bassi delle competenze. Inoltre, la pratica della valutazione spesso non tiene ancora conto di tali acquisizioni.

Cosa ancor più rilevante, un quadro simile è valido anche per attività tradizionali, praticate ogni giorno da centinaia di migliaia di docenti, come la valutazione di elaborati in ambito scolastico. I pochi tentativi di studio approfondito (tra cui spicca Serianni e Benedetti 2015) hanno mostrato che la pratica scolastica italiana è spesso arbitraria e individualizzata. Ciò non vuol dire che si tratti di un'attività inutile: in fin dei conti, quando si ha a che fare con esseri umani, anche una valutazione idiosincratICA e asistemica può essere uno stimolo importante, o addirittura essenziale per la maturazione degli apprendenti. Tuttavia, pratiche del genere sono ovviamente ben poco utili se usate come base per comprendere ciò che un sistema informatico è in grado o non è in grado di fare. È peraltro possibile che nel prossimo futuro i sistemi basati sull'intelligenza artificiale diventino in grado di duplicare efficacemente queste valutazioni, fornendo – sia in ambito scolastico sia altrove – giudizi che molti potrebbero essere tentati di ritenere tanto validi quanto quelli normalmente in uso (un esempio di queste spinte, a livello internazionale, è fornito da Khan 2024).

Al livello più alto, chi esamini in modo sistematico la manualistica per la scrittura universitaria italiana non può fare a meno di constatare l'ovvio: in tesi del genere non viene mai fatto riferimento a criteri definiti e riconosciuti che permettano di valutare una scrittura come di buon livello (Gallina, Orlando e Tavosanis in stampa). Assai più facile è costruire esempi di scrittura che *non* è di buon livello; un fortunato manuale di scrittura recente, per esempio, è dedicato fin dal titolo a presentare esempi del modo in cui *non* si deve scrivere (Giunta 2018). In generale, la valutazione della qualità dei testi di alto livello in L1 sembra una delle attività per cui la formalizzazione di criteri è meno avanzata.

Per valutare i prodotti dei sistemi di generazione automatica di testo, in sintesi, sembra poco produttivo fare un semplice riferimento a criteri preesistenti. Tali criteri hanno come minimo bisogno di essere valutati criticamente prima di poter essere considerati applicabili in modo soddisfacente.

3 Un caso particolare: la valutazione della traduzione automatica

La sintesi presentata nella sezione precedente potrebbe sembrare troppo pessimistica. Può quindi essere utile descrivere più in dettaglio una situazione in

cui i problemi di valutazione derivati dall'applicazione meccanica a un prodotto nuovo di criteri preesistenti sono stati notevoli: è il caso degli sviluppi recenti della traduzione automatica, che rappresenta un caso particolare di generazione automatica di testo. Prima della diffusione degli LLM, in effetti, la traduzione automatica è senz'altro stata il tipo di attività in cui i sistemi di generazione di testo hanno raggiunto risultati più interessanti e in cui al tempo stesso è stata rivolta la maggiore attenzione alle pratiche di valutazione. Ricostruire l'evoluzione storica di tali pratiche negli ultimi anni sembra dunque utile anche per capire quali potrebbero essere i limiti della valutazione di testi generati in modo più libero. Va comunque ricordato, in via preliminare, che nel caso della traduzione automatica l'esistenza di un testo originale, da tradurre, fornisce la possibilità di ancorare la valutazione a un punto di riferimento oggettivo e piuttosto definito.

3.1 Il fallimento dei sistemi automatici e semiautomatici

Da oltre vent'anni, molti prodotti della traduzione automatica sono stati valutati attraverso un sistema semiautomatico chiamato BLEU (Papineni et al. 2002). La definizione di "semiautomatico" è dovuta al fatto che, per funzionare, il sistema ha bisogno di una traduzione umana di riferimento. Al momento di valutare una traduzione automatica BLEU la confronta infatti con una traduzione umana dello stesso testo, calcolando la differenza tra le due in termini di n-grammi, o sequenze di parole, di lunghezza 4. Alla base di questa scelta si trova un postulato: due buone traduzioni avranno molti punti di contatto. Tale postulato ha però diversi limiti, che l'applicazione pratica ha portato alla luce.

Proposto nel 2002, BLEU è diventato in pratica uno standard per valutare i sistemi di traduzione all'avanguardia nel quindicennio successivo, che usavano le tecniche PBST (*Phrase-Based Statistical Translation*, 'traduzione statistica basata su sintagmi' – così chiamata anche se in effetti i segmenti di testo in cui venivano scomposti i testi di regola non erano veri sintagmi). All'origine, BLEU mostrava una buona correlazione tra valutazioni umane e valutazioni automatiche. L'impiego di BLEU non è mai stato considerato migliore della valutazione umana, ma rispetto a questa era un sistema molto più facile ed economico da applicare. La valutazione umana richiede infatti tempo e lavoro; la valutazione con BLEU è quasi istantanea e gratuita (Graham e Baldwin 2014).

Come in molti casi simili, è lecito dubitare che la correlazione fosse così stretta in tutte le situazioni in cui BLEU è stato applicato. Di sicuro, però, l'introduzione dei sistemi a reti neurali ha cambiato completamente le cose. Google ha annunciato l'introduzione di questi sistemi sui suoi servizi online a partire dal 2016 (Turovsky 2016); per esempio, Google Traduttore ha iniziato a supportare la traduzione a reti neurali tra inglese e italiano nell'aprile del 2017. Sono bastati i primi controlli dei sistemi, commerciali o sperimentali, per constatare l'ovvio: la valutazione di BLEU era del tutto inaffidabile quando applicata alle traduzioni generate da sistemi a reti neurali. Banalmente, traduzioni che al lettore umano risultavano di qualità molto diversa ricevevano punteggi simili, e viceversa,

traduzioni di qualità simile ricevevano punteggi molto diversi. Nel 2018 sono stati dunque pubblicati diversi contributi che mostravano che non c'era una correlazione utilizzabile tra i punteggi BLEU e le valutazioni umane dei prodotti di traduzione automatica (Bentivogli et al. 2018; per controlli che includono l'italiano: Shterionov et al. 2018).

La constatazione dell'inaffidabilità di BLEU non ha peraltro impedito il suo uso: ancora oggi continuano a uscire pubblicazioni di alto livello in cui le prestazioni di sistemi a reti neurali vengono valutate usando BLEU in modo del tutto meccanico. A volte, cosa epistemologicamente interessante, valutazioni del genere vengono accompagnate da giustificazioni in cui si mostra consapevolezza dell'inutilità dello strumento ma si dichiara che occorre comunque usarlo in quanto fornisce dati "oggettivi" o in quanto semplicemente è di uso normale nel settore (per un esempio: Rivera-Trigueros 2022: 597-598). Un comportamento simile sembra un'evidente manifestazione del "principio del lampione" (*streetlight effect*, Freedman 2010, cap. 2): quello per cui gli ubriachi, quando fanno cadere le chiavi, le cercano non nel luogo in cui le hanno perse, ma in quello meglio illuminato. È però chiaro che a monte, oltre alla riluttanza individuale, l'inerzia si spiega in base a condizioni istituzionali e strutturali, spesso collegabili proprio a uno specifico modello di attività di valutazione.

Constatata l'attuale inutilità dei sistemi di valutazione semiautomatica, ci si può chiedere quali strumenti alternativi siano a disposizione dei valutatori. Purtroppo, tutti i sistemi automatici o semiautomatici che sono stati proposti negli ultimi anni, incluse versioni modificate di BLEU o sistemi basati sulla somiglianza semantica, hanno mostrato molti limiti e al momento sembra impossibile indicare un sostituto di riferimento. Nelle pubblicazioni di settore si afferma spesso che la valutazione migliore è quella che combina valutazioni umane e valutazioni automatiche o semiautomatiche (per un esempio: Rivera-Trigueros 2022: 612-613); ancora più corretto sarebbe forse affermare che, con tutte le sue imperfezioni, la valutazione umana è *l'unica* che sia affidabile.

3.2 I limiti della valutazione umana della traduzione automatica

Tornare alla valutazione umana delle traduzioni automatiche mette in effetti in luce un forte limite. Neanche per la valutazione umana di traduzioni esiste infatti un consenso sul modo "corretto" per procedere (per un quadro d'insieme con alcune proposte pratiche: Moorkens, Castilho, Gaspari e Doherty 2018).

Negli ambienti in cui è stata sviluppata la traduzione automatica il problema è stato spesso gestito usando sistemi di valutazione relativi a compiti specifici e basati su indicatori molto semplici: scale Likert o scale di valutazione con punteggio da 1 a 5. Simili scale sono poi state spesso applicate a due soli parametri: correttezza del contenuto e scorrevolezza. Inoltre, invece di valutare testi interi sono state spesso valutate frasi in isolamento, visto che la correlazione tra la valutazione di testi interi e quella di singole frasi è stata ritenuta buona. Ciò ha permesso di eseguire attività di valutazione relativamente rapide e spesso utili (White 1995).

Il fatto che questi sistemi di valutazione siano elementari non vuol dire però che la loro applicazione sia sempre corretta. Per esempio, come valutatori delle traduzioni oggi vengono spesso impiegate persone prive di particolare competenza in materia, reclutate attraverso sistemi di crowdsourcing (Jiménez-Crespo 2018). In tale contesto, la qualità dell'esecuzione è molto variabile: se per alcuni tipi di compito il crowdsourcing funziona bene, in alcuni casi i risultati sono poco soddisfacenti. Data l'alta qualità delle traduzioni automatiche moderne, occorrono infatti spesso professionisti della valutazione, non semplici parlanti madrelingua, per valutare per esempio la correttezza o l'accettabilità di determinate frasi (Jiménez-Crespo 2018: 88).

Problemi del genere possono essere gestiti in vario modo. Una precauzione elementare è quella di selezionare valutatori competenti. Un'altra consiste nel ridurre la variabilità usando griglie di valutazione dettagliate e attività di formazione per i valutatori, in modo simile a quanto si fa nei contesti più strutturati per la valutazione di produzioni umane (come dettagliata per esempio in Krippendorff 2019). Tali pratiche sembrano però relativamente poco diffuse nella valutazione delle traduzioni automatiche, anche se purtroppo manca ancora un'analisi approfondita della situazione.

4 La valutazione della qualità dei testi generati

Se molti decenni di riflessioni sull'argomento sono arrivati a concludere che non esiste un modo unico e oggettivo per valutare la qualità di una traduzione, automatica o meno, si può ipotizzare che non esista nemmeno un modo unico e oggettivo per valutare la qualità di un testo generato. Di sicuro, è possibile partire da questa ipotesi per esaminare alcuni tipi di valutazione.

Per i testi scritti esistono naturalmente dei parametri oggettivi: per esempio, il rispetto delle regole specifiche di determinate varietà di lingua. Tuttavia, la qualità di un testo non dipende solo dalla sua correttezza ortografica e sintattica, e la distinzione tra i diversi livelli dovrebbe essere sempre descritta con attenzione. La valutazione può riguardare infatti aspetti più strettamente grammaticali (come il rispetto degli standard ortografici o morfosintattici), ma anche aspetti comunicativi: adeguatezza alla situazione e al destinatario, gestione della variazione di registro e così via. Anche in questo caso, un certo approfondimento delle difficoltà è utile.

4.1 I limiti nella valutazione dei testi scritti da esseri umani

In molti settori della produzione testuale si dà implicitamente per scontata la soggettività della valutazione e la difficoltà o l'impossibilità di formalizzare criteri in proposito. Al livello più alto, per esempio, tipicamente i premi letterari si basano in sostanza su valutazioni soggettive, discusse e discutibili, di persone che spesso non sono professioniste della valutazione. Non mi risulta invece che, per la lingua e la letteratura italiana, esistano premi letterari che vengano assegnati sulla base di un punteggio assegnato in base a un sistema esplicito e dettagliato.

In generale, per la lingua italiana L1 non esiste una griglia di valutazione di riferimento che possa essere applicata per esempio a un testo professionale. Non esiste inoltre nemmeno un quadro condiviso, ma solo una varietà di pratiche a volte in contraddizione tra di loro. La sistematicità di ogni valutazione, così come la sua confrontabilità, è quindi molto ridotta, perfino nei casi di maggiore rilevanza pratica.

Nell'Italia contemporanea, un esempio vistoso di questo stato di cose è fornito dalle discussioni sugli esami finali delle scuole secondarie superiori italiane. All'interno di queste prove, in cui la scrittura di testi nell'esame finale ha un ruolo importante, gli studenti delle regioni del sud ottengono infatti ogni anno voti più alti rispetto agli studenti delle regioni del nord, anche se i livelli medi di competenza misurati durante il percorso scolastico hanno una distribuzione geografica opposta. Anche trascurando le complesse dinamiche sociali dietro a una situazione del genere, è interessante notare che i mezzi di comunicazione italiani insistono con regolarità su questa discrepanza e che il governo italiano ha dichiarato in modo esplicito la volontà di uniformare le pratiche di valutazione. In attuazione della legge 105 del 2015 è stato quindi emanato nel 2017 il decreto legislativo 62 che al comma 6 dell'articolo 17 prevedeva appunto la creazione di una griglia di valutazione nazionale per le prove. La griglia è stata poi presentata con il decreto ministeriale 769 del 26 novembre 2018, ma è risultata talmente generica da rendere difficile immaginare che potesse avere un qualunque effetto di uniformazione (Cervesato 2019). Del resto, la complessità della valutazione della produzione è senz'altro uno dei fattori per cui molte importanti campagne internazionali di valutazione delle competenze non prendono in esame la capacità di produrre testi.

Questa ampia premessa è necessaria per inquadrare un punto: la competenza nella produzione scritta di alto livello in italiano L1 oggi non può essere valutata bene su scale che siano contemporaneamente approfondite, oggettive e condivise. In particolare, oltre un certo livello (livello che forse si può descrivere come "sorprendentemente basilare"), la valutazione è soggettiva e parziale – anche in verosimile rapporto con la rilevanza sociale delle pratiche. Ciò non vuol dire che la valutazione sia impossibile; vuol dire però che ogni istanza deve essere definita in modo preciso, indicando ciò che si valuta, chi valuta, in quali contesti e con quali limiti.

Sarebbe per esempio impossibile dire che il testo generato da un sistema come ChatGPT è al livello "di un elaborato di un esame di maturità italiano", o che "all'esame di maturità otterrebbe 17/20". Anche in assenza di dati pubblicati attendibili, è evidente che i diversi valutatori degli esami di maturità usano criteri diversi: sembra perfettamente possibile che una commissione valuti 14 ciò che un'altra commissione potrebbe valutare 18, e così via. Viceversa, non sarebbe impossibile (con diversi complessi accorgimenti pratici) sottoporre a una specifica commissione un elaborato generato e confrontare la valutazione con quella degli elaborati di studenti valutati nelle stesse circostanze dalla stessa commissione. Concepire la valutazione in questo modo delimitato, invece che in riferimento a

standard inesistenti, permette senz'altro di inquadrare la questione in modo più produttivo.

L'esempio di un esame di maturità ha anche un altro vantaggio: permette di ricordare che valutazioni simili, al momento, possono essere fatte solo da persone ("professori") con esperienza nella valutazione di elaborati simili (anche se oggi la realizzazione di sistemi informatici dedicati sembra come minimo possibile, nel senso presentato nella sezione 2). Allo stesso modo, la valutazione di testi di tipo professionale per l'editoria può essere fatta solo da redattori umani, con competenze nella redazione, valutazione e correzione di testi professionali.

4.2 La varietà degli oggetti da valutare

Valutare la generazione di testi mette in primo piano una domanda che nel caso della traduzione automatica ha avuto storicamente un ruolo marginale: in quali varietà o generi testuali deve rientrare il testo generato? Le verifiche di base mostrano infatti che i sistemi attuali – a differenza di quelli del recente passato – hanno una capacità di alto livello di generare testi rispettando per esempio le regole grammaticali di lingue come l'inglese o l'italiano, ma non riescono a fare altrettanto per lingue meno documentate, dialetti o specifiche varietà di lingua.

Giuseppe Antonelli (2023: 11) ha per esempio notato che a fine 2023 ChatGPT rispettava bene le regole della lingua italiana ma produceva errori grammaticali vistosi nella creazione di un testo in romanesco, usando per esempio sistematicamente come articolo determinativo *'l* invece del corretto *er*. Oppure, se a inizio 2024 si chiede a ChatGPT-4.0 di scrivere un sonetto sul tema "sentirsi oppressi dalla quantità di cose da fare" si può ottenere un prodotto che si avvicina al risultato desiderato ma che, per esempio, non rispetta gli schemi di rime della tradizione letteraria italiana e colloca in posizione di rima parole che in realtà non sono in rima, come *sòffocano* e *fanno*:

In questo mare vasto di doveri,
onde di impegni mi soffocano,
l'orologio gira, veloce, e oneri
si accumulano, inquieti mi fanno.

L'anima mia, stretta in ferree sbarre,
lotta per aria in questo frenetico volo,
ogni compito, una catena che mi sgarruppa,
in questo vortice, perdo il mio solo.

Ma in questo caos, una luce brilla,
un momento di pace, un dolce sfuggire,
nella quiete, la mente si distilla.

Forse in questa lotta, posso imparare,
a trovare equilibrio, a non soccombere,
nella tempesta, a navigare, a osare.

Tutto ciò sembra in evidente rapporto con la quantità di materiale usata per l'addestramento. Ma, al di là di ogni ipotesi di dettaglio sui meccanismi di generazione, è del tutto ragionevole prevedere che la qualità del testo generato sia diversa per le diverse varietà dell'italiano (descritte ora in Ballarè, Fiorentini e Miola 2024). Di conseguenza, andrebbe valutata separatamente la capacità di scrivere in italiano standard o neostandard, in italiano burocratico o italiano colloquiale (secondo le classificazioni di Berruto 2012), per non parlare poi della capacità di generare testi nel dialetto napoletano o nell'italiano regionale di Sardegna. Una naturale prosecuzione di questa suddivisione dovrebbe poi coinvolgere il livello successivo, relativo ai generi testuali: andrebbe infatti valutata la capacità di scrivere, oltre ai sonetti, comunicati stampa, tweet o telegrammi, e così via, per le decine o centinaia di generi le cui caratteristiche sono oggi note anche alle persone di media cultura. Né va trascurata la variazione diacronica: in fin dei conti, i sistemi hanno una certa, ma senz'altro limitata, capacità di imitare il fiorentino trecentesco o l'italiano del Cinquecento. Per esempio, si può chiedere a ChatGPT-4 di riformulare la celebre conclusione della *Questione privata* di Beppe Fenoglio in uno stile diverso (notando che il sistema estende il suo tentativo di imitazione linguistica anche all'interazione con l'utente):

D: Puoi riscrivermi questo testo nello stile di Pietro Bembo?

"Correva, con gli occhi sgranati, vedendo pochissimo della terra e nulla del cielo. Era perfettamente conscio della solitudine, del silenzio, della pace, ma ancora correva, facilmente, irresistibilmente. Poi gli si parò davanti un bosco e Milton vi puntò dritto. Come entrò sotto gli alberi, questi parvero serrare e far muro e a un metro da quel muro crollò."

R: Nel tessere le parole con la maniera che Pietro Bembo avrebbe potuto eleggere, un autore rinascimentale che con la sua opera "Prose della volgar lingua" plasmò il corso della lingua italiana, orientando essa verso una purità stilistica ed eleganza che riecheggia i classici latini, si potrebbe riforgiare il testo nel seguente modo:

Fuggiva, con le pupille dilatate, scorgendo assai poco del terreno e niente del cielo. Era pienamente conscio della solitudine, del silenzio, della quiete, tuttavia ancora fuggiva, con leggerezza, irresistibilmente. Poi, d'improvviso, si presentò davanti a lui un bosco e Milton vi si avventò senza esitazione. Nel momento in cui penetrò sotto le fronde, queste sembrarono chiudersi e formare una barriera e a breve distanza da quella barricata egli crollò.

Questo tentativo di imitazione mira a riflettere la predilezione di Bembo per una lingua raffinata, che aspira alla chiarezza, all'armonia e alla precisione, elementi che il Bembo sosteneva dovessero caratterizzare il volgare illustre, lingua in cui le opere letterarie dovessero essere composte per raggiungere un pubblico colto e raffinato. La scelta delle parole e la struttura della frase cercano di avvicinarsi allo stile letterario del Rinascimento, periodo durante il quale Bembo visse e operò, influenzando profondamente la letteratura e la cultura italiana.

Esercizi del genere potrebbero naturalmente essere ripetuti per tutta l'immensa gamma di variazione della lingua. Inoltre, sembra difficile prevedere in dettaglio le

capacità in ognuno di questi compiti sulla base di quelle misurate in altri, al di là dell'ipotesi generale che venga elaborato meglio ciò che è stato meglio rappresentato nei corpora di addestramento.

Tuttavia, si può decidere di concentrare la valutazione in base alla rilevanza. È infatti chiaro che un'area di particolare interesse è quella dell'espressione per così dire naturale del sistema: il testo prodotto quando l'utente non fornisce indicazioni precise sulla lingua da usare. Operativamente e intuitivamente, per evitare l'influenza delle classificazioni correnti, per i testi che saranno esaminati qui di seguito è stato quindi richiesto a ChatGPT di esprimersi a un "livello medio di formalità". La varietà di lingua prodotta con questa impostazione sembra in sostanza corrispondente alla lingua usata per gli articoli settoriali di quotidiani o periodici; le deviazioni da quest'ottica saranno viste come quelle che una persona con alle spalle studi linguistici contemporanei o un redattore esperto possono verosimilmente individuare e correggere.

4.3 Considerazioni preliminari alla valutazione di ChatGPT

Una volta constatato che l'italiano "a livello medio di formalità" di ChatGPT è all'ingrosso paragonabile all'italiano normalmente usato in alcuni tipi di testo, un punto di riferimento per la valutazione può essere il rispetto delle regole grammaticali previste in questa situazione. Occorre quindi assegnare il giusto peso alle variazioni rispetto allo standard.

È opportuno ricordare che l'evoluzione dei sistemi di generazione del testo rende molto difficile l'uso di un'unica scala per i sistemi del recente passato e per quelli contemporanei, così come è molto difficile usare un'unica scala per valutare elaborati di scuola elementare e tesi di laurea. Per vedere quanto siano cambiate le cose nel giro di pochi anni può essere utile tornare rapidamente alla traduzione con questo esempio del miglioramento spettacolare che si è avuto nel giro di pochi anni nella traduzione dall'inglese all'italiano di una frase tratta da un articolo di rivista:

Testo originale: The murder defendant, James Bates, agreed late Monday to allow Amazon to forward his Echo's data to Arkansas prosecutors.

Traduzione Microsoft Translator 2017: L'accusato di omicidio, James Bates, concordato late Lunedì consentire Amazon inoltrare i dati di sua Echo ai pubblici ministeri Arkansas.

Traduzione ChatGPT 2023: L'imputato per omicidio, James Bates, ha accettato lunedì sera di permettere ad Amazon di trasmettere ai procuratori dell'Arkansas i dati del suo Echo.

Nel caso del testo generato, la valutazione degli errori rispetto a una varietà di lingua di riferimento può essere un importante ancoraggio, sia di per sé sia in combinazione con altri elementi. Lo studio degli errori di apprendenti mostra peraltro che non è sempre facile descrivere in modo oggettivo quali e quanti errori siano presenti in un testo (Andorno e Rastelli 2009; Rosi 2009). Anche una valutazione approssimativa è però utile per iniziare un processo di taratura.

Particolarmente interessante, anche a fini di valutazione, sembra poi il distinguere tra gli errori che potrebbero essere commessi anche da esseri umani e quelli che invece è molto difficile che un essere umano commetta. Nella generazione di testo per esempio si ottengono occasionalmente frasi come questa, generata da ChatGPT-3.5 nel settembre 2023: “La "lingua franca" è un termine che deriva dall’antico italiano "lingua franca" che significa "lingua franca" o "lingua franca”.

La comparsa occasionale di frasi del genere nei testi da valutare è interessante dal punto di vista del confronto con le produzioni umane di alto livello, in cui strutture di questo tipo sono decisamente rare. Inoltre, ha anche una conseguenza pratica immediata sulla valutazione: è infatti sufficiente che all’interno di un testo lungo compaia anche una singola frase di questo tipo per far sospettare (correttamente) al lettore, e al valutatore, che tutto il testo sia stato generato da un sistema di intelligenza artificiale. Questo ha ovvi riflessi sull’oggettività della valutazione nel caso si voglia fare per esempio un confronto con la produzione di esseri umani, senza dare al valutatore informazioni sulla provenienza dei testi.

Per quanto riguarda poi le caratteristiche generali di questo e di altri sistemi simili, un’introduzione è stata fornita da Francesco Cicero (2023), che ha descritto diverse tendenze significative comuni agli LLM oggi in uso. Il lavoro ha mostrato tra l’altro che le risposte a domande in lingua italiana sono oggi spesso contenutisticamente inferiori rispetto a quelle a domande poste in lingua inglese, anche per quanto riguarda argomenti collegati alla lingua e alla cultura italiana.

5 ChatGPT a un “livello medio di formalità”

Qui di seguito si analizzeranno le risposte fornite da ChatGPT-3.5 e ChatGPT-4 durante una serie di interrogazioni condotte attraverso l’interfaccia web nel settembre del 2023. Per quanto i sistemi di interrogazione siano ormai ben noti, è forse utile ripetere che l’interfaccia web del sistema simula una conversazione, permettendo di presentare richieste cui ChatGPT replica rivolgendosi a volte direttamente all’utente prima di fornire la risposta vera e propria. Per esempio, un tipico scambio, con domanda dell’utente e inizio di risposta, è:

D: Quale differenza c’è tra "diritto d’autore" e "copyright"?

R: Mirko, "diritto d’autore" e "copyright" sono due concetti simili ma spesso utilizzati in contesti diversi, e la principale differenza risiede nell’uso linguistico e nella legislazione che li riguarda.

5.1 Descrizione del corpus

L’indicazione delle varietà di lingua e dei livelli di formalità richiesti può essere fornita al sistema in vario modo. Per i lavori descritti qui, attraverso il menu

Istruzioni personalizzate, in risposta alla domanda “Come vorresti che ChatGPT rispondesse?”, ho dato al sistema queste indicazioni in lingua inglese:

Medium level of formality.
 ChatGPT should remain neutral.
 Responses should be long.
 I want to be addressed as "Mirko".
 Do not use bullet points.

‘Livello medio di formalità.
 ChatGPT deve rimanere neutro.
 Le risposte devono essere lunghe.
 Voglio essere chiamato "Mirko".
 Non usare punti elenco.’

L’ultima indicazione serve per evitare l’impiego dei punti elenco, che ChatGPT usa normalmente per la presentazione di risposte complesse. L’obiettivo è stato quello di avvicinarsi ai modelli tipici della scrittura in lingua italiana; tuttavia, come si vedrà, ChatGPT segmenta comunque fortemente il testo.

Mantenendo le impostazioni indicate, ho poi portato avanti due conversazioni, una con ChatGPT-3.5 e una con ChatGPT-4. Nella prima ho seguito l’andamento della conversazione, riprendendo anche le parole usate dal sistema per rispondere; nella seconda ho riproposto al sistema le stesse 7 domande della prima, indipendentemente dall’andamento della conversazione. Le domande sono state quindi, in entrambi i casi:

1. Quale differenza c’è tra "diritto d’autore" e "copyright"?
2. Non ho capito: potresti spiegarmi la differenza con altre parole?
3. Quali sono le tecniche più efficaci per l’insegnamento della scrittura a livello universitario?
4. Quali modelli di scrittura di alta qualità si possono dare a questi studenti?
5. Puoi indicarmi un saggio accademico ben scritto?
6. Quali sono le caratteristiche di una scrittura di alta qualità?
7. Che cosa succede se scrivo un testo con molti errori ortografici?

I testi generati nel corso delle conversazioni hanno formato due piccoli corpora, rispettivamente di circa 4000 e 3000 token, con queste caratteristiche linguistiche, misurate con il già citato READ-IT:

Tabella 3: Statistiche linguistiche generate da READ-IT.

	ChatGPT-3.5	ChatGPT-4
GULPEASE	48,4	52,8
Numero totale periodi:	252	179
Numero totale parole (token):	4067	3100
Lunghezza media dei periodi (in token):	16,1	17,3
Lunghezza media delle parole (in caratteri):	5,2	5,3
Percentuale di lemmi appartenente al Vocabolario di Base (VdB):	67,8%	66,1%

<i>Ripartizione dei lemmi riconducibili al VdB rispetto ai repertori d'uso:</i>		
Fondamentale:	77,4%	76,6%
Alto uso:	19,3%	19,6%
Alta disponibilità:	3,3%	3,8%
Rapporto tipo/unità (calcolato rispetto alle prime 100 parole del testo):	0,660	0,610
Densità lessicale:	0,555	0,592
<i>Profilo sintattico</i>		
Sostantivi:	22,2%	21,1%
Nomi propri:	2,5%	5,5%
Aggettivi:	9,1%	9,4%
Verbi:	13,2%	12,5%
Congiunzioni:	5,5%	5,3%
Di cui coordinanti:	83,1%	81,7%
Di cui subordinanti:	16,9%	18,3%
<i>Struttura sintattica a dipendenze</i>		
<i>Articolazione interna del periodo</i>		
Numero medio di proposizioni per periodo:	1,702	1,737
<i>Proposizioni principali e subordinate (%)</i>		
Principali:	69,4%	66,7%
Subordinate:	30,6%	33,3%
<i>Articolazione interna della proposizione</i>		
Numero medio di parole per proposizione:	9,480	9,968
Numero medio di dipendenti per testa verbale:	1,828	1,945
<i>Misura della profondità dell'albero sintattico</i>		
Media delle altezze massime:	5,637	5,278
Profondità media di strutture nominali complesse:	1,270	1,250
Profondità media delle catene di subordinazione:	1,154	1,214
<i>Misura della lunghezza delle relazioni di dipendenza (calcolata come distanza in parole tra testa e dipendente):</i>		
Lunghezza media:	2,201	2,215
Media delle lunghezze massime:	6,020	6,838

Anche sulla base dei dati statistici, e in particolare del rapporto tra sostantivi e verbi, il testo è caratterizzabile come un normale testo espositivo professionale, lontano dalle caratteristiche del parlato. Al suo interno sono inoltre presenti anche tratti standard, anziché neostandard, come l'uso del pronome *loro* invece di *gli*, come per esempio in questa frase: “Se appropriato per il corso, mostri *loro* esempi di scritti creativi che sono stati premiati o riconosciuti per la loro qualità”.

Un esempio di risposta alla domanda “Che cosa succede se scrivo un testo con molti errori ortografici?” è stato:

Se scrivi un testo con molti errori ortografici, ci sono diverse conseguenze potenziali, a seconda del contesto e del pubblico. Ecco alcuni degli effetti che potrebbero emergere:

1. Perdita di Credibilità

Gli errori ortografici possono mettere in dubbio la tua competenza e professionalità, riducendo così la tua credibilità agli occhi del lettore. Questo è particolarmente critico in contesti formali come la scrittura accademica, i documenti professionali o la corrispondenza d'affari.

2. Distrazione per il Lettore

Gli errori possono distogliere l'attenzione del lettore dal contenuto principale del tuo testo. Invece di concentrarsi sulle tue idee o argomenti, il lettore potrebbe essere distratto nel cercare di decifrare il significato alterato dalle imprecisioni ortografiche.

3. Ambiguità e Confusione

Gli errori ortografici possono rendere il tuo testo ambiguo, rendendo difficile per il lettore comprendere il messaggio che stai cercando di trasmettere. In alcuni casi, un errore ortografico potrebbe cambiare completamente il significato di una frase.

4. Impatto sul Pubblico Obiettivo

Se il tuo pubblico è particolarmente attento alla precisione, come accademici o professionisti in un determinato campo, gli errori ortografici potrebbero portare a un giudizio negativo sul tuo lavoro nel suo complesso.

5. Problemi nelle Comunicazioni

In contesti in cui la chiarezza e la precisione sono fondamentali, come istruzioni mediche, contratti legali o documentazione tecnica, gli errori ortografici possono avere conseguenze serie, incluse incomprensioni che potrebbero portare a errori pratici.

6. Impatto SEO

Se stai scrivendo per una piattaforma online, gli errori ortografici possono influenzare negativamente il posizionamento del tuo contenuto nei risultati di ricerca, dato che i motori di ricerca come Google utilizzano la qualità del contenuto come uno dei fattori per determinare la rilevanza di una pagina.

7. Costi di Correzione

Se il testo deve essere pubblicato o stampato, gli errori ortografici possono portare a costi aggiuntivi per la correzione e la ristampa del materiale.

Per evitare queste conseguenze, è consigliabile dedicare tempo alla revisione e alla correzione del testo. L'uso di strumenti di correzione ortografica e grammaticale può aiutare, ma una revisione manuale è spesso necessaria per catturare errori che tali strumenti potrebbero non rilevare. Se possibile, avere un secondo paio di occhi per revisionare il tuo lavoro può fornire ulteriori garanzie contro gli errori.

Come si vede, il testo viene comunque segmentato, se non come lista puntata, come lista numerata. Tutte le risposte sono formulate in questo modo, e una simile suddivisione aiuta senz'altro la generazione (come notato per esempio da Cicero 2023: 753), mentre l'insegnamento scolastico italiano spesso mostra reticenza all'uso di queste strutture, poco diffuse nella prosa tradizionale. La coerenza del testo è comunque buona.

5.2 Errori linguistici

A un esame più ravvicinato si nota che i testi del corpus sono esempi di italiano standard o neostandard *quasi* perfettamente corretto dal punto di vista grammaticale e sintattico. Alcuni errori sono infatti presenti a diversi livelli, incluso quello ortografico. Per esempio, nei testi generati da ChatGPT-3.5 compare un'occorrenza di “un'autore”, con articolo indeterminativo apostrofato che precede un sostantivo maschile. Questo è un errore soggetto a forte sanzione sociale, ma nella realtà piuttosto frequente nella scrittura non revisionata anche di persone istruite (manca però, a oggi, una quantificazione).

Un altro aspetto di divergenza riguarda l'impiego della *-d* eufonica. Questa, oggi, nella scrittura professionale italiana, è usata solo tra vocali uguali; quando però le vocali sono uguali, va usata obbligatoriamente. Nel testo generato, invece, la *-d* eufonica a volte viene collocata anche tra vocali diverse, come nella sequenza “ad essere”; a volte viene evitata dove sarebbe necessaria, come nella sequenza “e Editoriali” (anche se a volte la scrittura professionale evita l'uso davanti a parola che inizia con *ed-*). In particolare, il mancato uso della *-d* eufonica si presenta per 5 volte nel testo generato da ChatGPT-3.5 e 2 volte nel testo generato da ChatGPT-4. Anche questo genere di incertezze è normale per gli esseri umani che non siano professionisti della scrittura o della revisione, e corrisponde in effetti a pratiche ampiamente diffuse. Nella stessa categoria rientra il fatto che nel testo generato da ChatGPT-3.5, contro la nota raccomandazione data da Luca Serianni a proposito dell'accentazione del pronome *sé* anche nella sequenza “sé stessi” e simili, compare la frase “in modo che gli studenti possano esporre *se* stessi a stili di scrittura diversi”; la grafia senza accento, oggetto di una norma trasmessa per lungo tempo dalla tradizione scolastica, è però talmente radicata e comune che è difficile considerare il suo impiego come un semplice errore – in effetti, è possibile che questa grafia sia ancora oggi prevalente anche nella scrittura colta.

Nel testo generato da ChatGPT-4 sono poi presenti due errori, a cavallo tra ortografia e lessico, che difficilmente un essere umano madrelingua potrebbe commettere: l'uso di *clarezza* invece di *chiarezza*, in evidente rapporto con l'ortografia della parola inglese *clarity*, e quello di *jargon* invece di *gergo* (nella frase “l'uso di un linguaggio complicato o di jargon non dovrebbe ostacolare la chiarezza”).

Per quanto riguarda la morfosintassi, ChatGPT-3.5 commette un solo errore nella gestione dei verbi. In un contesto in cui si rivolge al destinatario con il *tu*, all'interno di una frase genera una forma corrispondente alla seconda persona del presente indicativo o congiuntivo, *mostri*, invece del corretto imperativo *mostra*: “Se appropriato per il corso, *mostri* loro esempi di scritti creativi che sono stati premiati o riconosciuti per la loro qualità”.

Sempre per quanto riguarda la morfosintassi, ChatGPT-4 commette due errori. Sbaglia innanzitutto la gestione di singolare e plurale nella frase “Il termine “diritto d'autore” e “copyright” sono spesso utilizzati in modo intercambiabile” (per esseri umani con buone capacità di scrittura, questo sarebbe un normale errore di

esecuzione). Usa poi l'espressione "il lettore potrebbe essere distratto *nel* cercare di decifrare il significato alterato dalle imprecisioni ortografiche", in un contesto in cui l'uso italiano preferisce senz'altro una costruzione del tipo "potrebbe essere distratto dalla necessità" o "potrebbe distrarsi nel cercare" (per esseri umani, questo sarebbe un tipico errore di competenza).

5.3 Calchi dall'inglese

In sostanza, la maggior parte degli errori descritti fino a questo momento potrebbe comparire anche in un testo scritto da un essere umano con buone competenze linguistiche o addirittura da uno specialista, ma non revisionato – anche se i due errori ortografici riscontrati per ChatGPT-4 non rientrano in questa categoria. Non solo: anche dal punto quantitativo, su un corpus di questa dimensione, il numero di errori è paragonabile a quello che potrebbe trovarsi nella prima stesura di un testo di buon livello redatto da essere umano con buone competenze linguistiche o addirittura da uno specialista, ma non revisionato.

Nel bilancio non è stata però ancora inserita la categoria cui appartengono le deviazioni più consistenti: i calchi, cioè gli impieghi di parole o espressioni italiane nel senso che hanno le corrispondenti parole o espressioni inglesi. Ne è un esempio una frase come "Incoraggia anche la creatività nella scrittura, quando appropriato, per sviluppare uno stile *unico*", in cui in italiano si userebbe piuttosto l'aggettivo *personale*: è evidente il calco sull'aggettivo inglese *unique*, che ha appunto anche questo significato.

A volte questi calchi si configurano come veri e propri errori di traduzione. A monte si trova evidentemente, in qualche modo, un modello in lingua inglese, che viene trasposto in italiano usando parole o espressioni simili ma che in realtà hanno un significato tanto diverso da quello italiano da cambiare sostanzialmente il senso del testo. Per esempio, si vedano i calchi riportati nella Tabella 4:

Tabella 4: Calchi nella produzione di ChatGPT.

Testo di ChatGPT	
Mi dispiace, Mirko, ma non posso fornire saggistica specifica o estratti di saggistica a causa delle <i>restrizioni di copyright</i> .	Modello verosimile in inglese: <i>copyright restrictions</i> Traduzione corretta: <i>restrizioni dovute al copyright /</i> <i>restrizioni dovute al diritto d'autore</i>
Tenendo presenti queste caratteristiche e <i>praticando regolarmente</i> , gli studenti universitari possono sviluppare abilità di scrittura di alta qualità che li aiuteranno nella loro carriera accademica e professionale	Modello verosimile in inglese: <i>regularly practicing</i> Traduzione corretta: <i>facendo regolarmente esercizio</i>
Negli Stati Uniti, per esempio, la registrazione offre vantaggi legali, come la possibilità di	Modello verosimile in inglese: <i>statutory damages</i>

rivendicare *danni statutari* e spese legali in caso di violazione.

Traduzione corretta:

danni tabellari / compensi fissi per danni

Un esempio particolarmente vistoso, visti gli argomenti della conversazione, è quello dell'aggettivo *accademico*, che viene usato complessivamente 49 volte (29 per ChatGPT-3.5, 20 per ChatGPT-4), spesso in contesti in cui non è appropriato. Tale aggettivo, infatti, ha in italiano tra i suoi significati anche quello di “relativo all'università”, ma, a differenza dell'inglese, non quello di “scientifico” in generale, che non è registrato dal *GDLI* e dagli altri principali dizionari contemporanei. In linea con questo uso, fino a pochi anni fa erano rare espressioni come “riviste accademiche”, ma si parlava regolarmente di “riviste scientifiche”; il calco si è però senz'altro diffuso in tempi più recenti. Nel corpus, come è naturale, in molti casi l'uso di *accademico* è accettabile, ma in almeno 26 casi su 49 è decisamente poco accettabile. Un contesto in cui *accademico* andrebbe senz'altro sostituito da *scientifico* è costituito da queste frasi generate da ChatGPT-4: “La pubblicazione in una rivista accademica rinomata è spesso un segno di alta qualità. Riviste come "Nature", "Science", e "The Lancet" sono alcune delle più prestigiose nelle scienze naturali e mediche.” Altri esempi di calchi a diverso livello possono essere (sempre da ChatGPT-4):

Articoli scientifici sottoposti a revisione paritaria sono spesso considerati la “barra d'oro” per la scrittura accademica. *Essi* offrono esempi di come presentare argomenti complessi in modo chiaro e strutturato.

Saggi ben scritti da critici noti o accademici offrono esempi di *come* analizzare opere d'arte, letteratura o fenomeni sociali. *Questi* possono illustrare come utilizzare *evidenze* e argomentazioni per sostenere un punto di vista.

Colpiscono i calchi nei sostantivi, come per esempio *evidenze* (su cui: Gualdo 2020); tuttavia, colpiscono anche calchi ad altri livelli, come l'uso generalizzato di *come* (sulla base evidente di *how / as*) in contesti in cui l'italiano userebbe per esempio “del modo in cui”, “il modo in cui”, e simili; oppure, l'abbondante impiego di pronomi secondo il modello inglese, che porta a usare per esempio un pronome oggi raro in italiano come *essi*.

La quantificazione di tali fenomeni è molto soggettiva, ma al mio conto nei due corpus i calchi da modificare, e che spesso possono essere considerati veri e propri errori, sono in totale 132: un numero decisamente elevato, corrispondente a circa un'occorrenza ogni 53 token. Va però notato che un simile livello sembra paragonabile (anche dal punto di vista quantitativo) a quello che si può riscontrare nelle traduzioni fatte da non professionisti.

Inoltre, nei testi generati questi errori si incrociano al fatto che le informazioni fornite a volte non sono applicabili al contesto italiano, ma sembrano semplici traduzioni di informazioni relative al contesto statunitense (secondo un meccanismo notato anche da Cicero 2023: 737). Per esempio, alla richiesta di

esempi di scrittura universitaria, ChatGPT risponde con frasi come: “Molti dipartimenti universitari o docenti forniscono saggi campione di alta qualità scritti da studenti. Questi possono essere particolarmente utili perché sono direttamente pertinenti al tipo di scrittura che gli studenti dovranno svolgere.” In realtà, il tipo di pubblicazione descritto sembra ben poco comune in Italia: è possibile che non sia praticato da nessun dipartimento universitario o docente, e certamente non è praticato da “molti” di essi. Oppure, come modelli di scrittura, ChatGPT indica generi testuali che non esistono nella cultura editoriale italiana, come gli “Op-ed”, i cosiddetti “editoriali ospiti” dei quotidiani statunitensi.

Caratteri di questo genere avvicinano senz’altro il testo a produzioni di studenti che rielaborino un po’ meccanicamente fonti in lingua inglese per presentare informazioni in italiano. Un confronto sistematico potrebbe essere utile per individuare in modo più preciso somiglianze e differenze.

6 Conclusioni e possibili sviluppi

L’esame dei testi generati da ChatGPT mostra che al momento il sistema non è in grado di scrivere testi completamente corretti neanche dal punto di vista formale, anche se si avvicina molto a questo obiettivo. Per ottenere la correttezza formale è necessaria comunque una revisione umana. Più difficile è valutare l’aspetto quantitativo, cioè il livello di revisione richiesto.

Entrare nel dettaglio di questi aspetti quantitativi richiede però un confronto con lo stato attuale delle pratiche di valutazione, così come descritto nelle sezioni iniziali del contributo. L’esistenza di pratiche di valutazione sofisticate per diversi tipi di testo non implica purtroppo l’immediata applicabilità delle stesse pratiche a un settore come la valutazione dei prodotti dei testi generati. Per questo, sembra utile soprattutto una discussione interdisciplinare che consenta di mettere a frutto competenze di tipo diverso, in modo da poter tenere appropriatamente conto di caratteristiche come la possibilità di comparsa di frasi prive di senso all’interno di contesti generalmente corretti.

Dal punto di vista linguistico, la valutazione dell’italiano prodotto da questi sistemi ha però forse un pregio ancora maggiore: permette di riflettere sulla valutazione dell’italiano scritto L1 in generale. Una tale riflessione sulla qualità avrebbe senz’altro un notevole interesse scientifico.

Bibliografia

- Alva-Manchego, Fernando & Scarton, Carolina & Specia, Lucia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics* 47 (4). 861-889. https://doi.org/10.1162/coli_a_00418 (ultimo accesso 7 febbraio 2024)
- Andorno, Cecilia & Rastelli, Stefano. 2009. Un’annotazione orientata alla ricerca acquisizionale. In Andorno, Cecilia & Rastelli, Stefano (eds), *Corpora di italiano L2: tecnologie, metodi, spunti teorici*, 49-70. Perugia: Guerra Edizioni.

- Antonelli, Giuseppe. 2023. L'IA-taliano. *La Lettura, Corriere della Sera*, 3 dicembre 2023, 10-11.
- Bender, Emily M. & Gebru, Timnit & McMillan-Major, Angelina & Shmitchell Shmargaret. 2021. On the dangers of stochastic parrots: can language models be too big? 🦜. In *FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (March 3-10, 2021), 610-623. New York: Association for Computing Machinery.
<https://doi.org/10.1145/3442188.3445922> (ultimo accesso 31 gennaio 2024)
- Bentivogli, Luisa & Bisazza, Arianna & Cettolo, Mauro & Federico, Marcello. 2018. Neural versus phrase-based MT quality: an in-depth analysis on English–German and English–French. *Computer speech and language* 49. 52-70.
- Berruto, Gaetano. 2012. *Sociolinguistica dell'italiano contemporaneo*. Seconda edizione. Roma: Carocci.
- Cervesato, Ivan. 2019. Valutazioni uniformi all'Esame di Maturità: la beffa delle griglie nazionali. *Roars*, 20 marzo 2019. <https://www.roars.it/valutazioni-uniformi-allesame-di-maturita-la-beffa-delle-griglie-nazionali/> (ultimo accesso 31 gennaio 2024)
- Cicero, Francesco. 2023. L'italiano delle intelligenze artificiali generative. *Italiano LinguaDue* 15 (2). 733-761.
- Consiglio d'Europa. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press. <https://rm.coe.int/1680459f97> (ultimo accesso 31 gennaio 2024)
- De Mauro, Tullio. 1980. *Guida all'uso delle parole*. Roma: Editori Riuniti.
- De Mauro, Tullio. 2016. Il Nuovo vocabolario di base della lingua italiana. *Internazionale*, 23 dicembre 2016.
<https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana> (ultimo accesso 31 gennaio 2024)
- Dell'Orletta, Felice & Montemagni, Simonetta & Venturi, Giulia. 2011. READ-IT: assessing readability of Italian with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies* (Edinburgh, July 2011), 73-83. Stroudsburg: Association for Computational Linguistics.
- Drugan, Joanna. 2013. *Quality in Professional Translation: Assessment and Improvement*. London, New York: Bloomsbury.
- Freedman, David H. 2010. *Wrong: Why experts* keep failing us – and how to know when not to trust them* Scientists, finance wizards, doctors, relationship gurus, celebrity CEOs, high-powered consultants, health officials and more*. New York: Little, Brown and Company.
- Fulcher, Glenn & Harding, Luke (eds). 2021. *The Routledge handbook of language testing. Second edition*. London: Routledge.
- Gallina, Francesca & Orlando, Salvatore & Tavosanis, Mirko. In stampa. Le raccomandazioni per la scrittura universitaria: manuali, declaratorie, obiettivi

- formativi. *Univers-ITA. La scrittura degli studenti universitari*. Bologna: Bononia University Press.
- Giunta, Claudio. 2018. *Come non scrivere*. Torino: UTET.
- Graham, Yvette & Baldwin, Timothy. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the EMNLP* (Doha, Qatar, October 25-29, 2014), 172-176. Stroudsburg: Association for Computational Linguistics. <https://aclanthology.org/D14-1020.pdf> (ultimo accesso 31 gennaio 2024)
- Gualdo, Riccardo. 2020. È possibile *evidenziare* le *evidenze*? Storia e fortuna di un calco anglicizzante. *Accademia della Crusca – Lingua italiana – Consulenza linguistica – Risposte ai quesiti*. <https://accademiadellacrusca.it/it/consulenza/%C3%A8-possibile-evidenziare-le-evidenze-storia-e-fortuna-di-un-calco-anglicizzante/2785> (ultimo accesso 31 gennaio 2024)
- Jimenez-Crespo, Miguel A. 2018. Crowdsourcing and translation quality: novel approaches in the language industry and translation studies. In Moorkens, Joss & Castilho, Sheila & Gaspari, Federico & Doherty, Stephen (eds), *Translation Quality Assessment: From Principles to Practice*, 69-94. Berlin: Springer.
- Khan, Salman. 2024. *Brave New Words: How AI Will Revolutionize Education (and Why That's a Good Thing)*. London: Viking.
- Krippendorff, Klaus. 2019. *Content Analysis*. London: SAGE.
- Kuiken, Folkert & Vedder, Ineke. 2022. Measurement of functional adequacy in different learning contexts. *TASK: Journal on Task-Based Language Teaching and Learning* 2,1. 8-32.
- Lemoine, Blake. 2023. 'I worked on Google's AI. My fears are coming true'. *Newsweek*, 27 febbraio 2023. <https://www.newsweek.com/google-ai-blake-lemoine-bing-chatbot-sentient-1783340> (ultimo accesso 31 gennaio 2024)
- Lu, Sheng & Bigoulaeva, Irina & Sachdeva, Rachneet & Madabushi, Harish Tayyar & Gurevych, Iryna. 2023. Are emergent abilities in large language models just in-context learning? *Arxiv*. <https://doi.org/10.48550/ARXIV.2309.01809> (ultimo accesso 31 gennaio 2024)
- Lucisano, Pietro & Piemontese, Maria Emanuela. 1988. Gulpease: una formula per la predizione della leggibilità di testi in lingua italiana. *Scuola e città* 3. 110-124.
- Mahowald, Kyle & Ivanova, Anna A. & Blank, Idan A. & Kanwisher, Nancy & Tenenbaum, Joshua B. & Fedorenko, Evelina. 2023. Dissociating language and thought in large language models. *Arxiv*. <https://doi.org/10.48550/ARXIV.2301.06627> (ultimo accesso 31 gennaio 2024)
- Marazzini, Claudio. 2023. Intervista di Marcello Pelizzari a Claudio Marazzini. *Corriere del Ticino*, 12 ottobre 2023. <https://www.cdt.ch/lifestyle/tecnologia/chatgpt-ed-evoluzione-della-lingua-dove-sta-andando-litaliano-329827> (ultimo accesso 31 gennaio 2024)

- Moorkens, Joss & Castilho, Sheila & Gaspari, Federico & Doherty, Stephen (eds). 2018. *Translation Quality Assessment: From Principles to Practice*. Berlin: Springer.
- NIST. 2009. *Tools / NIST*. <https://www.nist.gov/itl/iad/mig/tools> (ultimo accesso 7 febbraio 2024)
- Papa, Sirio & Tavosanis, Mirko. 2020. Valutazione umana di Google Traduttore e DeepL per le traduzioni di testi giornalistici dall'inglese verso l'italiano. In Dell'Orletta, Felice & Monti, Johanna & Tamburini, Fabio (eds), *CLiC-it 2020 – Proceedings of the Seventh Italian Conference on Computational Linguistics* (Bologna, March 1-3, 2021), 422-428. Torino: Accademia University Press.
- Papineni, Kishore & Roukos, Salim & Ward, Todd & Zhu Wei-Jing. 2002. BLEU: a method for automatic evaluation of machine translation. In Isabelle, Pierre & Charniak, Eugene & Dekang, Lin (eds), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (Philadelphia, PA, July 2002), 311-318. Stroudsburg: Association for Computational Linguistics.
- Piemontese, Emanuela. 1996. *Capire e farsi capire. Teorie e tecniche della scrittura controllata*. Napoli: Tecnodid.
- Rivera-Trigueros, Irene. 2022. Machine translation systems and quality assessment: a systematic review. *Language Resources and Evaluation* 56 (2). 593-619. <https://doi.org/10.1007/s10579-021-09537-5> (ultimo accesso 31 gennaio 2024)
- Rosi, Fabiana. 2009. «Non ha saputo dove è stato»: come annotare le forme non-target in un corpus di italiano L2? In Andorno, Cecilia & Rastelli, Stefano (a c. di), *Corpora di italiano L2: tecnologie, metodi, spunti teorici*, 49-70. Perugia: Guerra Edizioni.
- Schaeffer, Rylan & Brando, Miranda & Sanmi, Koyejo. 2023. Are emergent abilities of large language models a mirage? *Arxiv*. <https://doi.org/10.48550/ARXIV.2304.15004> (ultimo accesso 31 gennaio 2024)
- Serianni, Luca & Benedetti, Giuseppe. 2015. *Scritti sui banchi*. Roma: Carocci.
- Shterionov, Dimitar & Superbo, Riccardo & Nagle, Pat & Casanellas, Laura & O'Dowd, Tony & Way, Andy. 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation* 32 (3). 217-235.
- Tavosanis, Mirko. 2018. *Lingue e intelligenza artificiale*. Roma: Carocci.
- Turovsky, Barak. 2016. Found in translation: more accurate, fluent sentences in Google Translate. *Google: The Keyword*. <https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/> (ultimo accesso 7 febbraio 2024)
- Wei, Jason & Tay, Yi & Bommasani, Rishi & Raffel, Colin & Zoph, Barret & Borgeaud, Sebastian & Yogamata, Dani & Bosma, Maarten & Zhou, Denny & Chi, Ed H. et al. 2022. Emergent abilities of large language models. *Arxiv*. <https://doi.org/10.48550/ARXIV.2206.07682> (ultimo accesso 31 gennaio 2024)
- White, James. 1995. Approaches to Black Box MT Evaluation. In *Proceedings of Machine Translation Summit V* (Luxembourg, July 10-13, 1995). Luxembourg.

Wu, Tianyu & He, Shizhu & Liu, Jingping & Sun, Siqi & Lui, Kang & Han, Qing-Long & Tang, Yang. 2023. A brief overview of ChatGPT: the history, status quo and potential future development. *Journal of Automatica Sinica* 10 (5). 1122-1136. <https://doi.org/10.1109/JAS.2023.123618> (ultimo accesso 31 gennaio 2024)