UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova

Dipartimento di Studi Linguistici e Letterari

Corso di Laurea Magistrale in Linguistica
Classe LM-39

Tesi di Laurea

Modelli linguistici generativi e prospettive didattiche: un'indagine sul processo di correzione dei testi

Relatrice
Prof. Dominique Pierina Brunato

Correlatrice
Prof. Alberta Novello

Laureanda
Adriana Mirabella
n° matr.2062926 / LMLIN

Anno Accademico 2023 / 2024

A Padova,
una parentesi piena di vita.

Sommario

Introduzione

Il panorama educativo odierno si trova ad affrontare molteplici sfide: classi eterogenee, carichi di lavoro crescenti per gli insegnanti e la necessità di garantire una personalizzazione dell'apprendimento che risponda alle esigenze di ogni studente. In questo contesto, l'intelligenza artificiale (IA) emerge come un potenziale alleato per migliorare l'efficacia e l'equità del sistema educativo.

Nonostante le tecnologie fondate su IA non siano state originariamente progettate per applicazioni nel mondo dell'istruzione, il progressivo miglioramento dell'intelligenza artificiale nello svolgimento di *task* legati al *Natural Language Processing* (NLP), combinato ad un accesso su larga scala alle piattaforme che le utilizzano, ha reso urgente la riflessione sugli scenari derivanti da un uso sistematico e mirato di questi strumenti in ambito educativo.

Con NLP si fa riferimento ad una branca dell'Intelligenza Artificiale indirizzata al miglioramento dell'interazione tra esseri umani e tecnologia. Combinando linguistica computazionale, informatica e machine learning, il NLP permette lo sviluppo di algoritmi e modelli in grado di comprendere e generare linguaggio umano, riuscendo conseguentemente a svolgere compiti che richiedono tali competenze.

Tra le possibili applicazioni dell'IA in ambito didattico, la valutazione automatica di testi scritti riveste un ruolo particolarmente interessante. Questa tecnologia, se ben sfruttata, può infatti offrire numerosi vantaggi, ad esempio in termini di efficienza e scalabilità, perché capace di agevolare e snellire il processo manuale di correzione dei compiti. Inoltre, essa introduce la possibilità di offrire un feedback personalizzato relativo ai singoli elaborati, aiutando gli studenti a identificare i propri punti di forza e di debolezza e a migliorare le proprie capacità di scrittura. Ancora, analizzando un gran numero di testi, l'IA può aiutare gli insegnanti a identificare modelli e tendenze nelle prestazioni degli studenti, permettendo loro di intervenire in modo più efficace.

Questa tesi si propone di contribuire all'inerente ed attuale filone di studi attraverso un'indagine che ha coinvolto studenti e insegnanti di una scuola secondaria di primo grado del territorio di Padova. L'obiettivo perseguito è stato esplorare le potenzialità degli strumenti basati sull'Intelligenza Artificiale in ambito didattico, focalizzandosi in particolare sul processo di valutazione di un corpus di testi

argomentativi redatti dagli studenti e sulla qualità dei feedback correttivi prodotti da ChatGPT in relazione ai suddetti testi, confrontando i risultati ottenuti con le valutazioni elaborate da una docente.

ChatGPT è un Large Language Model (LLM) sviluppato dall'azienda OpenAI nel 2022, in grado di generare e comprendere testi in modo automatico. La sua capacità di analizzare la coesione, la coerenza e la struttura di un testo lo rende un candidato promettente per la valutazione di elaborati argomentativi. La scelta di utilizzare ChatGPT è stata motivata dalla sua notorietà anche tra i non esperti (quali i docenti) e dall'indisponibilità di modelli altrettanto validi al momento dell'indagine.

Attribuendo parzialmente a ChatGPT il ruolo di un docente, l'obiettivo finale della ricerca consiste nel definire il contributo che l'Intelligenza Artificiale può apportare alla didattica, offrendo un quadro realistico delle capacità e dei limiti di questa risorsa.

La tesi è strutturata come segue: il primo capitolo presenta un'introduzione alle nozioni fondamentali di linguistica testuale, con particolare attenzione ai concetti di coesione e coerenza. Questi elementi sono ritenuti imprescindibili affinché un testo possa dirsi tale e pertanto rientrano in quelle competenze linguistiche valutate nell'analisi dei testi appartenenti al corpus. Verranno inoltre esaminati elementi lessicali e grammaticali come connettivi, anafore e deittici, che contribuiscono a rafforzare l'unità comunicativa del testo e guidano il lettore nella costruzione di una continuità di senso.

Nel secondo capitolo sarà offerta una breve panoramica sui Large Language Models, illustrandone le caratteristiche e l'evoluzione attraverso un excursus storico. Partendo dalle intuizioni di Turing in merito alle capacità cognitive delle macchine, passando per lo sviluppo di modelli statistici e a rete neurale, si arriverà alla moderna architettura *Transformer*, su cui si basa ChatGPT. Verrà quindi descritto il funzionamento di ChatGPT, con un focus sulla tecnica del *prompting* come modalità principale di interazione con lo strumento e sull'importanza di una corretta impostazione del *prompt* per ottenere risultati significativi.

Il terzo capitolo allarga la prospettiva d'indagine al dibattito attuale, relativo agli strumenti basati sull'Intelligenza Artificiale in ambito didattico. Verranno presentati alcuni sondaggi che riportano opinioni e perplessità dei docenti sul tema, accompagnate da esempi pratici di possibili utilizzi di queste tecnologie in aula.

Si discuteranno anche studi sull'uso di ChatGPT per la personalizzazione dell'apprendimento e le modalità attraverso cui è possibile conseguirla servendosi dello strumento. Tra le diverse applicazioni, verranno evidenziate quelle relative alla generazione e correzione di testi, introducendo così il tema principale della tesi, discusso nel capitolo successivo.

Il capitolo quarto è infatti dedicato alla presentazione del caso di studio. Come anticipato, la ricerca si inserisce in un progetto di collaborazione con una scuola locale, e ha coinvolto una classe di terza media. Verrà descritta la raccolta di un corpus di 34 testi argomentativi prodotti dagli studenti, che saranno analizzati e valutati dal chatbot in diverse condizioni sperimentali, ciascuna esemplificata da un particolare *prompt*. La valutazione dello strumento sarà confrontata con quella data dalla docente coinvolta nella ricerca. Questo capitolo approfondirà la metodologia dell'indagine, definendone tempi, strumenti e modalità.

Il quinto capitolo presenterà i risultati emersi, comparando i feedback ottenuti da ChatGPT con quelli formulati dall'insegnante. Tale approccio permette di focalizzare l'attenzione sul dialogo tra operatore umano e macchina, evidenziando le potenzialità e i limiti del modello. Le implicazioni di questi risultati verranno poi discusse in un contesto più ampio, riconoscendo l'efficacia della tecnologia ma ribadendo la centralità della figura del docente nei processi di apprendimento. Verranno infine esposte nuove regolamentazioni per l'uso etico e inclusivo di queste risorse.

1. Fondamenti di linguistica testuale

1.1 Cosa intendiamo per testo?

Prima di approfondire i parametri di valutazione di un testo, è necessario concordare una definizione di quest'ultimo.

Un primo tentativo di riflessione sistematica in questa direzione è rappresentato da Halliday (1976), dove già dalle prime pagine si sottolinea la capacità di ciascun essere umano di distinguere in maniera spontanea ed istintiva un testo da un non-testo.

Un testo si affermerebbe in quanto tale non sulla base della propria lunghezza, al pari di una sorta di 'super-sentence', bensì attraverso l'espressione di una proprietà definita *texture* e così identificabile: «The concept of texture is entirely appropriate to express the property of 'being a text'. A text has texture, and this is what distinguishes it from something that is not a text. It derives this texture from the fact that it functions as a unity with respect to its environment» (Halliday 1976: 2).

Ancora, in un volume di poco successivo, il testo viene a definirsi in contrapposizione al sistema teorico lingua, in quanto espressione di un concreto contesto comunicativo, coinvolgente tanto le volontà dell'emittente, quanto la capacità inferenziale del ricevente.

Si tratta di fare un passo ulteriore rispetto alla considerazione della «parole» saussuriana intesa come semplice esecuzione, scegliendo di approfondirla focalizzandosi sulle opzioni linguistiche che di volta in volta vengono selezionate dai parlanti per trasmettere o decodificare un determinato contenuto comunicativo:

> Mentre la lingua è un sistema VIRTUALE di selezioni possibili ma non ancora realizzate, il testo rappresenta un sistema ATTUALIZZATO in cui sono state eseguite e realizzate certe selezioni possibili per dar forma a una determinata STRUTTURA (una relazione fra elementi). Questa strutturazione viene ottenuta tramite procedure di ATTUALIZZAZIONE (Beaugrande, Dressler 1981: 50).

Un testo si caratterizza quindi in quanto tale soprattutto grazie alla sua natura **funzionale** e **semantica**; persegue uno scopo comunicativo e lo fa costruendo un'unità di significato che prescinde dalla lunghezza minima 'frase' e può arrivare a coincidere con qualunque messaggio dotato di senso compiuto e autosufficiente.

Diremo quindi che l'unità fondamentale dell'articolazione semantica del testo è l'*unità comunicativa*; essa può, facoltativamente, articolarsi in una serie di unità informative che si dispongono nel testo in modo gerarchico, permettendoci di individuare

un nucleo che ne definisce la funzione illocutiva primaria ed una serie di proposizioni poste in secondo piano volte ad arricchire o spiegare quest'ultimo.

Negli anni si sono succedute svariate proposte relative all'individuazione di tipologie testuali ben definite, prediligendo ora la prospettiva del locutore e il fine comunicativo del testo, come in Werlich (1975)[1], ora il rapporto che viene instaurandosi tra l'autore e il destinatario attraverso l'accettazione di ciò che prende il nome di *vincolo interpretativo*.

Quest'ultimo è stato definito in Sabatini (1999) in quanto principio guida per la formulazione e l'interpretazione di un testo, fondato sulla volontà dell'emittente "di regolare in maniera più o meno rigida l'attività interpretativa del destinatario" (ivi, pp. 147-8). A seconda dei casi si avranno allora testi poco, mediamente o fortemente vincolanti[2], con una ripercussione sulle loro proprietà semantico-comunicative e sull'aderenza dell'opinione del lettore a quella dello scrivente.

Per quanto riguarda il testo argomentativo, per noi maggiormente significativo in quanto oggetto del caso di studio, esso rientra nella categoria dei testi mediamente vincolanti; persegue la dimostrazione della validità di una tesi attraverso il ricorso ad opportune argomentazioni, al contempo garantendo la crescita progressiva delle conoscenze del lettore in merito a una specifica tematica e, laddove necessario, servendosi della persuasione al fine di far convergere le opinioni dei due interlocutori.

In Lo Cascio (1991) vengono individuati tre elementi costitutivi del testo argomentativo: l'opinione (anche denominata *tesi*), l'argomento e la regola generale.

Vediamone un esempio:

a) Poiché nessun candidato è onesto, non è giusto votare. (Cignetti, 2011a, p.1468)

[1] La tassonomia proposta da Werlich nel 1975 si fonda su una suddivisione dei testi in 5 tipologie, sulla base della funzione comunicativa dominante: testi narrativi, testi descrittivi, testi argomentativi, testi espositivi ed infine testi regolativi.

[2] Appartengono alla classe dei testi "poco vincolanti" i testi letterari, a quella dei "mediamente vincolanti" i testi espositivi ed argomentativi e a quella dei "fortemente vincolanti" i testi scientifici, normativi e tecnico-operativi.

In questo caso possiamo far coincidere la tesi con “non è giusto votare”, l’argomento con “nessun candidato è onesto” e la regola generale con la premessa implicita “si devono votare solo candidati onesti”.

I tre elementi possono infatti essere espressi in modo diretto o implicito e ricorrendo o meno a componenti aggiuntive al fine di dare maggior credibilità alla propria posizione; ecco allora che si citerà una fonte autorevole o si lascerà spazio a contro argomentazioni (puntualmente smentite in una fase successiva) per prevenire possibili critiche.

Quando rese esplicitamente, le stesse argomentazioni e le relazioni che le congiungono fanno affidamento sugli schemi linguistico-testuali tipici della formalizzazione logica, sfruttando quelle forme che più avanti individueremo con l’etichetta di *connettivi*, ottenendo così un’apparenza di rigore anche nei casi più opinabili e contingenti.

1.2 Coesione

Il principio di unità a cui finora abbiamo fatto riferimento in quanto elemento costitutivo di un testo, si configura specificamente come unità semantica, intesa come catena di significati in grado di mettere in comunicazione le diverse parti del testo.

Il raggiungimento di questa continuità di senso è coadiuvato dal ricorso ad una serie di espedienti strutturali e grammaticali, che convergono in un’ulteriore proprietà del testo denominata *coesione* e definita nel volume di Halliday come:

> The relations of meanings that exist within the text and that definite it as a text. Cohesion occurs where the INTERPRETATION of some element in the discourse is dependent on that of another. The one PRESUPPOSES the other in the sense that it cannot be effectively decoded except by recourse at it… the cohesion lies in the relation that is set up between the two (Halliday 1976: 4-5).

Il concetto di coesione viene ripreso ed ulteriormente analizzato in De Beaugrande, Dressler (1981), dove assieme a quello di coerenza si afferma in quanto parte delle condizioni preliminari per dichiarare l’efficacia comunicativa di un testo e rappresenta: «il modo in cui le componenti del testo di superficie, ossia le parole che effettivamente udiamo o vediamo, sono collegate fra di loro» (De Beaugrande, Dressler 1981: 26).

La coesione, così come la coerenza, se rapportata agli altri principi individuati dagli studiosi in quanto fondativi della testualità (e cioè: intenzionalità, accettabilità, informatività, situazionalità, intertestualità) e relativi al rapporto tra testo e contesto o testo ed interpreti, appare invece unicamente vincolata alla dimensione testuale.

Nella costruzione di un testo coeso un ruolo d'eccezione spetta quindi alla sintassi, inquadrata nel volume precedentemente citato in quanto sovrastruttura, architettura superficiale in grado di guidare il lettore nel suo percorso di interpretazione del testo in maniera molto agile.

Se egli fosse infatti ogni volta chiamato a rievocare i significati profondi dei termini (immagazzinati nella memoria a lungo termine) per avanzare nella lettura e nella comprensione di un testo, quest'ultima si conseguirebbe in tempi lunghissimi.

Diversamente la sintassi, e nello specifico le costruzioni di cui essa si serve, è conservata nella memoria a breve termine, ed è dunque più facilmente richiamata alla mente dal lettore laddove necessario, permettendogli di procedere nell'interpretazione del testo facendo affidamento su una serie limitata di alternative:

> It follows that text processing could not afford to run through the participants' vast stores of world knowledge immediately, there must be some ancillary organizational system with far more limited options and patterns. In natural language texts, this system is that of syntax, whose classes and structures though often more diversified than what is found in English, are still quite limited in number in comparison to the classes and structures for concepts and relations…This account is borne out by observations that surface structures are more predominantly maintained in a 'short-term' storage, and conceptual content in a 'long-term' storage (Wright 1968).3

Per meglio rendere conto dei diversi piani su cui agisce il medesimo principio di coesione in un testo, Conte ne propone una scomposizione in due sottocategorie: coesione e connessitá.

Se entrambe contribuiscono alla costruzione di quell'unità di senso citata all'inizio del capitolo, la prima ha da intendersi principalmente come: «la presenza fra le parti del testo di relazioni semantiche e tematiche: ad esempio, in un testo narrativo si riscontra coesione nel costante riferimento agli stessi personaggi, a luoghi, a sequenze temporali concatenate» Conte (1989).

La connessitá diventa invece espressione superficiale di queste relazioni, garantendo continuità attraverso l'impiego di strutture grammaticali di rinvio[4] e di

[3] Beaugrande R.A., Dressler W., Introduction to Textlinguistics, Londra, Longmann, 1981, p. 52
[4] Per una trattazione più approfondita si rimanda al § 1.4.

connessione che segnalano al lettore i legami presenti tra le varie componenti del testo e lo guidano nella sua attività interpretativa.

1.2.1 I connettivi

Oltre alla punteggiatura, alla morfologia verbale[5] e all'intonazione con cui si sceglie di pronunciare un enunciato, ulteriori elementi che contribuiscono al raggiungimento della coesione di un testo sono raggruppabili nella categoria dei *connettivi*; si tratta prevalentemente di forme invariabili quali congiunzioni (perché, dato che…) e avverbi (di conseguenza, quindi, in effetti…) che collegano porzioni più o meno ampie di testo stabilendo tra esse rapporti di coordinazione o dipendenza gerarchica ed esplicitando le relazioni logiche che esse sottendono: «Conjunctive elements are cohesive not in themselves but indirectly, by virtue of their specific meanings; they are not primarily devices for reaching out into the preceding (or following) text, but they express certain meanings which presuppose the presence of ohter components in the discourse» (Halliday, Hasan 1976: 226).

Alcuni esempi delle funzioni a cui assolvono sono: introdurre un nuovo tema, riprendere il tema dopo una digressione, addurre un esempio…

Trattandosi di una classe di espressioni molto ampia e variegata, nel tempo essa è stata sottoposta a numerose classificazioni. Per questioni di praticità scelgo di rifarmi alla suddivisione presente in Ellero (1986), a sua volta in stretta correlazione con la proposta di Halliday, Hasan (1976)[6] e di Bazzanella (1995).

L'autrice propone infatti una distinzione tra connettivi *semantici* e connettivi *pragmatici*; i primi sono quelli che cementano la relazione tra due porzioni di testo modificandone il senso attraverso le rispettive informazioni semantiche, implicitamente veicolate ed intuite dal lettore, i secondi hanno invece a che fare con il contesto conversazionale e sono quelli che non agiscono direttamente sul contenuto degli enunciati, modificandone il significato, ma ne sanciscono la coesione da un punto di vista comunicativo, collegando atti linguistici diversi (ad esempio una proposta ed il successivo rifiuto della stessa), chiarendo le intenzioni del parlante nello specifico contesto

[5] In italiano, quest'ultima permette di inferire informazioni relative al tempo dell'azione e al soggetto che la esegue.

[6] All'interno del volume viene proposta una suddivisione della categoria Conjunction in: Additive, Adversative, Causal, Temporal.

enunciativo (aiutandolo ad organizzare la sua esposizione) e richiamando l'attenzione dell'ascoltatore.

Quando rispondenti a questa seconda definizione, i connettivi si affermerebbero specificamente in quanto *segnali discorsivi*, ulteriormente riconoscibili dalla mobilità che assumono all'interno dell'enunciato (sebbene tendano a ricorrere in posizione iniziale) e dall'indipendenza prosodica.

Vediamo ora alcuni esempi di entrambe le tipologie di connettivi[7]:

a) Una tempesta di neve seguì la battaglia.
b) Una tempesta di neve precedette la battaglia.
c) Dopo la battaglia ci fu una tempesta di neve.
d) Prima della battaglia c'era stata una tempesta di neve.
e) Dopo che ebbero combattuto, nevicò.
f) Prima che combattessero, aveva nevicato.
g) Combatterono una battaglia. Poi, nevicò.
h) Combatterono una battaglia. Prima, aveva nevicato

Nelle prime tre coppie di frasi (a-f), la coesione tra gli enunciati e la relazione temporale che si instaura tra il verificarsi della tempesta di neve e la battaglia si realizza attraverso una pluralità di elementi strutturali: il ricorso ai segni paragrafematici che fungono da congiunzione, la semantica del lessico scelto, il rapporto tra diatesi attiva e passiva…

Nell'ultima coppia invece, la relazione temporale veicolata dall'avverbio si afferma in quanto unico collante tra i due enunciati, chiarendone i termini del rapporto semantico e permettendoci di individuarlo in quanto necessario elemento coesivo, riconducendolo così alla categoria dei connettivi semantici.

Per chiarire la distinzione tra questi e i connettivi pragmatici, ci avvaliamo delle seguenti coppie di esempi:

i) Roberto è all'ospedale perché ha avuto un incidente
j) Roberto ha avuto un incidente perché è all'ospedale

[7] Gli esempi sono tratti dai volumi precedentemente citati.

In i) *perché* svolge una funzione semantica; veicola la connessione causale tra i due avvenimenti. In j) *perché* svolge una funzione pragmatica, in quanto è il fatto che Roberto si trovi all'ospedale a consentire la mia affermazione (atto illocutivo).

In italiano l'uso del connettivo causale con funzione pragmatica è presente anche nelle frasi in cui *perché* può o deve essere sostituito da *che*:

k) Vieni a giocare a tennis, *che ho prenotato il campo per oggi?
l) Vieni a giocare a tennis *perché ho prenotato il campo per oggi?

In k) il fatto che io abbia prenotato il campo determina la richiesta di giocare a tennis. In l), rimanendo all'interno di un unico atto illocutivo (la mia domanda), io voglio sapere se è il fatto di aver prenotato il campo che comporta il tuo prendere parte alla partita (connessione causa-effetto tra due contenuti proposizionali).

La sola presenza di elementi coesivi in un testo non basta però per riconoscere a quest'ultimo una certa coerenza; un esempio esplicito di ciò è il seguente:

a) *Mio fratello* non studia a *questa università*. Egli non sa che *la prima università* tedesca fu Praga. In *tutte le università* c'è il numero chiuso. *L'università* ha un laboratorio linguistico.[8]

Pur essendo presenti ad un livello superficiale espedienti coesivi quali la ripetizione e la ripresa pronominale, riesce difficile attribuire a questo estratto un senso globale.

Viceversa, un testo in cui siano completamente assenti questi elementi, non vedrà in alcun modo condizionata la sua coerenza, confermandoli in quanto utili per il processo di comprensione ma non strettamente necessari.

1.3 Coerenza

Diversamente dalla coesione, equiparata ad altre proprietà che contribuiscono alla corretta formulazione di un testo ma che, se assenti, non ne minacciano tale natura,

[8] Ferrari (2014).

nell'opera di Conte la coerenza si impone invece in quanto quintessenza di un testo, elemento fondamentale per permettere di poter considerare tale un insieme di enunciati.

Tuttavia, «essa non è però una proprietà intrinseca del testo, ovvero non è intrinsecamente contenuta nelle espressioni che compongono il testo, ma proviene ad esso dall'attività interpretativa del ricevente» (Andorno 2003: 18-19).

Difatti, se le parole su un piano teorico contengono in sé numerose accezioni, acquisiscono invece un senso univoco quando calate in un determinato contesto; spetterà dunque al lettore scegliere il significato da attribuirgli sulla base del proprio orizzonte delle aspettative, determinando la probabilità che una specifica parola assuma un valore piuttosto che un altro in quel dato enunciato:

> In that perspective, the sense of an expression or the content of a concept are definable as an ordered set of hypotheses about accessing and activating cognitive elements within a current pattern. To describe such a sense or content, one would have to stand at that point in the configuration of concepts and relations and look out along all pathways (Quilban, 1966).

L'attivazione dei concetti relativi alle diverse voci lessicali è conseguente alla menzione di queste ultime in un testo; definiamo infatti *referente testuale:* «Un oggetto concettuale specifico, attuale, che viene evocato nel discorso da uno dei parlanti e a cui, una volta evocato, si possono attribuire proprietà, azioni, eventi» (Andorno 2003:28).

La capacità di un parlante di individuare il referente testuale a cui si riferisce una specifica espressione e di attribuirgli determinate qualità è influenzata dal grado di identificabilità e di attivazione di questo; tralasciando i casi in cui un referente risulta intrinsecamente identificabile poiché rimanda ad un'unica entità nel reale, in tutti gli altri spetterà agli interlocutori marcarne e coglierne la definitezza a seconda dei ruoli, contando su un bagaglio di conoscenze condivise e su specifici espedienti grammaticali.

Un referente testuale può infatti essere introdotto attraverso il ricorso ad un sintagma nominale, a un nome proprio o a un pronome.

Quanto al processo di attivazione, esso coinvolge quello che è stato definito "active storage", il quale rappresenta uno spazio mentale all'interno della nostra memoria in cui i concetti assumono la configurazione di nodi in una rete; ogni nodo corrisponde a un concetto e i legami tra i diversi nodi rendono conto delle relazioni di tipo semantico che intercorrono tra più concetti. Essi risultano al massimo grado di attivazione nel momento in cui vengono menzionati per la prima volta in un testo; a partire da quel

momento, e seguendo l'impostazione tradizionale della linguistica testuale, il *nuovo* si trasforma in *dato* focalizzando altrove l'attenzione dell'utente:

> In quest'ottica diventa estremamente utile la nozione di "universo di discorso" (cfr. Levelt, 1989): con questa nozione ci si riferisce all'insieme organizzato di informazioni, conoscenze e credenze che i partecipanti a una conversazione o gli interpreti di un testo possiedono, condividono, credono di condividere [...]. L'universo di discorso viene continuamente modificato a mano a mano che il testo si sviluppa [..] e, reciprocamente, i cambiamenti nell'universo di discorso influenzano il modo in cui l'informazione è codificata nel testo (perchè, ad esempio, un'informazione precedentemente segnalata come "nuova" viene in seguito segnalata come condivisa) (Andorno 2003: 24).

In italiano, un referente nuovo viene in genere introdotto dall'articolo indeterminativo, viceversa un referente già menzionato nel testo sarà ripreso da un articolo determinativo o da altri specificatori nelle sue successive occorrenze.

Secondo lo psicologo statunitense George Miller, l'active storage è in grado di processare solo sette elementi alla volta, spiegando come mai certe correlazioni richiedono un tempo di elaborazione più lungo di altre quando coinvolgono concetti semanticamente distanti tra loro.

È infatti in Collins, Loftus (1975) che si afferma definitivamente l'idea di un'attivazione simultanea di concetti semanticamente correlati (*spreading activation*), che coadiuverebbe i tentativi del lettore di identificare un senso globale nel testo, facilitandone l'interpretazione:

> When some item of knowledge is activated, it appears that other items closely associated with it in mental storage also become active [...]. This principle is often called spreading activation and mediates between the explicitly activated concepts or relations and the detailed richness which a textual world can assume. [...] In reception, spreading activation makes it possible to form elaborate associations, to create predictions and hypotheses [...] far beyond what is actually made explicit in the surface text. (De Beaugrande, Dressler 1981: 87-88).

La ricerca di un senso globale è altresì guidata da ciò che, sempre Miller, ha definito "global pattern", e che coincide con una serie di aspettative universali relativamente all'associazione di eventi, comportamenti, idee…

Un esempio di global pattern è costituito dai concetti di *frame* e *script*, rispettivamente relativi a rappresentazioni mentali di situazioni tipiche e di sequenze di azioni legate a specifici contesti, modellate sulla base dell'esperienza dei singoli individui.

Difatti, nell'attribuzione di un significato specifico alle diverse forme lessicali, gioca un ruolo importante anche quella che è stata definita in Sperber, Wilson (1986)

relevance (in italiano *pertinenza*), e che renderebbe conto del procedere per *inferenze*[9] del lettore sulla base dei dati contestuali a lui offerti e dell'importanza (variabile) da egli attribuita a questi ultimi relativamente alla situazione comunicativa in analisi.

Riassumendo, nel momento in cui all'interno di un testo il lettore, ricorrendo all'utilizzo dei processi mentali sinora citati, riesce a rinvenire una *continuità di senso*, allora tale testo potrà dirsi coerente:

> A text "makes sense" because there is a continuity of senses among the knowledge activated by the expressions of the text (cf. Hörmann 1976). A "senseless" or "nonsensical" text is one in which text receivers can discover no such continuity, usually because there is a serious mismatch between the configuration of concepts and relations expressed and the receivers prior knowledge of the world. We would define this continuity of senses as the foundation of coherence, being the mutual access and relevance within a configuration of concepts and relations (De Beaugrande, Dressler 1981: 84).

Con *continuità* intendiamo la presenza di collegamenti impliciti o espliciti tra le diverse unità di significato che compongono il testo, attraverso il costituirsi di un fil rouge logico-semantico che lo attraversa e congiunge. Tuttavia, un testo per dirsi coerente dovrà accompagnare alla continuità altre due proprietà: unitarietà e progressione.

L'unitarietà garantisce il dialogo tra il nucleo semantico centrale e il resto del contenuto del testo, considerato un'espansione di questo; la progressione è invece riconducibile ad una crescita dell'informazione trasmessa dal testo mano a mano che esso si sviluppa.

1.4 Dispositivi di riferimento e di rinvio

I diversi espedienti ai quali si può ricorrere per segnalare al lettore di un testo la ri-occorrenza di un referente precedentemente introdotto in esso, stabilendo così una relazione tra i due termini (prevalentemente di tipo co-referenziale) [10], possono essere inseriti nella macrocategoria di "dispositivi di rinvio e riferimento".

La tipologia di relazione alla quale più di frequente si fa ricorso è quella dell'anafora, con cui intendiamo: «la relazione fra due elementi linguistici in cui l'interpretazione di uno, detto anaforico, richiede in qualche modo l'interpretazione

[9] Definiamo inferenza il processo attraverso il quale i lettori riempiono le lacune informative di un testo facendo affidamento sulle premesse contenute nel contesto e nel co-testo e sulla propria background knowledge, cogliendone così i significati impliciti.

[10] Parliamo di co-referenza quando un'anafora rimanda allo stesso referente dell'elemento a cui è riferita.

dell'altro, detto antecedente» (Huang, 2000).[11] Non sono tuttavia rari i casi in cui la relazione tra i termini si stabilisce all'inverso, facendo cioè precedere l'elemento di rinvio all'effettiva menzione del concetto a cui esso si riferisce e che in qualche modo anticipa; in questo caso si parla quindi di *catafora*.

I mezzi linguistici di ripresa anaforica possono essere molteplici e variano a seconda della lingua e del contesto specifico d'utilizzo. Tra i principali riscontriamo i sintagmi nominali e le pro-forme; queste ultime sono parole funzionali (es: pronomi personali, relativi, dimostrativi ma anche aggettivi, avverbi e verbi…) che non rimandano **direttamente** ad un referente specifico nella realtà, ma ne richiamano la precedente occorrenza nel co-testo, "sostituendosi" ad esso:

a) Il *cavallo bianco* era in testa. *Lo* vedevamo sopravanzare gli avversari di almeno una lunghezza. [12]
b) Ho fatto piantare *due ippocastani* nel mio giardino. *Questi alberi* fanno molta ombra. [13]

La sostituzione può però essere effettuata anche mediante il ricorso a sinonimi, a sostantivi dal significato sovraordinato (es: *iperonimi*) o generico e a sintagmi nominali e pronomi che riprendono l'intero contenuto di uno o più enunciati (*incapsulatori anaforici*).

Con "iperonimo" intendiamo un sostantivo dal significato più ampio rispetto al termine di riferimento e dunque riconducibile ad una classe che ingloba quest'ultimo.

Il ricorso all'iperonimo garantisce la variatio del testo, mantenendo inalterate le informazioni semantiche del concetto espresso dai termini impiegati e assicurandone l'attivazione nella memoria testuale.

Un'esemplificazione di quanto detto è rappresentata da b).

Per quanto riguarda gli incapsulatori anaforici, se in parte rispondono alle stesse esigenze degli iperonimi, sono altresì utilizzati per sintetizzare il contenuto precedente, aggiungendovi ulteriori sfumature di significato e lasciando talvolta trapelare le opinioni dell'autore:

[11] Da Andorno 2003: 45.
[12] Da Andorno 2003: 49.
[13] Da Ferrari 2014: 188.

c) Alla Sbav quell'anno l'Ufficio Relazioni Pubbliche propose alle persone di maggior riguardo che le strenne fossero recapitate a domicilio da un uomo vestito da Babbo Natale. *L'idea* suscitò l'approvazione unanime dei dirigenti. (I. Calvino, in Lala, 2010 p. 641)

d) Il battello affondò nella notte, portando con sé in fondo all'oceano trecento persone. *Quel disastro* fu uno dei più gravi della storia della navigazione. (R. Simone)

In d) è infatti la stessa semantica della parola disastro a determinare la connotazione negativa dell'evento, permettendoci di inferire la valutazione dell'autore in merito a questo proprio attraverso la scelta lessicale operata.

Per riferirsi a casi del genere, Conte conia l'espressione "anafora empatica", in quanto il referente richiamato assume un giudizio di valore (emesso dallo scrivente) che viene immediatamente percepito dal lettore.

L'anafora empatica può realizzarsi non solo attraverso l'utilizzo di termini che rimandano intrinsecamente ad una componente valutativa (come nel caso di "disastro") ma anche attraverso la deformazione morfologica risultante da processi di alterazione (es: creazione di diminutivi, vezzeggiativi…) e il ricorso ad accezioni specifiche e impieghi particolari di termini generici e nomi comuni:

e) A: Guarda! Sta arrivando Giada
   B: Chi l'ha invitata *quella*?

L'utilizzo del pronome dimostrativo, genericamente adoperato per segnalare la distanza tra un referente e chi parla, in un contesto di comprovata vicinanza fisica permette di interpretare la scelta linguistica come un segno di lontananza affettiva.

Un esempio ancora più esplicito è rappresentato dal seguente passo tratto da *Madame Bovary*:

f) Tandis qu'il [il marito Charles] trottine à ses malades, *elle* reste à ravauder des chaussettes. Et *on* s'ennuie! *On* voudrait habiter la ville, danser la polka tuos les soirs! Pauvre petite femme. *Ça* bâille après l'amour […]. Avec trois mots de galanterie, *cela* vous adorerait; j'en suis sûr! […] mais comment s'en débarasser ensuite? (G. Flaubert in Conte 1999, p.79)

Il progressivo passaggio dal pronome personale *elle* all'impersonale *on* fino ad arrivare ai dimostrativi *ça* e *cela*, genericamente riferiti ad oggetti inanimati, tradisce il cambiare dei sentimenti e degli atteggiamenti del marito nei confronti della moglie.

Questo estratto può considerarsi anche un esempio di *catena anaforica*, con cui indichiamo l'insieme di rimandi anaforici (diretti o indiretti) presenti in un testo. L'elemento che per primo introduce un referente nel testo viene denominato "capo-catena" e forma, assieme alle successive occorrenze anaforiche, gli "anelli" di quest'ultima.

Sottolineando la continuità referenziale del testo, la catena anaforica diventa ulteriore espressione dell'unitarietà semantica di questo, affermandosi in quanto dispositivo di coesione.

Le frasi analizzate finora costituiscono un esempio di anafora diretta eseguita per sostituzione; è però importante sottolineare che un riferimento anaforico può istituirsi anche attraverso altri mezzi, quali ripetizione ed ellissi per l'anafora diretta e associazione per quella indiretta. [14]

L'anafora per ripetizione è da ricondurre ai casi in cui si verifica una ripresa parziale o totale dell'antecedente a cui l'anafora fa riferimento, spesso al fine di mantenere l'oggetto dell'enunciato in una posizione di focus per l'interprete del testo, facilitando la comprensione dello stesso.

Viceversa, l'ellissi rappresenta l'omissione del costituente a cui l'anafora si riferisce in quella che è considerabile la sua seconda menzione; pertanto, si afferma in quanto tale solo se confrontata con la forma completa dell'enunciato, che ne fa emergere l'assenza di specifici elementi.

[14] Per una descrizione dettagliata di tutte le tipologie anaforiche si rimanda a Conte, Maria-Elisabeth (19992), Deissi testuale e anafora, Determinazione del tema, Anafora empatica, in Ead., Condizioni di coerenza. Ricerche di linguistica testuale, Alessandria, Edizioni dell'Orso, pp. 11-28, 51-58, 75-82 (1a ed. La Nuova Italia, 1988).

Un esempio[15] utile a comprendere meglio quanto detto è il seguente:

a) Carla prende$_i$ la margherita, Sandra Ø[16] la napoletana.

Qui l'ellissi è riferita al predicato verbale, la cui informazione semantica viene facilmente recuperata dal co-testo precedente.

È da considerarsi ellissi anche l'omissione del soggetto che riprende anaforicamente il costituente esplicitato in prima battuta, come nel seguente esempio:

b) Carla ordinerà le pizze e Ø le porterà subito a casa

Tuttavia, nel caso specifico dell'italiano, il soggetto zero rappresenta un caso limite di ellissi, in quanto (trattandosi di una lingua pro-drop), le informazioni relative a persona e numero sono ricavabili a partire dalla morfologia del verbo.

Le connessioni referenziali che avvengono in modo indiretto sfruttano invece il meccanismo dell'anafora associativa. Diversamente dai casi analizzati finora, non vi è una ripresa esplicita del referente che le precede nel co-testo, bensì vi è l'introduzione di un elemento ad esso legato implicitamente, per via concettuale.

Si tratta dunque di un legame che il parlante è in grado di stabilire sfruttando le proprie conoscenze lessicali, enciclopediche e/o extralinguistiche, come nel caso seguente, in cui si riesce senza fatica ad individuare la tastiera come sottocomponente dell'elemento computer:

c) Mi ha riparato *il computer*. Non era nulla di grave: è bastato sbloccare *la tastiera*.

Spesse volte, il meccanismo dell'anafora associativa viene sfruttato per accrescere le conoscenze del lettore in un determinato contesto: "guidato dal principio di coerenza, egli scopre connessioni referenziali fino a quel momento a lui sconosciute" (Ferrari 2014: 183).

[15] Gli esempi che seguono sono tratti nuovamente da Andorno (2003).
[16] Il segno 0 indica il luogo da cui è stato cancellato il costituente ellittico.

1.5 Deissi e anafora

Potremmo considerare anafora e deissi due facce della stessa medaglia; se infatti l'anafora è comparabile a quello che viene definito in Halliday, Hasan (1976) "rimando all'interno", da intendersi come circoscritto allo spazio testo, la deissi rinvia invece al contesto extralinguistico, affermandosi in quanto elemento imprescindibile per l'interpretazione di determinati enunciati: «Per «deissi» si intende infatti quel fenomeno linguistico per cui determinate espressioni richiedono, per essere interpretate, la conoscenza di particolari condizioni contestuali che sono l'identità dei partecipanti all'atto comunicativo e la loro collocazione spazio-temporale» (Vanelli, Renzi 1995: 262).

L'insieme delle suddette condizioni prende il nome di "campo indicale" e risulta orientato rispetto ad un'origine (*origo*) che coincide col parlante e col suo punto di vista.

Pertanto, nell'esempio seguente:

d) *Io* resterò *qui* fino a *domani*

l' *io* è da riferirsi al parlante, il *qui* al luogo da lui occupato e il *domani* ad un tempo successivo al momento dell'enunciazione.

All'interno di uno stesso testo, il campo indicale preso in considerazione può però variare assumendo di volta in volta la prospettiva dei parlanti coinvolti; un caso tipico è rappresentato dal discorso riportato, in cui si sovrappongono i campi indicali di chi ha prodotto in un primo momento il discorso che viene riportato e di chi vi sta attualmente facendo riferimento:

e) «Spero che non piova» disse Luigi
f) Luigi disse che sperava che non piovesse

Se nel primo caso i due campi indicali restano separati e vengono opportunamente distinti anche attraverso il ricorso a determinati espedienti grafici, nel secondo, il campo indicale di Luigi viene traslato in quello del narratore, collocandosi in un tempo anteriore (momento del riferimento) rispetto al momento dell'enunciazione ed acquisendo così una

sfumatura anaforica, poiché riconducibile ad un momento di riferimento collocato nel passato rispetto all'origo primaria (quella del narratore).

In altri tipi di discorso riportato, quali l'indiretto libero, questa traslazione è meno netta e assume una configurazione parziale:

g) Improvvisamente s'interruppe per ordinare che, perdio, quel figliuolo se ne poteva andare a piangere di là (Pirandello, *Superior stabat lupus*, in *Novelle per un anno*).

In questo caso si riscontra una traslazione dal punto di vista temporale e personale, con l'utilizzo dei tempi verbali all'imperfetto e il coincidere del parlante con la terza persona, ma resta invece inalterata la componente spaziale (*di là*) che rinvia al campo indicale del personaggio.

Come emerso dagli esempi precedenti, gli elementi deittici (avverbi, pronomi, verbi) possono essere intrinsecamente tali oppure, come accennato nel §1.4, rimandare ad una funzione anaforica.

Partendo da questo presupposto, negli anni sono state avanzate numerose interpretazioni volte a rinvenire una maggiore somiglianza tra questi due elementi (anafore e deittici). Una delle più celebri risulta essere quella riportata in Andorno (2003), secondo la quale i deittici non farebbero direttamente riferimento al contesto (extralinguistico), bensì, al pari delle anafore, sarebbero riconducibili a sintagmi nominali impliciti almeno in una fase intermedia, risultando dunque anch'essi in qualche modo ancorati al co-testo.

Una riprova di ciò starebbe nell'accordo di numero e genere che si instaura tra il deittico e l'espressione lessicale indicante il referente reale. Ad esempio, in:

h) Togli*la* di mezzo!

supponendo che si stia parlando di una sedia, il pronome anaforico assume genere e numero del sintagma nominale (sedia), realizzando un riferimento, seppur implicito, a quello che viene considerato un referente già concettualmente attivo nell'universo del discorso.

Tuttavia, un'argomentazione del genere risulta troppo debole per permetterci di affermare che il referente possa dirsi *realmente* presente nel mondo testuale, al pari dei referenti richiamati attraverso le anafore.

Mantenendo dunque la distinzione tra questi due elementi, è comunque importante sottolineare che la deissi presenta una referenzialità meno diretta di quanto possa sembrare ad una prima analisi.

1.5.1 Deissi testuale

Un ulteriore dialogo tra deissi e anafora si instaura nell'ambito della deissi testuale, anche denominata "logodeissi" in Conte (1978). Essa si realizza quando espressioni intrinsecamente deittiche (dimostrativi, tempi verbali quali il passato prossimo o il futuro…) non fanno riferimento ad elementi del contesto situazionale, bensì ad elementi del testo, «ad una parte del discorso in corso («ongoing discourse»), ad un segmento o momento del discorso in atto» (Conte 1999: 13).

Quest'ultimo, diversamente da quanto accade con l'anafora, non viene però considerato come un semplice strumento attraverso il quale evocare, passando per una "mediazione" nominale (antecedente), un referente del mondo, bensì si afferma esso stesso in quanto referente designato.

Citando Andorno (2003: 67): «ci troviamo insomma di fronte a un campo indicale particolare, che è costituito dal testo stesso e ha come origo il punto del testo in cui il lettore si trova».

Le espressioni principalmente impiegate per realizzare la deissi testuale sono pertanto termini cronodeittici e topodeittici come: "qui, nel capitolo precedente, nel paragrafo successivo, sopra, prima, in precedenza, più avanti…".

I seguenti esempi, tratti da De Cesare (2010) chiariscono definitivamente l'ambiguità tra questo specifico utilizzo della deissi e l'anafora:

i) Nel capitolo 2 abbiamo presentato i mammiferi. *In questo capitolo* parleremo dei rettili [deittico testuale: *questo* = il capitolo che stiamo leggendo, ovvero il capitolo 3].

j) Nel capitolo 2 abbiamo presentato i mammiferi. *In questo capitolo*, in particolare, abbiamo presentato le caratteristiche dell'uomo [*questo* = il capitolo 2].

In i) il dimostrativo questo compare nella sua veste deittica, riferendosi materialmente ad una parte del testo in analisi; in j) viene invece usato anaforicamente come ripresa della forma lessicale introdotta all'inizio dell'enunciato.

Quanto detto finora offre un'ampia panoramica sugli espedienti che concorrono alla stesura di un testo coerente e coeso. L'utilizzo efficace di questi elementi risulta fondamentale per la fluidità del testo, facilitando lettura ed interpretazione da parte del lettore.

2. Intelligenza artificiale e Large Language Models: origini e stato dell'arte

2.1 Quando nasce l'Intelligenza Artificiale?

Con "Intelligenza Artificiale" facciamo riferimento ad una disciplina che affonda le sue origini nella seconda metà del Novecento e che riconosce come obiettivo primario lo sviluppo di tecnologie (algoritmi e sistemi) in grado di svolgere task complessi emulando in parte l'attività cognitiva umana.

Il termine "intelligenza" compare per la prima volta in relazione alle potenzialità delle macchine (da intendersi qui in riferimento ai dispositivi meccanici ed elettronici disponibili all'epoca) in Turing (1950), in cui il celebre matematico cerca di rispondere al quesito posto in apertura del paper ("Can machines think?") ragionando sulla possibilità di arrivare ad avere, in un futuro più o meno lontano, "thinking machines" definibili tali sulla base della loro riuscita nel gioco che egli stesso propone all'interno del testo e che prende il nome di "Imitation Game".

Il gioco prevede il coinvolgimento di tre partecipanti: due umani (uno incaricato di porre delle domande e l'altro di rispondervi) e una macchina (anch'essa soggetta alle domande dell'inquirente). Per l'interrogatore l'obiettivo è quello di riuscire a risalire all'identità di chi genera le risposte; se la macchina si dimostra in grado di confonderlo facendogli credere che si tratti di risposte provenienti da un essere umano, allora essa potrà dirsi intelligente.

La denominazione di "Intelligenza Artificiale" in relazione a questo ambito di studi è invece da ricondurre ai tentativi sperimentali effettuati durante il "Dartmouth Summer Research Project", iniziativa promossa tra gli altri da John McCarthy e Marvin Minsky nel 1955, attraverso il celebre manifesto nel quale compare la seguente dichiarazione di intenti:

> The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.[17]

[17] McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E., *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence* – august 31, 1955, "AI Magazine", 2006, vol. XXVII, (4).

Il workshop non portò a risultati particolarmente soddisfacenti…tuttavia, solo pochi anni più tardi verrà fondato il primo laboratorio dedicato all'intelligenza artificiale al MIT, in cui vedrà la luce il progetto *ELIZA*.

2.1.1 *ELIZA*: Il primo chatbot

Progettata da Joseph Weizenbaum, *ELIZA* costituisce il primo chatbot della storia.

Si tratta di un sistema che utilizza un set predefinito di regole (dunque è *rule-based*) per processare i dati che gli vengono somministrati, facendo coincidere ogni input con una determinata azione sulla base della regola a cui essa risponde:

> Instead of representing knowledge in a declarative, static way as a set of things which are true, rule-based system represent knowledge in terms of a set of rules that tells what to do or what to conclude in different situations. A rule-based system is a way of encoding a human expert's knowledge in a fairly narrow area into an automated system. A rule-based system can be simply created by using a set of assertions and a set of rules that specify how to act on the assertion set. Rules are expressed as a set of if-then statements (called IF-THEN rules or production rules): IF P THEN Q which is also equivalent to: P⇒Q (Abraham, Grosan 2005: 49).

L'analisi degli input viene condotta attraverso la ricerca e l'individuazione di parole chiave, sfruttando un meccanismo di *pattern matching*.

Nella sua versione più famosa, *ELIZA* veste i panni di uno psicoterapeuta riconducibile alla scuola di pensiero Rogeriana[18]. Le affermazioni dell'utente costituiscono il centro della conversazione; rovesciandole, il sistema le riprende come materiale per le successive domande, trasformando di fatto l'output in nuovo input.

Le risposte del bot, apparentemente calzanti, appaiono ad un'analisi più profonda ripetitive e superficiali; il sistema su cui è basato gli impedisce infatti una reale comprensione del contesto e della semantica delle parole impiegate, lasciando intravedere parte dei limiti dell'approccio *rule-based*. Tra questi rientra la scarsa scalabilità, che interviene quando all'aumentare delle regole il sistema riscontra difficoltà nel disambiguarle e nell'applicarle in tempi relativamente brevi, andando inoltre a gravare sulle limitate risorse di memoria.

Ciononostante, possiamo considerare *ELIZA* espressione concreta dei primi passi mossi nel campo del *Natural Language Processing*.

[18] Facciamo riferimento al particolare tipo di psicoterapia person-centered, introdotto dallo psicologo Carl Rogers tra gli anni quaranta e cinquanta del '900.

2.2 Machine Learning e tipologie di apprendimento

È in quegli stessi anni che assistiamo ai primi sviluppi del Machine Learning, un ambito di studi che punta alla realizzazione di algoritmi in grado di apprendere determinati comportamenti a partire da un periodo di training effettuato servendosi di un corpus limitato (training corpus) e più o meno *rappresentativo* di un dato fenomeno. Quanto appreso viene poi generalizzato a nuovi dataset e contesti, portando il sistema a svolgere determinati task senza bisogno di essere riprogrammato o ulteriormente modificato dal programmatore, spingendosi dunque oltre i limiti dei sistemi *rule-based*: «Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort» (Samuel 1959: 211).

Un primo tentativo di applicazione di questo metodo viene effettuato nel 1956 proprio da Arthur Samuel, il quale sceglie di servirsi del gioco degli scacchi come strumento di apprendimento per la macchina, portandola ad impararne le regole e a predire gli esiti delle partite attraverso l'osservazione delle mosse e delle strategie messe in atto durante queste ultime. Si tratta quindi di apprendere attraverso l'esperienza: «Rather than program the computer to solve the task directly, in machine learning, we seek methods by which the computer will come up with its own program based on examples that we provide» (Abraham, Grosan 2005: 261).

Sulla base della tecnica utilizzata possiamo individuare quattro tipologie di apprendimento: supervised learning, unsupervised learning, semi-supervised learning e reinforcement learning.

Nel *supervised learning* l'algoritmo è allenato ricorrendo ad un training set che prevede il coinvolgimento di coppie di input e output opportunamente marcate mediante il ricorso ad etichette (*labels*) che segnalano al sistema la natura degli elementi e le relazioni che intercorrono tra essi, facilitandone il successivo riconoscimento. Attraverso l'analisi di questi esempi l'algoritmo riesce infatti a costruire un modello eventualmente applicabile anche a dati mai visti prima.

Con *unsupervised learning* facciamo invece riferimento al caso opposto, in cui i dati si presentano privi di annotazione e non c'è supporto all'algoritmo da parte del programmatore; il sistema provvederà autonomamente ad identificare un pattern per il riconoscimento e il raggruppamento delle informazioni.

Se il *semi-supervised* learning costituisce una soluzione intermedia fondata sulla combinazione delle due tecniche prima citate, il *reinforcement learning* prevede invece la presenza di feedback positivi o negativi associati alle diverse azioni svolte dal sistema, il quale tenterà di migliorare la propria performance al fine di garantire sempre il raggiungimento di un risultato positivo.

2.3 Modelli probabilistici del linguaggio

Il ricorso a tecniche di Machine Learning per la costruzione di algoritmi deputati allo svolgimento di task relativi al NLP passa inizialmente attraverso la creazione di modelli linguistici stocastici (*MLS*), utilizzati per stimare la probabilità d'occorrenza di una parola in un determinato contesto.

Diversamente da altri tipi di fenomeni, in cui il verificarsi dei singoli eventi che li compongono può essere visto in partenza come ugualmente possibile (si pensi al lancio di un dado a sei facce e al suo ipotetico risultato), le parole presenti in un testo non hanno tutte la stessa probabilità di essere prodotte, in quanto risultano soggette a specifici vincoli linguistici. Difatti: «La scelta di una parola *modifica* lo spazio di probabilità relativo alla scelta lessicale successiva» (Lenci et al., 2016: 166). Parliamo quindi di *probabilità condizionata*, in riferimento al verificarsi di una parola (W_2) data una parola (W_1) già presente all'interno di un enunciato.

Calcolare la probabilità di occorrenza di una frase equivale dunque a calcolare il prodotto delle singole probabilità condizionate delle parole che la compongono; conseguentemente la probabilità che si verifichi l'intera stringa è calcolata attraverso la seguente formula, che richiama la regola della catena del prodotto di probabilità (*Chain Rule*):

$$P(W_1W_2...W_n)=P(W_1)*P(W_2|W_1)*P(W_n|W_1,...,W_{n-1})$$

Un esempio pratico che può aiutarci nella comprensione del modello è il seguente:

> Considering S as the sentence "I am very happy" [...] To calculate this probability, the conditional probability can be employed:
>
> P(I,am,very,happy)=P(I)·P(am|I)·P(very|I,am)·P(happy|I,am,very)

> where P(I) represents the probability of the word "I" appearing and P(am|I) stands for the probability of "am" appearing given that "I" has appeared. When we multiply P(am|I) by P(I), it fulfills the condition of "I" appearing in P(am|I), resulting in the probability of "I am" appearing together as P (I, am) = P (I) · P (am|I).19

Uno dei principali limiti di questo approccio sta nel fatto che più si allarga il contesto da considerare per il calcolo della probabilità, più, a causa della scarsità di dati (*data sparsity*) e dell'andamento Zipfiano del vocabolario, si rischia di imbattersi in sequenze grammaticalmente possibili e tuttavia non presenti all'interno del corpus. Il sistema, faticando a riconoscerle come tali, le identifica quindi in quanto agrammaticali e gli attribuisce una probabilità pari a zero.

Una soluzione marginale al problema è rappresentata dal modello Markoviano (o *N-gram model*):

> I modelli markoviani stimano la probabilità di una parola a partire da un certo numero di parole che la precedono direttamente nel testo. Il cosiddetto *ordine* del modello markoviano specifica il numero esatto di parole che il modello prende in considerazione: nel caso del primo ordine solo la parola immediatamente precedente, nel modello del secondo ordine le prime due, nel modello di terzo ordine le prime tre e così via. L'ordine definisce anche il grado di complessità del modello markoviano (Lenci et al., 2016: 163).

Il presupposto sul quale si fonda il modello è dunque la possibilità di **stimare** la probabilità di una parola basandosi su un **contesto ristretto** di parole precedenti:

$$P(W_1 W_2 \ldots W_n) \approx \prod P(W_i \mid W_{1-k} \ldots W_{i-1})$$

Ciò che viene preso in considerazione per il calcolo è la frequenza relativa di una data sequenza (o n-gram), riprendendo quindi la definizione *frequentista* di probabilità per la quale è possibile stimare la probabilità di un determinato evento partendo dalla frequenza relativa dello stesso in un dato numero di osservazioni condotte nelle stesse condizioni (*Maximum Likelihood Estimation)*.

Diremo quindi che, dato un campione di osservazioni (corpus), la probabilità che occorra una determinata sequenza di parole corrisponde alla frequenza relativa di quest'ultima nel corpus, espressa come il rapporto tra il numero di occorrenze della specifica sequenza (dunque la sua frequenza assoluta) e il numero totale di sequenze che condividono n-1 parole.

[19] Chu, Z., Ni, S., Wang, Z., Feng, X., Li, C., Hu, X., ... & Zhang, W. (2024). History, Development, and Principles of Large Language Models-An Introductory Survey.

$$P(W_n|W_{n-1}) = \frac{C(Wn-1Wn)}{C(Wn-1)}$$

Un'ulteriore soluzione consiste nel ricorrere a metodi matematici di redistribuzione della probabilità (es: Laplace Smoothing). Tuttavia, neanche questa tecnica risulta particolarmente efficace, portando alla necessità di sviluppo di modelli più complessi, in grado di gestire contesti sintattici più ampi e di incorporare una conoscenza profonda della lingua.

2.4 Deep Learning e sviluppo delle reti neurali artificiali

Un passo in avanti in tal senso è garantito dall'avvento del Deep Learning.

Si tratta di una branca del Machine Learning che si sviluppa nei primi anni del nuovo secolo e che sfrutta le *reti neurali artificiali* (ANN) per costruire architetture complesse che rimandano al funzionamento della mente umana per quanto riguarda le connessioni stabilite tra i neuroni attraverso le sinapsi:

> Artificial Neural Network: is a network of artificial neurons which communicate with each other through edges (equivalent to the synapses of the biological neurons). The edges and the neurons are weighted and the weights can be adjusted through the learning process of the machine. As long as the output is not the desired and there is a difference (error) the machine will adjust the weight of the neuron in order to reduce that error. The neurons are arranged in layers. The first layer corresponds to the input layer and the last to the output (Koltsakis et al., 2023: 6).

Il primo livello (*input layer*) è costituito da neuroni deputati alla ricezione dell'informazione in ingresso che, una volta acquisita, viene trasmessa ai neuroni presenti negli strati successivi (*hidden layers*) e da essi elaborata prima di raggiungere l'ultimo livello denominato *output layer* e responsabile della risposta finale del sistema.

L'attivazione dei diversi nodi risente del peso (*weight*) attribuito ai neuroni che li precedono nel processo di elaborazione e trasmissione dei dati; tale peso viene all'occorrenza riconsiderato grazie al meccanismo di *backpropagation* che interviene quando il sistema produce un output diverso dalle aspettative (*error*). Attraverso la *backpropagation* viene dunque effettuata un'analisi di tutti i layer che compongono la struttura procedendo a ritroso e calcolando il contributo di ogni nodo all'errore, che verrà poi superato operando una redistribuzione del peso delle connessioni tra i diversi neuroni.

2.4.1 Word Embeddings

A partire dai primi anni 2000 il NLP ha conosciuto un grande sviluppo grazie alla combinazione dei modelli a rete neurale e di nuove proposte relative alla struttura e all'uso dei *Word Embeddings*. Definiamo *Word Embeddings* i vettori numerici utilizzati per rappresentare l'informazione semantica e sintattica delle parole al fine di facilitarne la comprensione da parte della macchina.

L'intuizione che guida il cambiamento nell'impostazione della sequenza vettoriale si fonda sull'ipotesi distribuzionale, per la quale parole semanticamente simili tendono ad occorrere in contesti simili e dunque condividono la medesima distribuzione: «The degree of semantic similarity between two linguistic expressions A and B is a function of the similarity of the linguistic contexts in which A and B can appear» (Lenci, 2008: 3).

Partendo da questo presupposto, in studi quali Bengio et al. (2003) si sottolinea la necessità di utilizzare vettori che tengano in maggior considerazione la somiglianza semantica e grammaticale delle parole per poter sviluppare algoritmi predittivi più efficienti, in grado di generalizzare quanto visto nel training corpus a strutture nuove ma dalle caratteristiche simili:

> There are at least two characteristics in this approach which beg to be improved upon, and that we will focus on in this paper. First, it is not taking into account contexts farther than 1 or 2 words,second it is not taking into account the "similarity" between words. For example, having seen the sentence "The cat is walking in the bedroom" in the training corpus should help us generalize to make the sentence "A dog was running in a room" almost as likely, simply because "dog" and "cat" (resp. "the" and "a", "room" and "bedroom", etc...) have similar semantic and grammatical roles (Bengio et al., 2003: 1139).

Un reale cambiamento si ha quindi con l'introduzione di word embeddings *densi*, in grado di concentrare in un'unica sequenza vettoriale una molteplicità di tratti che rendono conto delle caratteristiche formali delle parole e delle relazioni che si instaurano tra esse. La densità del vettore si traduce nell'assegnazione di un valore numerico significativo ad ogni sua componente, superando in questo modo le rappresentazioni proposte precedentemente quali i *one-hot vector*, in cui il vettore ha una lunghezza pari a quella del vocabolario del testo ma viene associato il valore 1 solo alla cella corrispondente alla parola in analisi, attribuendo alle altre il valore 0. Conseguentemente, per creare un unico vettore che rappresenti l'intero enunciato si rende necessaria una concatenazione dei singoli *one-hot vector* , ciascuno rappresentante la parola di volta in volta analizzata.

Uno dei principali modelli neurali fondato su sequenze vettoriali dense è il Word2Vec.

2.4.1.1 Word2Vec

Introdotto in Mikolov et al. (2013) il modello Word2Vec genera word embeddings partendo da un input testuale. Ad ogni parola viene associato un vettore collocato in uno spazio vettoriale multidimensionale. La vicinanza di vettori rappresentanti parole diverse all'interno dello stesso spazio vettoriale lascia trapelare le somiglianze profonde che le legano, permettendone il riconoscimento e la predizione da parte degli algoritmi attraverso il ricorso ad operazioni algebriche coinvolgenti i vettori:

> A well-known example involves predicting the vector queen from the vector combination king − man + woman, where linear operations on word vectors appear to capture the lexical relation governing the analogy, in this case OPPOSITE-GENDER. The results extend to several semantic relations such as CAPITAL-OF (paris−france+poland ≈ warsaw) and morphosyntactic relations such as PLURALISATION (cars − car + apple ≈ apples).[20]

La distanza tra i due vettori verrà calcolata attraverso la tecnica della similarità del coseno e dunque valutando l'ampiezza dell'angolo che intercorre tra essi.

Il modello dei Word2Vec si articola in due tipologie di architettura: *Continuous Bag-of-Words* (*CBOW*) e *Skip-Gram*. Esse possono considerarsi complementari; difatti se l'obiettivo della prima consiste nel predire una parola basandosi sulle informazioni contestuali a disposizione, quello della seconda è individuare le forme che accompagnano una parola partendo solo dalla conoscenza della stessa.

Il principale parametro utilizzato per cogliere le relazioni che si instaurano tra le parole è la loro frequenza di co-occorrenza in una frase. Tali frequenze vengono organizzate in matrici a loro volta utilizzate poi dal modello per generare gli embeddings.

Attraverso l'analisi delle frequenze di co-occorrenza, è possibile inferire la similarità semantica tra due parole o l'ambiguità delle stesse; ad esempio, le parole "re" e "regina" potrebbero avere frequenze di co-occorrenza elevate in resoconti storici, lasciando trapelare la loro similarità semantica. Ancora, parole con frequenze di co-occorrenza elevate in contesti diversi potrebbero avere significati ambigui proprio perché difficilmente ancorabili ad un'unica situazione comunicativa.

[20] Vylomova, E., Rimell, L., Cohn, T., & Baldwin, T. (2015). Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. *arXiv preprint arXiv:1509.01692*.

L'utilizzo della co-occorrenza come parametro principale per individuare le relazioni tra le diverse parole rappresenta anche il limite intrinseco del sistema, il quale non ha accesso all'intera sequenza in analisi e all'informazione contestuale da essa veicolata. Conseguentemente, non è in grado di reperire tutte le sfumature semantiche di una parola utili a ricostruirne il significato specifico nei diversi enunciati. Ulteriori limiti consistono nell'incapacità di gestire in modo ottimale le dipendenze a lunga distanza e le parole rare o dal significato ambiguo.

Un netto miglioramento in tal senso è garantito dal passaggio all'architettura Transformer, alla base degli odierni Large Language Models.

2.6 Large Language Models

Con il termine 'Large Language Models' (*LLM*), facciamo riferimento a modelli linguistici relativi all'ambito del NLP allenati su vaste porzioni di testo al fine di sviluppare al meglio la comprensione e la produzione di linguaggio human-like; questo secondo obiettivo li rende parte di ciò che è stata definita *Intelligenza Artificiale Generativa*, vale a dire un tipo di intelligenza artificiale in grado di generare testo, immagini o altri media in risposta a determinati comandi (*prompt*).

L'emergere di questi modelli nel corso degli ultimi anni è riconducibile all'avvento di algoritmi sempre più innovativi e complessi e alla maggior disponibilità di grandi quantità di testo appartenente alle più svariate categorie (*data diversity*) e facilmente accessibile (basti pensare al materiale presente in rete), di cui il modello può servirsi per migliorare le proprie prestazioni.

Alla base dello sviluppo delle potenzialità di questi modelli vi è un periodo di *pre-training* durante il quale il sistema viene allenato attraverso il ricorso ad un dataset linguistico molto ampio, che gli consente di raccogliere informazioni sulle strutture fondamentali della lingua (vocabolario, sintassi…). Il processo di apprendimento, che consente al sistema di acquisire consapevolezza anche in merito ai rapporti grammaticali e semantici che si instaurano tra le parole, passa attraverso le stesse tipologie presentate in §2.2.

Nel caso di task riguardanti la generazione di testo, il periodo di addestramento consente al sistema di imparare a predire la parola successiva in una frase basandosi sul contesto precedente, attribuendo dunque una probabilità più o meno elevata alla

ricorrenza di ciascuna parola; ne sono un esempio i suggerimenti proposti dai moderni portali di posta elettronica durante la scrittura di una mail.

Per migliorare la performance del modello su task specifici, alla fase di pre-training può seguire un ulteriore periodo di addestramento in cui il sistema lavora con un dataset più piccolo funzionale alla specializzazione in un determinato compito, quale può essere la traduzione o la sintesi di testi; questo passaggio prende il nome di *fine-tuning*.

Tuttavia, i Large Language Models sono frequentemente utilizzati sfruttando approcci di *few-shot o zero-shot learning*.

Il primo prevede la somministrazione al modello di un ristretto numero di esempi relativi al task che esso sta per svolgere, al fine di favorirne una familiarizzazione. Il secondo invece, non coinvolge alcuna fase di adattamento specifico per il sistema, che è chiamato ad eseguire il compito sfruttando unicamente la capacità di generalizzazione delle sue conoscenze pregresse.

All'interno delle due macrocategorie trovano posto anche il *Chain of Thought prompting* (CoT, riconducibile al few-shot learning) e lo *zero-shot CoT* (riconducibile allo zero-shot learning), teorizzato in Kojima, Gu, Reid, Matsuo, & Iwasawa (2023).

Il primo approccio implica il fornire al modello alcuni esempi di ragionamento step-by-step al posto di semplici scambi coinvolgenti domanda e risposta. Il fine ultimo consiste nell'elicitare la capacità analitica del sistema, guidandolo nella scomposizione di processi complessi in passaggi più semplici, che verranno poi esposti sequenzialmente all'utente affinché ne valuti la correttezza.

Il secondo approccio dimostra che è possibile ottenere simili risultati anche senza ricorrere ad esemplificazioni, lavorando direttamente sul prompt. Difatti, integrando in esso formule volte a stimolare lo svolgimento graduale di un task quali "Let's think step by step", il modello produce ragionamenti coerenti arrivando alle giuste conclusioni, diversamente da quanto avviene affidandosi al tradizionale paradigma di *zero-shot*.

Per comprendere meglio il funzionamento dei Large Language Models è però necessario descriverne l'architettura alla base.

2.6.1 Da RNN all'architettura *Transformer*

Ciò che rende possibile la generazione di una parola sulla base dell'input ricevuto è la rete neurale a fondamento dell'intero modello, l'algoritmo che ne influenza l'output.

La *recurrant neural network* (RNN), è una tipologia di rete neurale (facente parte della macrocategoria di ANN presentata nel paragrafo 2.4) in grado di analizzare un determinato input procedendo sequenzialmente e affidandosi a connessioni ricorrenti, grazie alle quali l'output di un passaggio precedente diviene input di quello successivo.

Tali connessioni ricorrenti sono garantite da due componenti principali: l'*encoder* e il *decoder*. L'encoder processa la parola e la "traduce" in una rappresentazione vettoriale che diventa poi il nuovo input preso in carico dal decoder; esso a sua volta lo utilizzerà per genere un output che tenga conto delle informazioni derivanti dal passaggio immediatamente precedente conservate in un *hidden layer*. Il processo di generazione degli output viene reiterato più volte finché non si completa l'intera sequenza.

Il limite intrinseco di questo modello è rappresentato dalla sua incapacità di processare parallelamente più input, obbligandolo a procedere parola per parola e comportando un notevole aumento delle tempistiche. A ciò si aggiunge l'incapacità di gestire lunghe sequenze di parole legata all'emergere del *vanishing gradient problem*; durante la fase di backpropagation il peso assegnato ai gradienti dei layer iniziali tende a decrescere/aumentare esponenzialmente, risultando nella quasi totale impossibilità di riconoscimento e risoluzione dell'errore da parte della rete.

A partire dal 2017 viene però introdotta l'architettura conosciuta col nome di *Transformer*. Pur presentando anch'essa le componenti di *encoder* e *decoder*, esse risultano arricchite da una molteplicità di livelli caratterizzati da meccanismi di attenzione avanzati, che consentono il superamento delle difficoltà sopracitate e dunque la gestione simultanea di ampie porzioni di testo.

Come esposto in Vaswani et al. (2017), il modello riesce infatti ad elaborare informazioni relative ai singoli elementi della sequenza in modo indipendente e simultaneo, attuando così un processo denominato *parallelizzation*.

Non procedendo più sequenzialmente, e dunque abbandonando il meccanismo tipico delle RNN, l'architettura è in grado non solo di guardare contemporaneamente a tutti gli elementi che compongono una struttura ma anche di riuscire a capire a quali di essi attribuire più importanza sfruttando il meccanismo dell'*Attention* nelle sue diverse configurazioni: *Self-Attention*, *Masked Self-Attention* e *Multi-Head Self-Attention*.

La *Self-Attention* si afferma in quanto funzione fondamentale per il modello Transformer, poiché costituisce ciò che permette al sistema di comprendere su quali

elementi presenti all'interno dell'input focalizzarsi maggiormente per la generazione dell'output: «For instance, when analyzing the input "I am very" and predicting the output "happy", the relative contributions of the three words "I am very" to the output "happy" might be quantified as 0.7, 0.1, and 0.2, respectively» (Chu et al., 2024: 9).

La *Masked Self-Attention* impedisce invece al modello di accedere alle parole successive rispetto a quella in analisi adottando il mascheramento dei livelli, rendendo questa tecnica particolarmente utile per implementare la realisticità e solidità delle predizioni.

Infine, vi è la *Multi-Head Self-Attention*; essa costituisce un'amplificazione del normale meccanismo di attenzione che, attraverso la creazione di teste indipendenti, consente un'analisi simultanea e differenziata dell'input, i cui risultati vengono poi condensati in un unico output.

Un ulteriore meccanismo sfruttato dal sistema è quello del *Positional Encoding*, che consente al Transformer di ottenere informazioni sul posizionamento di una parola in una frase. Diventa necessario per preservare l'informazione temporale, che in questo tipo di architettura risulta intrinsecamente assente a causa del venir meno dell'analisi sequenziale, rendendo obbligata una marcatura esplicita delle diverse parti del discorso.

Il *Positional Encoding* è espresso attraverso un valore vettoriale che viene aggiunto all'embedding iniziale della parola, arricchendone il significato e consentendone la disambiguazione semantica da parole uguali ma che tuttavia ricorrono in posizioni diverse. Ad esempio, considerando le frasi[21] "He is a good person and does not do bad things" e "He is a bad person and does not do good things", un minimo cambiamento nell'ordine delle parole altera completamente il significato della frase.

2.6.1.1 BERT

Un esempio di applicazione dell'architettura Transformer è rappresento da BERT, sviluppato dai ricercatori di Google AI e presentato per la prima volta in Devlin et al. (2018).

[21] Gli esempi sono tratti da Chu, Z., Ni, S., Wang, Z., Feng, X., Li, C., Hu, X., ... & Zhang, W. (2024). History, Development, and Principles of Large Language Models-An Introductory Survey. *arXiv preprint arXiv:2402.06853*.

Il nome risulta essere un acronimo per “Bidirectional Encoder Representations from Transformers”, con riferimento alla natura bidirezionale del sistema che gli consente di processare una parola presente in una frase tenendo conto e del contesto precedente e del successivo.

Diversamente dai suoi predecessori, allenati sfruttando campioni di dati ridotti e relativi ad un unico task, l’addestramento di BERT avviene su un dataset molto ampio di testo e codice non etichettato (estratto dall’English Wikipedia e dal BooksCorpus), che gli permette di acquisire una conoscenza profonda del linguaggio e delle sue relazioni semantiche, rendendolo particolarmente efficace in compiti di natura analitica quali la disambiguazione dei significati di parole, il sentiment analysis e la classificazione di testi.

Il periodo di pre-training verte sullo svolgimento di due task: il *Masked Language Model (MLM)* e il *Next Sentence Prediction (NSP)*. Il primo, al quale in letteratura ci si riferisce spesso col nome di “Cloze task”, consiste nell’oscurare parte dei token presenti nell’input al fine di valutare la capacità predittiva del sistema, a cui è chiesto di individuarli. Il secondo vede il modello chiamato a scegliere tra due coppie di frasi per identificare quella seguente l’input originario.

2.6.1.2 ChatGPT

Pur riprendendo l’architettura Transformer, GPT (Generative Pre-trained Transformer) rappresenta invece, almeno inizialmente, un modello di tipo unidirezionale, in grado di generare contenuto processando le informazioni solo da sinistra a destra.

Questa precisa configurazione risponde alla diversa natura del modello, contrapposta a BERT e di tipo generativo, per la quale il modello si rivela particolarmente utile in compiti che non richiedono una comprensione profonda dei testi.

Dal primo lancio da parte di OpenAI nel 2018 ad oggi, il modello ha conosciuto numerosi miglioramenti, direttamente proporzionali all’espansione del dataset usato per il proprio addestramento. Quest’ultimo, nel passaggio da GPT 1 a GPT 4, non solo ha raggiunto dimensioni notevoli ma ha anche permesso al modello di specializzarsi nella generazione di testi appartenenti alle più svariate tipologie (dalla scrittura di mail a codici per la programmazione) grazie ai dati a cui aveva accesso e senza necessità di un addestramento più specifico.

Difatti, la fase di *fine-tuning* è stata applicata al modello solo a partire dalla terza generazione, attraverso l'utilizzo di un dataset più ristretto ma coinvolgente conversazioni reali; «this supervised learning phase involves training the model to generate appropriate responses in a conversational setting. The model learns from human-generated input-output pairs, refining its ability to provide contextually relevant and coherent responses» (Plevris et al., 2023: 4).

Nel novembre 2022 OpenAI rilascia anche ChatGPT, un agente conversazionale basato sull'omonima architettura e addestrato attraverso il ricorso al *Reinforcement learning from Human Feedback* (RLHF), una tecnica che prevede l'utilizzo di giudizi umani per la valutazione dell'accuratezza degli output emessi dal sistema al fine di migliorarne le performance.

Ad oggi ChatGPT è in grado di intrattenere conversazioni estremamente realistiche, producendo output accurati sulla base del contesto e del pubblico di riferimento grazie alla combinazione di pre-training su larga scala e fine-tuning con dati specifici.

L'interfaccia intuitiva della piattaforma contribuisce a rendere le interazioni user-friendly; l'utente ha infatti a disposizione un campo di testo in cui inserire richieste ed istruzioni sottoforma di input scritti, definiti *prompt.* Si tratta propriamente del comando che l'utente dà in input alla macchina affinché questa possa svolgere un determinato task.

La qualità dell'output emesso dal modello risente fortemente dell'accuratezza con cui è stato progettato il prompt, portando alla necessità di dedicare particolare attenzione al processo di "prompt engineering", approfondito nel paragrafo successivo.

Ricevuto l'input, il chatbot lo analizza confrontandolo con i pattern presenti nei dati su cui è stato addestrato, utilizzandoli per predire e generare una risposta coerente emessa quasi istantaneamente.

Inoltre, la piattaforma supporta anche la possibilità di rivedere e modificare le interazioni precedenti (facilitando un flusso di dialogo continuo e coerente) e di personalizzare le impostazioni predefinite relative al tono delle risposte (più o meno formale) e alle informazioni personali che desideriamo far avere al modello.

Queste funzionalità rendono ChatGPT non solo uno strumento potente per la comunicazione, ma anche una piattaforma versatile che può essere adattata a una vasta gamma di esigenze professionali e personali. La facilità d'uso e l'accessibilità

dell'interfaccia hanno infatti contribuito alla sua rapida adozione e diffusione in molteplici settori, dall'assistenza alla sfera dell'istruzione.

Ulteriori limiti e potenzialità di questo modello soprattutto in relazione al contesto didattico saranno oggetto di analisi nei prossimi capitoli.

2.6.1.2.1 Prompt Engineering

Il prompt rappresenta quindi il principale strumento attraverso cui ci interfacciamo con modelli fondati sull'Intelligenza Artificiale generativa, ed è per questo che il processo di *prompt engineering* relativo alla sua progettazione merita un'attenzione particolare.

In Eager, Brunton (2023) si sottolinea l'importanza per i docenti di padroneggiare l'arte di scrivere prompt efficaci, al fine di collaborare efficientemente con l'Intelligenza Artificiale, ottenendo contenuti diversificati e pedagogicamente validi. Presentando una panoramica di prompt concretamente utilizzati all'interno di contesti scolastici, lo studio ne illustra la costruzione procedendo di passaggio in passaggio, garantendo la replicabilità della procedura da parte dei lettori.

Innanzitutto, è necessario determinare lo scopo per cui ci si rivolge al sistema, identificando gli obiettivi che si vogliono raggiungere e la forma nella quale si desidera ricevere l'output. Si passa poi all'effettiva scrittura del comando, utilizzando un linguaggio semplice e privo di ambiguità, dotato di tutti i parametri ed i riferimenti necessari allo svolgimento del compito. Infine, il prompt viene dato in input al software e sulla base del risultato ottenuto si valuta la possibilità di modificarlo per migliorare ulteriormente la prestazione.

L'indagine individua sei componenti fondamentali da integrare nel prompt al fine di guidare ulteriormente il modello verso l'elaborazione di un output in linea con le proprie aspettative. Nello specifico si tratta di: *Verb*, *Focus*, *Context*, *Focus and Condition*, *Alignment, Constraints and Limitations*.

Il verbo utilizzato nel prompt definisce l'azione specifica che si richiede al modello, come "creare" o "analizzare". Il focus indica il tipo di contenuto che si vuole ottenere, ad esempio un report, una simulazione d'esame o una batteria di domande. Il contesto delimita lo scopo o i parametri dell'attività, specificandone l'ambito o il pubblico di riferimento; ad esempio, si potrà chiedere al modello di creare contenuti per un corso

avanzato di informatica o per un progetto di ricerca in ambito medico. Con “Focus and Condition” si precisa ulteriormente l’obiettivo del task, specificando gli elementi che l’output dovrebbe includere, come il riferimento a determinati fonti. Gli ultimi due parametri contribuiscono invece a definire eventuali limiti a cui il feedback del modello deve rifarsi (ad esempio un numero massimo di caratteri utilizzabili).

Un’esemplificazione di quanto detto è la seguente:

> Write a case study for a first-year undergraduate marketing class. The case study should illustrate the challenges faced by a small business in developing a social media marketing strategy for attracting new clients.
>
> Context: This case study will be used to support students' attainment of intended learning goals [insert learning goals].
>
> Case Study Requirements: The case study should be approximately 2000-words long, include a brief description of the business, describe the challenges faced by the business in relation to designing and delivering a social media strategy, and possible solutions, and include case questions for the students to discuss in class (Eager & Brunton, 2023: 5).

Un ricorso assiduo all’Intelligenza Artificiale da parte dei docenti richiede una grande flessibilità nei prompt da essi utilizzati, al fine di riuscire facilmente ad adattarli al contesto e agli obbiettivi didattici che di volta in volta si trovano a dover affrontare.

A tal proposito, lo studio ipotizza uno scenario futuro in cui biblioteche digitali, consultabili liberamente da tutti, inizieranno ad accogliere esempi di prompt suddivisi per tipologia e già testati nella pratica, a cui i docenti interessati possano attingere al bisogno, modificandoli ulteriormente in base alle loro esigenze.

Una riflessione più approfondita in merito al possibile utilizzo di queste risorse all’interno dell’ambiente didattico verrà condotta nel capitolo successivo.

3. Intelligenza Artificiale e didattica

3.1 Strumento utile o potenziale ostacolo?

La possibilità di un accesso gratuito e di massa a ChatGPT ha portato con sé lo sviluppo quasi istantaneo di un dibattito relativo all'utilizzo di questo strumento nella didattica, inserendosi nel più generale clima di discussione riguardante il ricorso sempre più frequente all'intelligenza artificiale anche in questo ambito.

Il confronto ha visto l'imporsi prevalente di quanti riconoscono in questa innovazione la causa del futuro declino dell'insegnamento, profetizzando un ridimensionamento del ruolo e dell'importanza dei docenti nel processo di formazione degli studenti.

Tra le persone pronte a sostenere questa visione figurano anche gli insegnanti, nella maggior parte dei casi ostili nei confronti di queste nuove tecnologie.

Le principali preoccupazioni riguarderebbero l'impossibilità di riuscire a distinguere un elaborato frutto dell'inventiva di uno studente da uno generato in pochi minuti dalla macchina, compromettendo conseguentemente i tradizionali metodi di valutazione. Un'ulteriore perplessità deriverebbe invece dall'eccessiva fiducia che i ragazzi ripongono in questi strumenti; se da un lato ciò si traduce nel rischio di credere a notizie spesso infondate (visti i bias e le allucinazioni di cui è vittima il sistema), dall'altro (secondo i più estremisti) comporterebbe un progressivo impigrimento delle loro funzioni cognitive, rendendoli sempre più proni al delegare alla macchina operazioni da essi considerate troppo dispendiose.

Per ridimensionare questa prospettiva è utile riprendere alcune argomentazioni esposte in Steele (2023), in particolare l'idea che ChatGPT, al pari di qualsiasi altro chatbot attualmente disponibile, non sia né il primo né l'ultimo strumento presente in rete ad essere potenzialmente inaffidabile; basti pensare agli articoli clickbait o ai commenti lasciati da troll e bot nei forum e nelle piattaforme social, fino ad arrivare ad alcune voci presenti in Wikipedia e sfuggite alla revisione.

Tuttavia, se per certi versi la celebre enciclopedia online può essere considerata la diretta predecessora di ChatGPT per la quantità di informazioni a cui consente di accedere liberamente (ad esempio le voci relative alla trama di un dato film o libro di cui riassumono gli aspetti salienti), allo stesso tempo si distacca da esso per la propria policy,

la quale obbliga il creatore di una voce a citare le fonti utilizzate per generarla, di fatto limitando la possibilità di inserire informazioni non veritiere.

Viceversa, ChatGPT oltre a non prevedere un sistema del genere, utilizza anche uno stile di scrittura *human-like* che contribuisce ad accrescere la sua credibilità (diversamente dagli errori grammaticali generalmente spia dei bot prima citati), alimentando la difficoltà dell'utente nel mettere in discussione l'output che gli viene proposto.

Ne deriva che ora più che mai la figura del docente si rende necessaria per insegnare ai ragazzi come valutare correttamente le fonti a propria disposizione, non demonizzando a priori gli strumenti digitali ma comprendendone potenzialità e limiti: «In the internet age, schools and universities have already learned that they must teach information literacy–the ability to check and evaluate sources[…], we must help students learn not to fear automated writing tools, but to understand and leverage them responsibly» (Steele, 2023: 2-3).

Un metodo utile al conseguimento di questo obiettivo consiste proprio nell'invitare gli studenti a sperimentare in prima persona l'utilizzo di queste piattaforme; l'insegnante potrà ad esempio chiedere ai propri alunni di far realizzare a ChatGPT l'analisi di un testo da loro precedentemente studiato, per poi dar vita ad un dibattito in classe in cui si discutono i punti deboli dell'elaborato. Ciò rende possibile e la valutazione della capacità di giudizio critico dello studente e la sua conoscenza dell'argomento in questione:

> My proposal is that we teach students to generate and critique ChatGPT summaries of key texts that we want them to understand. In my research methods courses for master's students, I ask students to read and discuss difficult research articles. What if I prompted them to ask ChatGPT to summarize an article they had already read, in terms of its methods, data sources, findings, and limitations? I could then ask them: What did GPT get right? What did it misconstrue, overlook, or invent out of thin air? Why are the misconstrued details hard for AI to understand? The purpose of such an activity is to deepen students' own understanding of the text. Sometimes ChatGPT's mistakes resemble the mistakes students make. Once students are tasked with fact-checking a computer, such mistakes become depersonalized data points. Students are now in the position of close textual readers, of experts (Steele, 2023: 3-4).

Se dunque da un lato appare chiaro che lo sviluppo di tecnologie didattiche basate sull'IA rappresenta un universo in continua ed inarrestabile espansione, dall'altro questo processo non ha da intendersi come nocivo a priori ma anzi, se opportunamente orientato, si conferma in quanto vera e propria risorsa.

3.2 Verso percorsi didattici personalizzati

Il seguente paragrafo si propone di tracciare una panoramica dei principali utilizzi dell'intelligenza artificiale in contesti didattici al fine di approfondire le potenzialità della stessa e di fornire alcuni spunti di riflessione a quanti interessati alle sue applicazioni.

Partendo dal presupposto che essa dimostra la sua utilità come strumento di supporto tanto agli alunni quanto ai docenti, divideremo i possibili impieghi sulla base del pubblico di riferimento.

Prima di procedere all'esemplificazione, è tuttavia necessario puntualizzare che molti degli strumenti digitali attualmente utilizzati in aula non sono stati inizialmente progettati con questa specifica finalità ma sono stati piuttosto adattati a questo scopo in un secondo momento.

Un esempio calzante è rappresentato da piattaforme collaborative quali Google Docs e Google Sheets o anche da utilizzi specifici di applicazioni di messaggistica come Zoom e Teams, le cui funzionalità integrate hanno consentito una simulazione dell'ambiente classe durante la pandemia, conoscendo successivamente a questa una rapida accelerazione del loro processo di implementazione.

Il principale obiettivo legato all'impiego dell'IA nella didattica è il raggiungimento di una personalizzazione dell'apprendimento fondata sull'adattamento dei contenuti e dei metodi alle esigenze dei singoli studenti, resa possibile grazie al monitoraggio costante del processo d'acquisizione e all'emissione di feedback correttivi puntuali.

La necessità di investire in un traguardo del genere inizia a delinearsi a partire dagli anni '80, grazie agli studi condotti dallo psicologo e pedagogista Benjamin Bloom, i cui risultati più rilevanti furono riportati in un articolo del 1984 dell'*Educational Researcher*.

La ricerca parte dal confronto di tre contesti di apprendimento diversi: *Conventional*, *Mastery Learning* e *Tutoring*.

La prima tipologia rappresenta una realtà scolastica di tipo convenzionale, che vede la classe composta da 30 alunni seguiti da un solo docente e periodicamente interrogati con l'unica finalità di assegnargli un voto. La seconda differisce dalla prima per l'introduzione di ulteriori test volti a valutare l'effettiva comprensione dell'argomento da parte degli alunni, a cui succedono feedback e interventi correttivi mirati. Infine

l'ultima prevede la presenza di un tutor per singolo studente o per piccoli gruppi, garantendo agli alunni la possibilità di essere seguiti nel processo di apprendimento passo dopo passo.

Ciò che emerge dall'indagine è che le prestazioni degli studenti istruiti attraverso il sistema di tutoraggio uno-a-uno mostrano una *deviazione standard* pari quasi al doppio della media ottenuta analizzando le prove del gruppo appartenente alla classe tradizionale; questo tipo di risultato si traduce non solo in performance migliori in termini di tempo e obiettivi raggiunti da parte degli alunni, ma anche in una loro partecipazione più attiva alla lezione e nello sviluppo di una maggior sicurezza nelle proprie capacità.

La principale causa della disparità è da rinvenirsi nell'impossibilità da parte dell'unico insegnante all'interno di una classe convenzionale di riservare la stessa quantità di attenzioni ad ogni alunno, finendo col non accorgersi dei bisogni di determinati elementi ed incoraggiando solo un gruppo ristretto:

> Observations of teacher interaction with students in the classroom reveal that teachers frequently direct their teaching and explanations to some students and ignore others. They give much positive reinforcement and encouragement to some students but not to others, and they encourage active participation in the classroom from some students and discourage it from others. The studies find that typically teachers give students in the top third of the class the greatest attention and students in the bottom third of the class receive the least attention and support. These differences in the interaction between teachers and students provide some students with much greater opportunity and encouragement for learning than is provided for other students in the same classroom (Brophy & Good, 1970).[22]

Diversamente, nell'approccio di tutoraggio uno-a-uno il docente riesce a diagnosticare prontamente eventuali incomprensioni ed è a sua volta "esaminato" e corretto dall'alunno qualora i metodi da lui utilizzati dovessero risultare poco efficaci:

> If the explanation is not understood by the tutee, the tutor soon becomes aware of it and explains it further. There is much reinforcement and encouragement in the tutoring situation, and the tutee must be actively participating in the learning if the tutoring process is to continue. In contrast, there is less feedback from each student in the group situation to the teacher- and frequently the teacher gets most of the feedback on the clarity of his or her explanations, the effect of the reinforcements and the degree of active involvement in the learning from a *small* number of high achieving students in the typical class of 30 students (Bloom, 1984: 11).

Ne consegue che la strada da percorrere per un miglioramento dell'esperienza didattica coincide con l'investire in un approccio che si avvicini quanto più possibile al sistema di tutoraggio uno-a-uno.

[22] Benjamin Bloom, The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring, *Educational Researcher*, 1984, p 11.

Se all'epoca della ricerca non si riusciva ad immaginare una soluzione che garantisse questi risultati senza comportare spese ingestibili e deformare l'assetto tradizionale delle classi, ad oggi lo sviluppo della tecnologia e ancor più dell'intelligenza artificiale applicata al linguaggio rende possibile l'integrazione in aula di sistemi in grado di offrire ad ogni alunno feedback costanti, personalizzati e puntuali.

3.2.1 Principali utilizzi per gli studenti

Rientrano nella categoria di *Intelligent Tutoring System* (*ITS*) tutte le piattaforme che progettano un percorso formativo personalizzato in base alle esigenze di ciascun utente, proponendo ad ogni studente che aderisce al sistema lezioni ed esercizi specifici sulla base delle informazioni raccolte attraverso attività da egli svolte precedentemente:

> While the student engages with a particular activity, the system captures thousands of data points such as what is clicked, what is typed, which tasks have been answered correctly, and any misconceptions that have been demonstrated. This data is analysed to determine the next information, activity and quiz to be delivered, thus generating a personalised pathway through the material to be learned, and the process is repeated.[23]

Un primo esempio è rappresentato da "Spark", applicazione realizzata dalla compagnia francese Domoscio. Vi è un test iniziale che serve a stabilire il posizionamento dell'alunno all'interno di una specifica scala di competenze; una volta superato, il sistema restituisce gli obiettivi futuri, i tempi e le modalità adatti a conseguirli. Oltre allo spazio dedicato allo studente, nel quale vengono inseriti i materiali generati sulla base delle proprie esigenze, ve n'è anche uno pensato per il docente, in cui sono presenti dati e feedback utili a monitorare i progressi dell'alunno.

Il web offre però anche un'ampia gamma di piattaforme deputate alla spiegazione e alla semplificazione di contenuti specifici (ad esempio paper scientifici), in cui l'utente può caricare il materiale da sottoporre ad analisi ed interfacciarsi poi con un assistente virtuale in grado di dissipare i suoi dubbi in merito, ottenendone una spiegazione sintetica e ricca di riferimenti al testo (es: MyLessonPal, Unriddle, TutorAI…). Ancora, a partire da quello stesso input è possibile chiedere al sistema di generare un campione di domande ad esso ricollegate, da utilizzare per esempio come materiale di studio in vista di una verifica.

[23] Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. European Journal of Education, 57, 551.

Funzioni simili sono garantite anche dal ricorso agli attuali Chatbot; dal più competitivo prodotto da OpenAI e in grado di supportare l'utente nella scrittura dei contenuti più disparati coinvolgendolo in scambi stimolanti, passando per *Gemini*, agente conversazionale sviluppato da Google e capace di reperire rapidamente informazioni citando fonti attendibili, fino ad arrivare a quelli integrati nei sistemi operativi (es: *Copilot* per Microsoft) e predisposti all'assistenza in diversi ambiti.

Lo studio condotto da Chen et al. nel 2022, rivela l'apprezzamento nutrito dagli studenti nei confronti di queste risorse. Essi ne confermano l'utilità identificandone numerosi punti di forza, primo fra tutti la possibilità di un'interazione costante, che garantisce feedback immediati a qualsiasi tipo di quesito (anche di carattere personale, considerata la natura intima di alcune affermazioni condivise dagli alunni) e che evita allo studente di permanere nel dubbio in attesa di qualcuno che possa aiutarlo. A ciò si aggiunge il carattere informale della comunicazione, che spinge l'alunno ad esporsi superando ostacoli quali la timidezza o la paura di avanzare domande banali, che appaiono invece frenarlo nelle interazioni dal vivo. In ultimo, alla macchina viene riconosciuta la capacità di fornire esempi concreti e molteplici fonti in relazione agli argomenti analizzati, contribuendo ad una loro maggior comprensione da parte degli studenti, i quali possono eventualmente richiedere al sistema di ricorrere ad espedienti più interattivi (come l'utilizzo di immagini e video) per soddisfare le esigenze dettate dal proprio stile cognitivo.

Le capacità generative dei chatbot si rivelano particolarmente utili agli studenti anche per ricevere stimoli in merito alla redazione di brevi testi scritti; supportati dal sistema nelle fasi di brainstorming e costruzione della scaletta tramite opportuni suggerimenti, gli alunni possono anche scegliere di affidarsi completamente ad esso, testando l'*Automatic Essay Writing* (*AEW*).

In questo caso, sviluppando la traccia propostale attraverso il *prompt*, spetterà alla macchina generare l'intero testo; esso, fungendo da modello, faciliterà poi l'identificazione e l'apprendimento delle caratteristiche relative a particolari tipologie testuali da parte degli studenti.

Sfruttando queste medesime potenzialità, lo studente può anche decidere di rivolgersi ad uno di questi sistemi per intraprendere un dibattito stimolante in merito ad un determinato argomento, chiedendo alla macchina di sposare opinioni contrarie alle

proprie. Un esercizio del genere può aiutare l'alunno a sviscerare nel dettaglio il topic in vista di un test scritto (es: saggio breve) o di un'esposizione orale, preparandolo a sostenere eventuali controargomentazioni e contribuendo al rafforzamento delle proprie idee e della capacità di argomentarle.

Un impiego che potremmo definire complementare rispetto all'*AEW* è invece l'*Automated Writing Evaluation (AWE)*, macrocategoria che fa riferimento ai processi di correzione automatica di un testo da parte dei sistemi di IA prima menzionati e che costituisce il focus della presente ricerca. Al suo interno troviamo anche l'*Automated Essay Scoring (AES)*, che comprende tecnologie orientate verso il medesimo fine valutativo ma predisposte principalmente all'attribuzione del solo punteggio.

In Foster (2019) vengono indagate le opinioni degli studenti relative all'utilizzo della piattaforma "Openessayist", progettata da alcuni alunni della Open University con l'obiettivo di fornire un feedback automatico ed istantaneo alle produzioni scritte generate dai propri utenti. Tra i principali vantaggi riconosciuti al sistema troviamo la creazione di una scaletta che funge da struttura di base del testo aiutando i ragazzi nell'organizzazione delle proprie idee. A questa segue la funzione di "ricapitolazione", che permette al software di identificare errori o ripetizioni nel lessico usato revisionando velocemente l'intero scritto e al contempo fornendo suggerimenti su come modificarlo, assicurandosi che segua correttamente la traccia assegnata.

A riprova di quanto detto, gli alunni attestano una maggior velocità nei tempi di produzione e correzione dell'elaborato:

> One unexpected finding was that whilst use of OpenEssayist did not significantly change the way students planned and wrote their assignments, it did potentially make that process quicker. RP2 commented that they actually wrote less drafts of their essay "….because it's [OpenEssayist] given me the feedback to be able to get straight to where I need to change, whereas before I didn't have that so I just relied on other people reading it and thinking I needed to change so it drastically reduced the amount of drafts I did". Indeed, three of the four students who used OpenEssayist felt that its summarisation of their assignment helped them to complete the assignment more quickly. For instance, both RP2 and RP6 commented that they always checked their assignments to ensure they had covered the points required and that OpenEssayist's summarisation process made a check of the essay quicker (Foster, 2019: 3).

Ad oggi, l'impiego dell'IA a questo scopo è ancora soggetto a sperimentazioni vista la complessità derivante dalla conduzione di un'analisi semantica da parte della

macchina. Tuttavia, in ulteriori applicazioni dell'*Automatic formative assessment* [24] quali la correzione di verifiche a risposta chiusa o la revisione delle scelte grammaticali operate in un testo, il sistema ha già dato prova della sua efficienza.

Una delle piattaforme di supporto alla scrittura più celebri è sicuramente *Grammarly*, attualmente disponibile anche come estensione web e applicabile a programmi di terze parti quali Gmail o Word.

Oltre ad effettuare una revisione superficiale del testo scritto, intervenendo su eventuali errori ortografici, Grammarly propone anche diverse riformulazioni dello stesso adattate allo stile desiderato e al contesto di riferimento.

Rimanendo nell'ambito dell'*AFA* e dell'*ITS* troviamo anche "Kaligo", applicazione utilizzata in Bonneton-Botté et al. (2020) al fine di indagare le potenzialità derivanti dall'impiego di ambienti di apprendimento digitali all'asilo. La piattaforma consente ai bambini di esercitarsi autonomamente nella scrittura delle diverse lettere dell'alfabeto servendosi unicamente del display del dispositivo (tablet) e delle proprie dita. Alla fine di ogni tentativo effettuato viene emesso un feedback da parte del sistema fondato sul riconoscimento e l'analisi dei diversi parametri relativi alla calligrafia (es: forma, dimensioni e direzioni delle linee…). A seguito della valutazione, consultabile dagli insegnanti, vengono stabiliti i task che l'alunno dovrà svolgere successivamente, calibrandoli sulla base dei risultati raggiunti fino a quel momento:

> During a written activity on Kaligo, the app's artificial intelligence checks whether the child has been successful. If not, the requested task is simplified. For example, if a child fails to join two letters in a bigram, the app automatically suggests working on each letter separately first. Pupils each work at their own pace. The children's successful and unsuccessful productions are stored in a notebook and their teachers can visualize them at any time via a dashboard. This dashboard enables teachers to manage the pedagogical scenarios they want to follow in their teaching activities. They can select several exercises (writing letters/words in capital/cursive letters, writing numbers), and in each exercise they can either choose from a stock of available words or add new words, depending on the topics tackled in class (Bonneton-Botté et al., 2020: 5).

Anche in questo caso, risorse del genere permettono di avvicinarsi sempre di più ad un'istruzione di tipo personalizzato (*one to one*) e ai vantaggi che ne conseguono, facilitando l'apprendimento da parte degli studenti e al contempo agevolando il lavoro del docente, il quale si vede normalmente limitato nella possibilità di valutare contemporaneamente e in maniera accurata tutte le performance del gruppo classe.

[24] Macroarea comprendente tutte le tecniche e le piattaforme in grado di offrire un feedback agli output degli studenti.

Il ricorso all'IA in ambito didattico garantisce inoltre un incremento dell'accessibilità, consentendo di superare eventuali barriere derivanti da difformità linguistiche e/o riconducibili a disabilità di vario genere; accanto alle traduzioni istantanee fornite e dai Chatbot e da piattaforme specializzate in grado di cogliere le diverse sfumature semantiche sfruttando reti neurali molto estese (es: DeepL), troviamo software di trascrizione automatica quali "Notas Meas", progetto promosso dall'Università degli Studi di Padova e attualmente in fase di sperimentazione.

Mediante questo tool si riesce a trascrivere in tempo reale quanto spiegato dai docenti a lezione sfruttando un sistema di riconoscimento vocale e di trasmissione dei dati di alta precisione; il sistema genera automaticamente un file contenente gli appunti in formato Google Docs, garantendo l'accesso a docenti e studenti dai loro molteplici dispositivi.

Le note, suddivise per insegnamenti e data delle lezioni, potranno essere successivamente modificate ed integrate con ulteriori informazioni, facilitando la vita accademica a tutti quegli studenti che soffrono di DSA o presentano disabilità visive e uditive.

3.2.2 Principali utilizzi per i docenti

Prima di illustrare i possibili utilizzi dell'Intelligenza Artificiale da parte dei docenti, è bene specificare che ancor prima dell'avvento del Deep Learning, e conseguentemente di modelli quali ChatGPT, l'Intelligenza Artificiale aveva già iniziato a lasciare un segno significativo nell'ambito del *NLP* e dell'*ITS*, sebbene le tecniche fossero meno avanzate rispetto a quelle odierne.

Fondata prevalentemente sui modelli statistici presentati in §2.3 e su regole e conoscenze esplicitamente codificate, l'IA veniva perlopiù utilizzata in compiti di traduzione automatica, question answering, informational retrieval o sentiment analysis. In relazione alle sue applicazioni didattiche, essa consentiva ad esempio di tenere traccia delle prestazioni degli studenti nel tempo, registrandone le risposte. Ciò permetteva il successivo adattamento da parte del docente dei contenuti da proporre in classe, attenendosi a quanto registrato dall'algoritmo.

Ad oggi, sia che si tratti del noto chatbot che di tool appositi fondati sulla medesima architettura (GPT), l'intelligenza artificiale permette ai docenti di semplificare

e migliorare attività quotidiane quali la pianificazione delle unità di apprendimento e la valutazione di test ed elaborati svolti dagli studenti, favorendo la generale implementazione delle metodologie impiegate nell'insegnamento.

Nell'ambito della pianificazione, un docente può ad esempio avvalersi di ChatGPT per farsi aiutare nella costruzione di un'unità didattica. Il sistema guiderà l'utente nella compilazione definendo le diverse fasi di progettazione e suggerendo eventuali contenuti e obiettivi sulla base delle informazioni da egli condivise, velocizzando così l'intero processo di stesura.

Ancora, si può pensare di sottoporre a questa e ad altre piattaforme (tra cui API create ad hoc) gli eventuali testi che si vogliono proporre in classe, al fine di valutarne il grado di leggibilità ed assicurarsi dell'adeguatezza degli stessi rispetto al pubblico di riferimento. Tale potenzialità può essere utilizzata anche per programmare la somministrazione periodica di letture dalla complessità incrementale.

Un'ulteriore opportunità offerta ai docenti dal ricorso all'IA consiste nel riuscire a monitorare attivamente il livello di attenzione degli alunni durante la lezione, al fine di intervenire prontamente qualora dovesse calare.

Un primo esempio di applicazione dell'intelligenza artificiale a questo scopo si ritrova in Su et al. (2014). Lo studio elabora un sistema di rilevazione della concentrazione durante l'apprendimento (l*earning concentration detection system*), avente come parametri di riferimento la direzione dello sguardo (*eyes gaze*) ed eventuali cambiamenti relativi alla posizione adottata dallo studente mentre è seduto (*sitting position*). Integrando queste informazioni in un software fondato sull'IA viene generata una profilazione dei singoli alunni, ottenendo un quadro relativo al loro livello di attenzione mentre assistono ad una spiegazione in aula.

Indagini più recenti introducono la possibilità (sebbene più invasiva e controversa da un punto di vista etico) di monitorare questo stesso parametro sfruttando dispositivi portatili quali i caschi EEG, rendendo i dati successivamente visibili a docenti e genitori.

Tramite gli elettrodi opportunamente localizzati, i caschi EEG sono in grado di captare le diverse tipologie di onde cerebrali emesse durante un'attività senza interferire con quest'ultima; i dati raccolti vengono poi identificati ed analizzati da complessi algoritmi di machine learning, ricostruendo lo stato cognitivo del soggetto durante il periodo di osservazione:

> We experimented the performance of machine learning algorithms in distinguishing attentive, distracted, and drowsed states of the individual based on EEG signal processing. [...] it was necessary to employ Fast Fourier transform algorithms to extract Delta, Theta, Alpha and Beta brainwaves. Delta waves are related to deep sleep, unconsciousness, anesthesia, and lack of oxygen; Theta waves activity occurs when a person experiences emotional pressure, unconsciousness, or deep physical relaxation; Alpha waves are instead visible when an individual is in a state of consciousness, stillness, or rest, whereas when one is thinking, blinking or otherwise stimulated, this wave type disappears (alpha block); finally, Beta waves is evident when a person thinks or receives sensory stimulation (Massa et al., 2023: 1-2).

Strumenti del genere consentono ai docenti di monitorare l'intera classe, superando le difficoltà riscontrate in Bloom (1984) relativamente alla tempestività e all'estensione degli interventi di correzione da essi effettuati. Si rende infatti possibile identificare prontamente eventuali criticità nei metodi didattici adoperati, risolvibili attraverso il ricorso ad un trattamento egualitario ed incisivo:

> Teachers are frequently unaware of the fact that they are providing more favorable conditions of learning for some students than they are for other students. Generally, they are under the impression that all students in their classes are given equality of opportunity for learning. One basic assumption of our work on teaching is the belief that when teachers are helped to secure a more accurate picture of their own teaching methods and styles of interaction with their students, they will increasingly be able to provide more favorable learning conditions for more of their students, rather than just for the top fraction of the class. […] We attempt to provide teachers with a mirror of what they are now doing and have them develop techniques for equalizing their interactions with the students. […] The major emphasis in this work was not to change the teachers' methods of instruction, but to have the teacher become more aware of the ways in which he or she could more directly teach to a cross section of the students at each class section (Bloom, 1984: 10-11).

Allo stesso modo, le risorse precedentemente citate nell'ambito dell'*AFA* (§3.2.1) permettono ai docenti di velocizzare l'intero processo di correzione e di ottenere un "double check" dal sistema, circoscrivendo eventuali errori dovuti alla fallibilità umana (es: stanchezza, distrazione…).

Rientrano nei tools di cui un insegnante può avvalersi per correggere un compito assegnato ai propri alunni anche i software utilizzati per assicurarsi dell'originalità dello stesso e dunque programmati per la rilevazione di un eventuale plagio.

Gli algoritmi posti a fondamento delle piattaforme citate non si limitano a processare il contenuto di un testo confrontandolo con le fonti disponibili in rete attraverso il meccanismo di *pattern matching*[25], ma analizzano la semantica dei periodi al fine di evidenziarne la similarità.

[25] Il pattern matching consiste nell'esaminare una sequenza di token (pattern) per rilevarne la presenza all'interno di una sequenza più lunga (testo).

L’intelligenza artificiale si è inoltre affermata in quanto motore per lo sviluppo di numerosi ambienti di apprendimento immersivi (*Immersive Learning Environments*). In studi come Kuhail et al. (2023), le viene riconosciuta la capacità di aumentare la partecipazione degli studenti durante le lezioni, offrendogli un’esperienza di apprendimento che li aiuti a visualizzare e comprendere determinati concetti calandoli all’interno di uno scenario più realistico e dinamico: «As such, immersive learning facilitates learning using technological affordances, inducing a sense of presence (the feeling of being there), co-presence (the feeling of being there together), and the building of identity (connecting the visual representation to the self)» (Kuhail et al., 2023: 2).

Un approccio del genere contribuisce a promuovere il ruolo degli alunni da osservatori passivi a figure concretamente coinvolte in un processo di apprendimento che incontra le esigenze dei diversi stili cognitivi e che coadiuva il mantenimento a lungo termine delle informazioni nella memoria come conseguenza dell’esperienza diretta.

L’intelligenza artificiale è infatti in grado di amplificare le potenzialità di tecnologie già diffuse anche in ambito didattico quali la Realtà Virtuale (VR) e la Realtà Aumentata (AR), rispettivamente intese come ciò che consente agli utenti di muoversi all’interno di un ambiente artificiale e ciò che “altera” momentaneamente la percezione dell’utente, senza tuttavia impedirgli di continuare ad interagire con il mondo esterno nel corso dell’attività (es: Google Maps).

Un esempio relativo all’applicazione delle suddette risorse a scopo educativo è rappresentato dal ricorso a proiezioni 3D di oggetti liberamente manipolabili dagli utenti al fine di esplorarne da vicino la configurazione nello spazio e le diverse componenti. Così una lezione di biologia potrebbe prevedere l’osservazione diretta di strutture molecolari e cellule, laddove una di matematica quella di grafici a torta e solidi geometrici, contribuendo a rendere più comprensibili concetti che, se relegati alla dimensione teorica, rischiano di risultare eccessivamente complessi.

In un'accezione più ristretta, l’etichetta di “ambienti d’apprendimento immersivi” è estendibile anche a quei sistemi che consentono simulazioni interattive pur non ricorrendo all’ausilio di strumenti quali i visori.

Ad esempio, utilizzando ChatGPT o altre piattaforme specializzate allo scopo e basate sull’intelligenza artificiale quali “HelloHistory” (che dispone di un catalogo di personaggi preimpostati), si può attuare una finzione storica chiedendo al sistema di

impersonare una figura nota per aver influenzato il corso degli eventi. A partire da quel momento l'alunno potrà interagirvi ponendogli domande di ogni tipo (magari sulle imprese compiute o sulle opere scritte), a cui la macchina risponderà "tenendo a mente" lo stile del personaggio e le sue vicissitudini biografiche.

Questo tipo di interazione consente agli alunni di calarsi appieno nell'atmosfera di una data epoca, contribuendo ad accorciare le distanze in termini geografici e temporali, al contempo consentendo un'assimilazione più rapida e divertente delle informazioni ad essa relative.

3.3 Opinioni sull'uso di ChatGPT nell'istruzione

Lo scenario presentato all'inizio del capitolo in relazione alle opinioni dei docenti sull'utilizzo dell'intelligenza artificiale merita un ulteriore approfondimento. Difatti, seppur gran parte delle preoccupazioni nutrite da questi ultimi appaiano largamente condivise, esse non coincidono con la totalità delle opinioni presenti. Inoltre, quando sottoposte ad un'analisi più dettagliata volta ad indagarne le ragioni alla base, lasciano trapelare la volontà degli insegnanti di continuare a sperimentare e mettersi alla prova.

Quanto detto trova conferma nello studio di Rakowski et al. (2023), in cui, attraverso la somministrazione di un sondaggio in scala Likert[26], viene testata la familiarità dei docenti con l'intelligenza artificiale di tipo generativo (*GAI*) e più nello specifico con ChatGPT, analizzando le opinioni sull'efficacia dello strumento e la predisposizione ad incrementarne l'utilizzo in classe attraverso la proposta di impieghi concreti.

Diversamente da quanto emerso in studi precedenti, relativi ad altre tipologie di supporti digitali, l'indagine rivela una generale apertura degli insegnanti verso l'adozione di ChatGPT e non mancano proposte sul suo possibile impiego in aula.

Questo risultato si spiega in parte con il venir meno dei principali ostacoli normalmente relativi all'implementazione di *ILE* e *ITS*, vale a dire i costi elevati per l'acquisto delle apparecchiature e l'oneroso processo in termini di tempo ed energie funzionale alla familiarizzazione del docente con queste ultime.

[26] Ideata dallo psicologo statunitense Rensis Likert nel 1932, è una tecnica psicometrica di misurazione dell'atteggiamento. Definite un certo numero di affermazioni (*items*), ad ognuna di esse verrà attribuito un valore che riflette l'accordo/disaccordo da parte dell'utente rispetto a quanto detto.

La gratuità di ChatGPT grazie al rilascio della versione free, il facile accesso tramite registrazione web alla piattaforma e l'utilizzo indipendente da supporti terzi (si necessita solo di una connessione internet stabile e di un dispositivo in grado di connettervisi), incentiva infatti i docenti all'uso, contribuendo ad influenzarne in positivo la percezione dello strumento.

Nonostante permangano alcune perplessità relative al confine da tracciare tra impegno individuale dell'alunno e intervento del chatbot, ciò che emerge è la consapevolezza da parte degli insegnanti di trovarsi di fronte a un cambiamento epocale che essi si dimostrano in grado di gestire ed apprezzare solo quando opportunamente guidati nell'esperienza, sviluppando una conoscenza dello strumento che consenta la rassicurazione dei loro principali timori:

> Akin to earlier technologies, a lack of teacher training and preparation leads to trepidation and fear in terms of integrating technologies and AI in practice (Ally, 2019; Wang et al., 2021, Yang & Chen, 2022). However, as Kuleto et al. (2022) suggested, by being increasingly exposed to AI (e.g., through professional development training), educators might become more confident and therefore eager to incorporate AI into their teaching practice. Teachers need not fear that technology will replace them, but they should be keenly aware of the implications of adopting emerging technologies in education. New technologies should transform teaching in creative and innovative ways. Many educators currently possess a moderate level of awareness regarding GAI. However, by elevating this collective level of understanding, it is possible to mitigate certain initial apprehensions (Rakowski et al., 2023: 331).

Ancora, una survey condotta in Hosseini et al.(2023) durante un convegno relativo all'utilizzo dell'IA, e nello specifico di ChatGPT, in ambito sanitario, scolastico e di ricerca, rivela, attraverso le opinioni di 420 votanti, i numerosi benefici che essi attribuiscono all'utilizzo del chatbot nella sfera dell'istruzione.

Non solo vengono elencate specifiche applicazioni dello strumento nella didattica di determinate discipline (quali la possibilità di semplificare le leggi giuridiche per quanti interessati al diritto o di guidare nella programmazione informatici e studenti di design avvezzi alla generazione di codice), ma se ne riconoscono i possibili impatti positivi soprattutto nel livellamento delle difficoltà linguistiche riscontrate ad esempio da quanti costretti ad interagire in una lingua diversa da quella nativa.

In scenari del genere, il software contribuirebbe da un lato ad una semplificazione dei contenuti proposti agli studenti, facilitandone la comprensione, e dall'altro ad un arricchimento di quanto gli alunni stessi producono con le loro limitate risorse:

> Since students' scientific abil-ities should not be overshadowed by their insufficient language skills, ChatGPT was seen as asolution that could help fix errors in writing and accordingly, an

> instrument that can supportstudents who might be challenged by writing proficiency—specifically those not writing intheir native language. Another useful application was "adding the fluff" to writing (i.e., detailsthat could potentially improve comprehension), especially for those with communication chal-lenges. Structuring and summarizing existing text or creating the first draft of letters of appli-cation with specific requirements were also mentioned among possible areas where ChatGPTcould help students. Another mentioned possibility was to use ChatGPT as a studying tool that (upon further improvements and approved accuracy) could describe medical concepts at aspecific comprehension level (e.g., "explain tetralogy of fallot at the level of a tenth grader") (Hosseini et al. 2023: 6-7).

Le stesse tendenze sono confermate attraverso un'indagine incentrata sull'analisi della letteratura attualmente disponibile in merito all'utilizzo dell'Intelligenza Artificiale Generativa nella didattica. La ricerca, condotta da Locky Law nel 2023 attraverso il metodo *PRISMA*[27], prende in considerazione i report pubblicati dal 2017 al 2023, con l'obiettivo di restituire un quadro dettagliato dell'intero processo di evoluzione e diffusione della risorsa.

L'analisi dei paper lascia emergere un atteggiamento di apertura verso l'utilizzo di queste tecnologie, alle quali viene riconosciuto il merito di consentire un'esperienza di studio "su misura", attraverso la generazione di contenuti adatti al proprio livello e per questo tanto più utili quando coinvolti in processi di apprendimento di una L2, come avviene per l'ESL (English as a foreign language):

> The potential benefits of the use of AI dialog systems for EFL education, includes customisable input and complexity and instant feedback. […] AI-powered collaborative and interactive language learning tools can enhance student engagement in EFL teaching, improve student learning outcomes, and increase teacher satisfaction. In addition, ChatGPT has the potential to promote autonomy and personalized learning in language education via its human-like conversation and tailored language learning assistance. Learners can even regulate their learning processes, set objectives, and make decisions about their language acquisition.[…] ChatGPT can enhance conventional pedagogies and improve the efficiency of EFL teachers in grading while providing more accurate and insightful feedback (Law 2023: 4-5).

Anche in Jeon et Lee (2023), attraverso un'indagine condotta sul lungo periodo in una scuola elementare coreana, vengono approfondite le opinioni dei docenti in merito all'utilizzo di ChatGPT nella didattica, confermando i trend precedentemente evidenziati.

Lo studio risulta particolarmente significativo in quanto coinvolge docenti che prima di quel momento non avevano mai utilizzato lo strumento, offrendo un quadro delle sfide e dei traguardi legati ad una prima familiarizzazione con esso.

[27] La sigla "Prisma" sta per "Preferred Reporting Items for Systematic Reviews and Meta-Analyses" e si riferisce ad una serie di linee guida stilate da esperti al fine di aiutare i ricercatori a realizzare report in modo trasparente e rigoroso.

Dopo la partecipazione ad un seminario tenuto da esperti e volto a fornire ai docenti le linee guida necessarie ad interfacciarsi col chatbot, essi vengono lasciati alla propria sperimentazione in classe (della durata di due settimane) con l'unico obbligo di tenere nota degli scopi per i quali si ricorre a ChatGPT e dei prompt utilizzati per espletarli.

Al termine del periodo di utilizzo, i docenti ne confermano l'efficacia soprattutto in quanto interlocutore in giochi di ruolo finalizzati all'apprendimento della lingua e generatore di materiale didattico personalizzato sulla base del pubblico di riferimento e dei contenuti da affrontare. Esempi di prompt generati a questo scopo sono: «I want to do a two-member role play. This role-play is for practicing the question-and-answer structure, "What are you doing?" and "I'm …" Can you make some scripts to do the role-play using only simple English?» e « I'm an elementary student in Korea and I don't use English so well. Can you use only simple words if you have the vocabulary list of Key Stage 1?».[28]

Alcuni insegnanti dichiarano inoltre di aver testato il software anche come assistente personale degli studenti (i quali erano liberi di utilizzarlo per sciogliere dubbi relativi a grammatica e uso del vocabolario) e ne confermano l'utilità.

Nonostante tutti i partecipanti alla ricerca riconoscano le potenzialità didattiche di ChatGPT, appaiono ugualmente concordi nell'affermare che si tratta di un tool capace di supportare l'insegnante ma non di sostituirlo. I docenti saranno tuttavia chiamati a ripensare la didattica tradizionale per poter usufruire al meglio dei benefici garantiti dal software:

> All participants agreed that using ChatGPT would require teachers to become highly skillful at managing the plethora of teaching resources aforded by the technology so as to design creative, organized, and engaging lessons. With ChatGPT, teachers would now have signifcantly more control over creating and revising lesson materials than in the past, when they had access to fewer teaching resources and had to manually develop and revise materials. However, the teachers also noted that the increased availability of helpful resources would not automatically raise the quality of instruction. They mentioned that it depended on human teachers' pedagogical knowledge and judgment in selecting materials appropriate to their students' needs and how efectively they used them within their instructional contexts. (Jeon et Lee 2023: 15884-15885).

Essi risultano infatti consci di dover anzitutto imparare come formulare in modo corretto una domanda per poter ottenere il miglior risultato dal chatbot, ed evidenziano la necessità di educare anche gli studenti allo stesso scopo. Questi ultimi non dovranno

[28] Gli esempi sono tratti da (Joan et Lee 2023).

temere né nascondere l'utilizzo del software, ma piuttosto imparare come servirsene per incrementare consapevolmente le loro conoscenze:

> Teachers have to make a classroom atmosphere where using the technology itself is not ethically wrong but where the behavior of hiding how they got information from it and pretending that the knowledge gained from it is entirely their own is wrong. Students do not need to feel that they should hide their use of the chatbot. Rather, we have to acknowledge students' eforts to use the chatbot but only in a way that they can further their own learning. To do this, I believe we have to frst make it acceptable for students to openly talk about what they obtained from the chatbot and how they used it to facilitate their learning experience (Jeon et Lee 2023: 15887).

Alla luce delle riflessioni appena condotte riguardo ai potenziali utilizzi di ChatGPT all'interno di ambienti didattici, nel capitolo successivo verrà presentato l'oggetto di indagine descrivendone la metodologia impiegata e le opinioni dei docenti coinvolti nel lavoro.

4. Presentazione del caso di studio

4.1 Valutazione automatica del testo scritto e studi correlati

Come accennato nel capitolo precedente, lo sviluppo di sistemi fondati sui LLM e la loro contemporanea diffusione ha consentito un incremento degli studi relativi all'utilizzo di queste risorse in ambito educativo, focalizzandosi in particolar modo sulla loro efficacia in relazione a task di *AWE* (§ 3.2.1).

Ne sono un esempio ricerche quali Mizumoto et al. (2023) e Naisimith et al. (2023), in cui si sceglie di valutare rispettivamente la capacità di GPT-3.5 e GPT-4 di attribuire un voto ad una breve produzione scritta. Al contempo, si indaga la sensibilità del sistema in relazione a parametri linguistici quali la coerenza; questi parametri sono inclusi nelle rubriche di valutazione somministrate al software per la formulazione del giudizio finale, eventualmente giustificato tramite *rationale*.

Le rubriche, costruite ad hoc, sono relative tanto a dati quantitativi quanto qualitativi, come ad esempio la percentuale di fenomeni di overlapping e coreferenza presenti nel testo o la chiarezza dell'esposizione e la capacità di argomentazione mostrata dallo scrivente:

> Following previous research on linguistic correlates of human rating scores, we considered a range of linguistic features at the levels of lexis, phraseology, syntax, and cohesion. […] We selected these measures based on previous studies in each domain, which demonstrated utilities of the features in each of the following domains: lexical diversity (Kyle et al., 2021; McCarthy & Jarvis, 2010; Zenker & Kyle, 2021), lexical sophistication (Crossley et al., 2018), syntactic complexity (e.g., Lu, 2010; Wolfe-Quintero et al., 1998), fine-grained syntactic features (Kyle & Crossley, 2018), verb-argument construction measures (Kyle & Crossley, 2017), and textual cohesion measures (e.g., Crossley et al., 2019) (Mizumoto et al. 2023: 5).

Al fine di determinare il peso che la conoscenza di questi parametri detiene nell'emissione di un output che sia il più preciso possibile, in Mizumoto et al. (2023) l'esperimento viene riprodotto includendo o escludendo i suddetti parametri di volta in volta. Il medesimo approccio verrà adottato anche all'interno della nostra indagine.

Tra gli step fondamentali alla buona riuscita del task rientra una corretta formulazione del *prompt*, cioè il comando attraverso il quale si indirizza il sistema all'esecuzione di un compito. Nel caso specifico, quest'ultimo coincide col chiedere alla macchina di esprimere un giudizio (esteso o sottoforma di punteggio) relativo ai testi sulla base delle indicazioni fornitele, eventualmente riformulandoli apportando delle

correzioni. I giudizi vengono poi confrontati con le revisioni proposte da annotatori umani, al fine di verificarne qualità ed affidabilità.

Quanto emerge dallo studio è una moderata congruenza tra il feedback generato dal sistema e quello umano. Oltre al confronto dei punteggi attribuiti da entrambe le parti, in Naisimith et al.(2023) viene operata anche una comparazione dei *rationali*, analizzando contenuto e struttura delle analisi proposte e dal software e dagli annotatori umani dimostrandone la similarità:

> Comparing the content of the two rationales, there is a great deal of consistency, with both addressing the clarity, flow, structure, and effect on the reader. For example, both rationales describe how the writer's position is initially presented and provide a specific example. The two rationales also note the same main weakness relating to the lack of development of the second point. The two rationales then move on to describe how discourse markers are used to achieve local coherence, even highlighting the same two examples of Firstly and Secondly. Examples of coherence negatively affected by language inaccuracies are then given, though different examples are used to exemplify this point in the two rationales. Finally, both rationales summarize the reason for the overall satisfactory effect on the reader. Likewise, in terms of style, the GPT-4 rationale has clearly adopted the examples and followed the guidelines from the prompt. The rationales use terminology such as the writer (rather than the author/student/learner), are written in the 3rd person, and are within the desired length range. The overall format of the rationale is also consistent, starting with an overall statement of coherence, moving to discuss each of the coherence subconstructs in turn, then closing with an overall description of the effect on the reader (Naisimith et al. 2023: 398).

Inoltre, le predizioni in termini di punteggio si rivelano più efficaci laddove vengono tenuti in considerazione i fenomeni linguistici prima citati.

Tuttavia, le motivazioni offerte dal software in relazione all'attribuzione di un determinato voto risentono talvolta dei bias di cui esso stesso è vittima, rivelando delle falle nella *chain-of-thoughts* (CoT) e culminando nell'assegnazione di un voto sbagliato in determinati casi.

Se da un lato ciò limita l'efficacia della correzione, dall'altro il sistema non manca di restituire riferimenti immediati e puntuali al testo, rendendo il giudizio emesso particolarmente utile nei task finalizzati all'apprendimento della lingua e confermandone l'utilità non in quanto strumento sostitutivo dell'insegnante ma in quanto supporto al lavoro di quest'ultimo.

4.2 Metodologia di ricerca

Ispirandogli agli studi presentati nel paragrafo precedente, l'indagine oggetto di questa tesi propone a sua volta di testare il ruolo di ChatGPT nel processo di *AWE*, sottoponendo alla valutazione del software un corpus composto da elaborati raccolti ai

fini della ricerca all'interno di una classe terza dell'Istituto secondario Giacomo Zanella, situato a Padova.

La ricerca, condotta nell'arco di tutto l'anno scolastico, si è articolata in diverse fasi:

- **Familiarizzazione con lo strumento**

  Appurata la volontà dei docenti di prendere parte all'indagine, è stata in primis valutata la loro familiarità nei confronti dell'Intelligenza Artificiale, attraverso la compilazione di un sondaggio che verrà presentato in §4.2.1. Alla luce dei risultati ottenuti, si è ritenuto opportuno procedere con una lezione guidata volta ad illustrare agli insegnanti modalità ed utilizzi del software.

- **Introduzione ai temi oggetto d'analisi**

  Parallelamente, gli studenti della classe selezionata per la ricerca sono stati introdotti dagli insegnanti al testo argomentativo (banco di prova per lo sviluppo di argomentazioni coerenti e coese), preparandosi alla successiva fase di scrittura. La scelta di questa precisa tipologia testuale ai fini dell'indagine è da ricollegare alla volontà di testare non solo l'abilità del sistema nell'offrire un giudizio adeguato al testo datogli in input, ma anche la capacità di individuazione e valutazione del livello di coerenza caratterizzante gli scritti, rendendo possibile un approfondimento sull'idea stessa che la macchina mostra avere di questo elemento.

- **Raccolta del corpus oggetto di analisi**

  Stabilito il campione di studenti su cui effettuare l'analisi in accordo con l'insegnante, sono stati raccolti 34 elaborati, due per studente. Gli elaborati sono stati prodotti dagli studenti come compito per casa e valutati dalla docente di riferimento secondo una griglia di valutazione predefinita, che sarà discussa nel paragrafo 4.6. I testi raccolti sono stati in seguito trascritti in formato digitale per permetterne la valutazione automatica.

- **Selezione e configurazione degli strumenti**

  La valutazione automatica del corpus tramite ChatGPT è stata condotta ricorrendo a una delle versioni più recenti e performanti del software (4.0), utilizzando un approccio *zero-shot* (§2.6) per lo svolgimento del task. Se da un lato questa scelta si è data come obbligata, viste le dimensioni circoscritte

del corpus a nostra disposizione, dall'altra si è rivelata particolarmente interessante per le implicazioni relative all'esperimento e alla sua riproducibilità. Difatti, tale approccio è quello che meglio riflette le normali dinamiche di utilizzo tra utente non esperto e macchina, permettendoci di evidenziarne aspetti importanti e di fornire un pattern replicabile a quanti interessati a testare il software in prima persona. Definiti i parametri di giudizio dei testi in analogia alla griglia di valutazione definita dalla docente, sono stati poi costruiti tre diversi prompt da fornire al modello per consentire lo svolgimento del task.

- **Analisi dei risultati**

  I risultati forniti dallo strumento, secondo le diverse tipologie di prompt, sono stati poi confrontati con quelli della docente. Questo confronto ha fatto emergere similarità e differenze, lasciando spazio a riflessioni relative all'efficacia e all'affidabilità dello strumento, nonché alle potenzialità derivanti dal suo impiego in questo ambito.

I paragrafi successivi illustreranno nel dettaglio la composizione della popolazione e del corpus, nonché le fasi preliminari dell'indagine e la scelta e l'degli strumenti utilizzati in quest'ultima.

4.3 Il contesto: Istituto Secondario Giacomo Zanella

Quanto detto finora in merito all'opinione che i docenti nutrono nei confronti dell'IA, trova conferma in un questionario proposto ad alcuni insegnanti della scuola secondaria di primo grado "Giacomo Zanella", parte dell'Istituto Comprensivo Rosmini con sede a Padova.

Rispetto ad altre strutture presenti sul territorio, la Zanella si distingue per l'impiego di metodologie didattiche innovative e la creazione di contesti d'apprendimento sperimentali. Avvalendosi del supporto di strumenti digitali (quali i tablet offerti in dotazione agli alunni), l'istituto si impegna a garantire un approccio all'istruzione quanto più concreto e stimolante possibile (da intendersi nell'ottica presentata in §3.2.1 e §3.2.2), favorendo lo sviluppo delle competenze specifiche e trasversali dei singoli studenti.

Tali presupposti, sommati alla disponibilità e all'interesse dei docenti verso la materia, hanno reso questo istituto il luogo perfetto in cui condurre l'indagine sul tema, preludio del successivo progetto di ricerca.

Il sondaggio somministrato[29] è volto a testare le posizioni degli insegnanti e la loro familiarità nei riguardi dell'Intelligenza Artificiale, rapportandola al proprio ruolo e alle eventuali esperienze vissute in aula.

Il campione si compone di un ristretto gruppo di docenti impegnati nel monitoraggio e nella coordinazione delle attività svolte dalla classe coinvolta nello studio.

Essi dimostrano di essere consapevoli delle potenzialità didattiche dello strumento, rivelandosi quasi del tutto favorevoli al suo utilizzo (60%) e asserendone l'efficacia in questo ambito (80%).

Due degli intervistati dichiarano inoltre di averne usufruito per la progettazione di lavori da proporre in classe. Ad esempio: «Ho utilizzato AI per redigere un testo di argomento storico (Risorgimento) per offrirne ai miei studenti una sintesi» (Intervistato anonimo, 2024, Utilizzo dell'IA nella didattica).

Appaiono tuttavia concordemente preoccupati dalla possibilità che gli studenti possano farne un utilizzo improprio, ed è la volontà di impedirlo che li sprona a comprendere appieno il funzionamento dello strumento.

A tal proposito, la maggior parte dei docenti intervistati lamenta la mancanza di una formazione specifica al riguardo da parte degli organi competenti, in parte giustificando lo scetticismo relativo alla possibilità di sfruttare l'intelligenza artificiale per offrire un'esperienza didattica personalizzata agli alunni (80%).

Quanto riportato rende evidente la necessità di educare gli insegnanti, ancor prima che gli studenti, all'uso dello strumento, evidenziandone limiti e potenzialità al fine di garantirne un utilizzo consapevole in classe da parte di tutti.

4.4 Popolazione e campione

Il campione coinvolto nella ricerca comprende 19 studenti facenti parte di una classe terza di secondaria di primo grado. La selezione di questo specifico campione

[29] Il sondaggio è consultabile al seguente indirizzo: Utilizzo delle IA nella didattica: - Moduli Google.

rispetto all'intera popolazione di riferimento (tutti gli studenti dell'istituto) è avvenuta sulla base delle indicazioni e della disponibilità dei docenti.

Gli studenti, aventi differenti nazionalità e inseritisi in classe in periodi diversi, rendono l'ambiente molto variegato dal punto di vista linguistico e socioculturale.

Al fine di tracciarne al meglio i profili in quanto parlanti, è stato loro proposto un questionario basato sul modello *INVALSI* e focalizzato sulle loro origini e abitudini linguistiche. Le domande, riportate integralmente nell'Appendice A, mirano ad approfondire la nazionalità degli studenti e degli elementi stretti del loro nucleo familiare, al fine di ricostruirne la storia linguistica.

Il quadro che emerge dall'analisi delle risposte rivela che circa la metà degli alunni sono nati da genitori non italiani. Tra questi, vi è chi è nato in Italia e chi invece vi è approdato in un secondo momento, tanto nella prima infanzia quanto nella prima adolescenza.

Questo secondo scenario, che si afferma in quanto maggioritario, vede la prima scolarizzazione del bambino avvenire nel paese d'origine, con conseguenti implicazioni sulla sua successiva integrazione scolastica. Difatti, egli dovrà non soltanto abituarsi ad un nuovo sistema educativo, ma anche sottoporsi ad una valutazione di titoli e competenze che spesso lo costringe alla permanenza in una classe di grado inferiore, andando potenzialmente ad alimentare atteggiamenti di isolamento e frustrazione.

La natura bilingue dei ragazzi si afferma poi nella netta distinzione da essi operata tra la lingua L2 parlata all'interno della dimensione scolastica ed amicale (italiano) e quella alla quale ricorrono per interagire nel proprio contesto familiare (L1).

Ulteriore conseguenza di ciò è la necessità di ricorrere ad aiuti esterni (es: doposcuola) quando in difficoltà con lo svolgimento dei compiti, a causa delle barriere linguistiche che si interpongono tra i genitori e le materie oggetto di studio.

Tra le nazionalità maggiormente presenti sul numero totale degli studenti di seconda generazione che compongono la classe, troviamo quella indiana e quella cinese. In particolar modo, la classe ha accolto una ragazza indiana all'inizio dell'anno accademico, mentre un'altra coppia di studenti cinesi (un ragazzo e una ragazza) si è aggiunta nel corso dell'ultimo semestre.

Al momento del loro arrivo nella scuola (coincidente col loro arrivo in Italia), nessuno dei tre studenti era in grado di parlare, leggere o scrivere in italiano.

Se ad oggi, grazie a dei corsi pomeridiani di potenziamento, la ragazza indiana dimostra di aver acquisito le basi della lingua italiana, in un primo momento è stato tuttavia indispensabile ricorrere ad una terza lingua (l'inglese) per comunicare con lei. Lo stesso mezzo non si è invece rivelato efficace con gli studenti cinesi, i quali, a distanza di mesi, continuano a necessitare l'ausilio di strumenti di traduzione automatica direttamente dalla loro lingua madre.

A causa delle sopracitate difficoltà linguistiche, alcuni studenti sono stati pertanto esclusi dalla redazione dei testi raccolti nei campioni.

4.5 Raccolta del corpus

Ai fini dell'indagine è stato costruito un corpus composto da due gruppi di testi relativi alla tipologia "testo argomentativo" e raccolti a distanza di tre mesi l'uno dall'altro (marzo-giugno).

Il primo gruppo (gruppo A) conta 17 testi non vincolati ad una traccia precisa; ogni studente era infatti libero di scegliere la tematica da affrontare sulla base dei propri interessi. Ciò nonostante, la maggioranza degli alunni ha comunque dimostrato una forte sensibilità nei confronti di tematiche sociali attualmente al centro dei dibattiti pubblici.

L'obiettivo perseguito dalla docente nell'assegnare un compito del genere era quello di far esercitare gli studenti nell'argomentazione rigorosa delle proprie idee, al contempo permettendo la familiarizzazione con gli elementi fondamentali di un testo argomentativo (tesi, antitesi, argomentazioni a favore e controargomentazioni...) in vista della prova d'esame da sostenere alla fine dell'anno scolastico.

Il secondo gruppo (gruppo B), composto anch'esso da 17 testi, vede i ragazzi chiamati a scegliere tra due tracce: una relativa al ruolo attualmente attribuito agli influencer nella nostra società e l'altro focalizzato sull'avvento e la diffusione degli ebook, a discapito dei libri cartacei.

La possibile difficoltà riscontrata dagli studenti nell'esecuzione del compito va rapportata non solo al vincolo tematico, ma anche alla necessità di utilizzare all'interno del proprio lavoro le fonti riportate nei documenti contenenti la traccia. Questo step si dimostra infatti fondamentale per riflettere sulla capacità dell'alunno di costruire un'argomentazione lineare e ricca di riferimenti puntuali ai dati di cui è in possesso.

4.5.1 Trascrizione dei testi

Una volta raccolti, i testi sono stati anonimizzati e digitalizzati in vista delle successive fasi di analisi. L'operazione è resa possibile dall'esistenza di strumenti di *Automatic Speech Recognition* (*ASR*), alla base di tecnologie quali Siri, Alexa, Google Home, Cortana…

In particolare, parliamo di *Speech to Text Tools* in relazione a strumenti che permettono una "traduzione" dal parlato allo scritto attraverso una decodifica del segnale acustico emesso dal parlante.

Per la presente ricerca si è scelto di ricorrere alla funzionalità "Dettatura Vocale" integrata in Google Docs. Tramite l'accesso del software al microfono del dispositivo in utilizzo, vengono captate ed analizzate le onde sonore. Segue l'estrazione delle caratteristiche del segnale (ad esempio l'ampiezza, la frequenza e la durata dell'onda) e la successiva conversione di queste in vettori processabili dal sistema.

I vettori vengono poi confrontati con sequenze acustiche predefinite, rappresentanti i vari fonemi di una lingua, sfruttando gli algoritmi probabilistici alla base del software per identificare la natura del suono e gli eventuali matching con i pattern già noti.

Man mano che avviene la ricostruzione dei fonemi e delle parole, il sistema elabora predizioni in merito alla probabilità di una data sequenza, analizzandone la frequenza in un determinato contesto linguistico sfruttando i modelli presentati in §2.3.

Terminata la decodifica e combinati i risultati e del modello acustico e del modello linguistico per ottenere la sequenza con probabilità di verifica più elevata, il software procede alla sua trascrizione testuale.

La resa dell'operazione dipende da diversi fattori, quali: la qualità del suono emesso, su cui può incidere il rumore ambientale o l'inadeguatezza degli strumenti utilizzati per la registrazione, l'incapacità del parlante di pronunciare chiaramente le parole, talvolta a causa della presenza di un forte accento o di fenomeni tipici del parlato (es. parlato continuo), ed infine la lingua nella quale avviene la registrazione, poiché l'output della macchina risulta più accurato nella trascrizione di lingue o varietà che ha già avuto modo di osservare durante il periodo di pre-training.

Ciò nonostante, l'avanzamento di tecnologie fondate sul deep learning, la possibilità di trattare i dati sfruttando sistemi in grado di migliorarne la qualità,

l'addestramento su dati linguistici diversificati e l'integrazione di feedback umano nella valutazione delle analisi hanno portato ad una maggiore robustezza dei sistemi di ASR moderni.

Nel caso specifico, le prestazioni del modello si sono rivelate abbastanza efficaci, riuscendo ad individuare con precisione la maggior parte delle parole a patto che venissero scandite con particolare attenzione e in assenza di rumori di fondo. Ciononostante, si è talvolta reso necessario intervenire manualmente per correggere errori legati perlopiù alla punteggiatura e all'ortografia di determinati calchi e forestierismi inerenti ai temi trattati.

4.5.2 Profiling linguistico dei testi

Prima di procedere alla valutazione dei testi tramite lo strumento, si è deciso di effettuarne un'analisi linguistica, al fine di restituire un quadro quanto più dettagliato possibile dei fenomeni linguistici riscontrabili al loro interno.

A questo scopo si è scelto di ricorrere all'applicazione web di **Profiling-UD**, risultante dalle ricerche e dalle sperimentazioni condotte presso l'Istituto di Linguistica Computazionale "Antonio Zampolli". Ispirata dalla metodologia di monitoraggio linguistico dei testi esposta in Montemagni (2013), la piattaforma si sviluppa prendendo come riferimento il framework delle **Universal Dependencies**[30].

Universal Dependencies (UD) è un progetto collaborativo internazionale volto ad individuare un insieme di linee guida per l'annotazione morfosintattica delle lingue naturali, al fine di raggiungere una standardizzazione nel processo di annotazione marcatura che supporti indagini tipologiche e faciliti il confronto e lo sviluppo di strumenti di analisi linguistica anche per lingue a più basse risorse, tramite approcci "cross-lingual".

Ponendo al centro del modello le relazioni di dipendenza funzionali che intercorrono tra i diversi elementi presenti in una frase, vengono quindi definite delle etichette universali relative alla funzione sintattica che mette in relazione due termini, specificandone la natura.

[30] De Marneffe, Marie-Catherine, et al. "Universal dependencies." Computational linguistics 47.2 (2021): 255-308.

L'applicazione dei principi del framework è ravvisabile nelle vaste *treebanks* annotate, relative a lingue diverse e liberamente fruibili. Esse vengono primariamente utilizzate per addestrare modelli funzionali al NLP.

Principale obbiettivo di Profilig-UD è garantire un'analisi linguistica attraverso il ricorso a due diverse fasi: una prima, relativa all'annotazione linguistica ed espletata da **UDPipe**[31] e una seconda, eseguita dalla componente di profilazione linguistica del software e basata su script scritti in Python, che si avvale delle informazioni ricavate dall'annotazione per monitorare la presenza di determinati fenomeni linguistici.

L'interfaccia web permette all'utente di copiare ed incollare il testo di cui si desidera effettuare l'analisi, di condividere un file in formato .txt o addirittura un'intera cartella zippata. Prima di avviare la scansione, l'utente dovrà assicurarsi di selezionare correttamente la lingua del file dato in input.

Ad analisi avvenuta, il software restituirà un file contenente il testo annotato in formato .CoNLLU separato da tabulazioni, una tabella relativa al profiling linguistico in formato .csv, con ogni valore assegnato ad una cella a sé, ed infine un file in formato .txt contenente la legenda utile all'interpretazione delle *features* (tratti linguistici) evidenziate.

Le informazioni ottenute tramite il secondo step consentono di monitorare aspetti che spaziano dalla lunghezza del testo e delle parole in esso presenti alla sua varietà e densità lessicale, con un focus anche sulle strutture e le dipendenze sintattiche.

Nel caso specifico, si è deciso di selezionare un numero limitato di parametri in confronto alla moltitudine offerta dal software, prediligendo i tratti più rilevanti allo scopo della ricerca.

I parametri tenuti in considerazione sono: numero di frasi presenti nel testo, numero di token, lunghezza media del periodo (espressa in token), lunghezza media della parola (espressa in caratteri), indice di Gulpease, densità lessicale, rapporto tra type e token (calcolato sulla base delle prime 100 parole che compongono il testo) ed infine lunghezza media delle relazioni di dipendenza (calcolata come distanza in parole tra testa e dipendente).

[31] UDPipe, presentata in (Straka et al., 2016), è una pipeline riconducile al framework di Universal Dependencies (UD) (e dunque predisposta all'analisi di molteplici lingue) in grado di effettuare operazioni funzionali ad una successiva analisi linguistica più mirata, quali: segmentazione del testo in frasi, tokenizzazione, POS tagging, lemmatizzazione e dependency parsing.

Tutti i parametri sono calcolati sfruttando **Profiling-UD**, ad eccezione dell'indice di Gulpease, calcolato tramite **READ-IT**[32]. Si tratta di un indice di leggibilità automatica concepito per la lingua italiana da un gruppo di linguisti dell'università di Roma "La Sapienza" negli anni '80. Nonostante si basi su caratteristiche molto superficiali del testo, quali appunto la lunghezza media della frase e la lunghezza media delle parole, è in grado di fornire una prima approssimazione del livello di complessità lessicale e sintattico dei testi, ed è molto usato anche in ambito di costruzione di risorse didattiche.

L'indice, espresso in percentuale, permette di evincere il grado di leggibilità di quest'ultimo, rapportandolo a delle scale di valori prestabilite.

Difatti, testi con un indice inferiore a 80 sono difficili da leggere per chi ha la licenza elementare, quelli con un indice inferiore a 60 risultano complessi a chi ha la licenza media ed infine quelli con un indice inferiore a 40 rappresentano letture ostiche anche per chi detiene un diploma superiore.

I dati che emergono dal profiling linguistico dei testi precedentemente anonimizzati (raccolti in Appendice G) ci restituiscono un quadro del campione che combacia appieno con le nostre aspettative. Guardando al numero delle frasi e dei token, ci accorgiamo subito che quest'ultimo risulta superiore nel primo gruppo, consentendoci di affermare che la scrittura dei ragazzi risente in positivo della padronanza dell'argomento che si è scelto di trattare. L'indice di Gulpease rimane pressoché invariato nelle due raccolte, con un range che oscilla da un massimo di 60 ad un minimo di 43,4 nel secondo gruppo. La difficoltà di lettura, se rapportata al contesto d'appartenenza (dunque una classe terza in procinto di effettuare il passaggio alle scuole superiori), appare abbastanza in linea con i valori predefiniti. Tuttavia, soprattutto nei testi del secondo gruppo, soggetti ad una traccia e pertanto più tecnici, si registra un leggero incremento del grado di complessità del testo.

Anche densità lessicale e rapporto tra type e token, rispettivamente indicativi della predominanza delle parole contenuto su quelle funzionali e della varietà lessicale presente nello scritto, si mantengono pressoché costanti nei due gruppi, con una media di 0,6 nel primo caso e 0,7 nel secondo. Entrambi i valori sono da rapportare a degli intervalli che

[32] Tool di analisi testuale elaborato dall'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) del Consiglio Nazionale delle Ricerche (CNR) con sede a Pisa. È liberamente accessibile tramite il link: http://www.italianlp.it/demo/read-it/.

spaziano da 0 ad 1, dove 1 rappresenta il maggior grado di variabilità e densità lessicale acquisibile dal testo.

Infine, la misura relativa alla lunghezza media delle catene di dipendenza, calcolata sulla base del numero di parole che si interpone tra la testa di un periodo e gli elementi da essa dipendenti (esclusa la punteggiatura), oscilla variabilmente tra una minima di 2,002 e una massima di 2,951, sottolineando ulteriormente la complessità sintattica di alcuni periodi e la conseguente incidenza sull'indice di lettura.

Metodologie di monitoraggio linguistico del testo, che permettono di estrarre caratteristiche linguistiche quali quelle offerte da Profiling-UD, sono utilizzate in diversi contesti applicativi, che spaziano dalla stilometria alla valutazione automatica della complessità di un testo. Ad esempio, studi hanno dimostrato la possibilità di determinare la lingua madre di un individuo attraverso lo studio di testi da lui prodotti nella propria L2, sulla base dell'estrazione di caratteristiche linguistiche identificative di determinati tratti stilistici:

> Since the organization of the First Shared Task on Native Language Identification (Tetreault et al., 2013), stylistic characteristics of L2 writings have been used to model L1 features and predict the native language of the writer of a given document. Also in this scenario, Profiling– UD features turned out to be effective in i) classifying the L1 of the writer and ii) reconstructing the linguistic profile of L1 starting from L2 productions (Cimino et al., 2013) (Brunato, Cimino, Dell'Orletta, Montemagni, & Venturi, 2020, p. 7150).

Ciò risulta quantomai interessante alla luce della multiculturalità che caratterizza il gruppo classe coinvolto nell'indagine. Tuttavia, la limitata estensione del dataset a nostra disposizione impedisce un approfondimento di questo aspetto all'interno dello studio attuale.

4.6 Griglia di valutazione

La griglia di valutazione[33] realizzata a cura della docente di riferimento e consultabile in Appendice B, riporta i parametri da tenere in considerazione nella correzione dei testi.

Vengono identificati cinque indicatori che spaziano dalla sfera ortografica e semantica (relativa alla correttezza e alla precisione del lessico adoperato) a quella

[33] Si tratta di una griglia risultante da un libero adattamento della docente dell'articolo "Un'esperienza concreta di didattica cooperativa a distanza" di Patrizio Vignola, pubblicato sul sito www.scintille.it nell'aprile 2021.

sintattica e contenutistica, volta a valutare la capacità dello studente di condurre un'esposizione in modo chiaro ed approfondito, con opportuni riferimenti ad elementi che sostengano la propria ipotesi.

Ad ogni indicatore viene associato un punteggio da 0 a 2, che corrisponde ad un diverso livello di competenza dimostrato dallo studente. Nello specifico, i quattro livelli individuati sono: in via di acquisizione, di base, intermedio ed avanzato.

La griglia è stata utilizzata come strumento di correzione di entrambi i gruppi di testi dalla docente, attribuendo ai singoli scritti un voto complessivo, risultante dalla media dei valori ottenuti nelle diverse voci.

Per quanto riguarda la correzione operata da ChatGPT, si è ricorso all'utilizzo di tale griglia solo in una delle tre valutazioni effettuate dal sistema, al fine di testarne l'influenza sull'efficacia dell'output finale attraverso il confronto coi risultati raggiunti nelle altre casistiche. La griglia è stata dunque opportunamente integrata all'interno del prompt necessario allo svolgimento del task.

4.7 Configurazione dei Prompt

Per condurre l'indagine si è resa necessaria la costruzione di tre prompt principali, ciascuno riadattato in una seconda versione che ne assicurasse l'efficacia anche nell'analisi dei testi a traccia vincolata. Operata la scelta del sistema da utilizzare (ChatGPT 4.0), si è dunque proceduto alla redazione dei comandi.

Il primo prompt rende conto della volontà di osservare il comportamento "naturale" del sistema nell'esecuzione di task di *Automatic Writing Evaluation (AWE)*; difatti, chiedendogli di operare una correzione dei testi senza tuttavia fornirgli dei criteri di valutazione specifici a cui rifarsi, si riceve un insight sul suo modus operandi e sui parametri presi in considerazione, successivamente confrontandoli con quelli selezionati dalla docente:

- **Prompt 1a:**

  "Assegna un voto a ciascuno dei testi argomentativi che ti darò in input. I testi sono in totale 17. Si tratta di testi argomentativi a traccia libera. Puoi attribuire un voto da 0 a 10, dove 0 corrisponde al valore più basso e 10 al più elevato. Il voto deve essere l'espressione di un giudizio complessivo.

Non ti verrà chiesto di motivare l'attribuzione di un determinato voto. Ricordati di associare ad ogni voto il titolo del testo a cui fa riferimento."

- **Prompt 1b:**
  "Assegna un voto a ciascuno dei testi argomentativi che ti darò in input. I testi sono in totale 17. Si tratta di testi argomentativi vincolati ad una traccia. Ti verrà fornito il documento contenente le tracce tra cui gli alunni erano liberi di scegliere. Puoi attribuire un voto da 0 a 10, dove 0 corrisponde al valore più basso e 10 al più elevato. Il voto deve essere l'espressione di un giudizio complessivo. Non ti verrà chiesto di motivare l'attribuzione di un determinato voto. Ricordati di associare ad ogni voto il titolo del testo a cui fa riferimento."

Il secondo prompt si configura come una "soluzione intermedia" rispetto agli altri due comandi. Mentre da un lato chiede al software di elaborare giudizi basati sull'analisi di molteplici parametri, come avviene nel terzo prompt, dall'altro questi parametri non coincidono con quelli selezionati dalla docente e presenti nella sua rubrica valutativa.

Gli aspetti considerati dal software in questo caso emergono, infatti, da un'analisi approfondita del feedback implicito ottenuto attraverso il primo prompt, che riflette le conoscenze incorporate nel sistema. Sotto specifica richiesta, è lo stesso ChatGPT ad elencare i criteri automaticamente tenuti in considerazione per la valutazione di un testo, identificando cinque categorie principali: **chiarezza** e **coerenza**, **argomentazione**, **originalità** e **profondità**, **stile** ed **espressione** ed infine **impatto emotivo** e **persuasione**.

Una volta individuati questi parametri, è stato chiesto al sistema di costruire una griglia ad hoc (consultabile in Appendice B) che richiamasse la struttura di quella utilizzata dalla docente in merito ai punteggi attribuiti a ciascun livello di competenza. Pur non essendo perfettamente sovrapponibili, entrambe le griglie sembrano focalizzarsi sui medesimi elementi.

Ecco allora che ci sarà perfetta aderenza tra i **Prompt 2a** e **2b** e i **Prompt 3a** e **3b**, con l'unica differenza relativa alla rubrica di valutazione a cui essi devono fare riferimento.

Infine, il terzo prompt mira ad equiparare IA e insegnante, offrendo alla macchina gli stessi criteri di riferimento utilizzati dall'umano nella formulazione dei giudizi, così da poter condurre un'analisi in parallelo:

- **Prompt 3a:**

  "Assegna un voto a ciascuno dei testi argomentativi che ti darò in input. I testi sono in totale 17. Si tratta di testi argomentativi a traccia libera. Puoi attribuire un voto da 0 a 10, dove 0 corrisponde al valore più basso e 10 al più elevato. Il voto dovrà essere elaborato sulla base della rubrica di valutazione che ti fornirò. Non ti verrà chiesto di motivare l'attribuzione di un determinato voto. Ricordati di associare ad ogni voto il titolo del testo a cui fa riferimento."

- **Prompt 3b**:

  "Assegna un voto a ciascuno dei testi argomentativi che ti darò in input. I testi sono in totale 17. Si tratta di testi argomentativi vincolati ad una traccia. Ti verrà fornito il documento contenente le tracce tra cui gli alunni erano liberi di scegliere. Puoi attribuire un voto da 0 a 10, dove 0 corrisponde al valore più basso e 10 al più elevato. Il voto dovrà essere elaborato sulla base della rubrica di valutazione che ti fornirò. Non ti verrà chiesto di motivare l'attribuzione di un determinato voto. Ricordati di associare ad ogni voto il titolo del testo a cui fa riferimento."

Definite le tipologie di prompt da utilizzare nell'indagine, esse sono state somministrate a ChatGPT. I diversi output prodotti dal modello saranno oggetto d'analisi nel capitolo successivo.

5. Discussione

5.1 Giudizi dell'insegnante

Servendosi della griglia di valutazione (Appendice B) realizzata appositamente per l'indagine, la docente coinvolta nella ricerca ha valutato i due gruppi di testi argomentativi, soffermandosi prettamente sulla capacità degli alunni di riuscire ad esporre in modo coerente e solido (supportato da riferimenti mirati) le proprie posizioni in merito ad un determinato argomento.

I giudizi, culminanti nell'attribuzione di un voto compreso nella fascia 0-10, sono stati riportati in forma anonima all'interno dell'Appendice C.

L'assegnazione di un punteggio per un determinato parametro non è accompagnata da feedback personalizzati o da riferimenti al testo, ma si limita a giustificare la scelta riportando le indicazioni di riferimento presenti all'interno della rubrica di valutazione. Lo stesso vale per l'attribuzione del giudizio finale.

L'analisi del primo gruppo di testi conferma la forte polarizzazione della classe, che si traduce in una netta contrapposizione tra gravi insufficienze (relative ad un numero ristretto di soggetti in difficoltà) e ottimi voti (raggiunti dalla restante parte degli alunni). Questa tendenza risente in parte del pregresso biografico degli alunni, messo in luce attraverso il questionario sulle abitudini linguistiche (Appendice A).

L'analisi dei giudizi attribuiti al gruppo B riconferma questa condizione, segnalando tuttavia un miglioramento generale delle prestazioni degli studenti e conseguentemente delle valutazioni ad esse relative, ulteriormente giustificabile se ricondotto al periodo di chiusura dell'anno accademico.

5.2 Output del sistema

Di seguito sono riportati i feedback del sistema in relazione all'oggetto di indagine. I risultati, distinti in base al prompt utilizzato, verranno presentati e discussi nei relativi sottoparagrafi.

5.2.1 Risultati derivanti dal primo prompt

Il primo prompt mira ad ottenere dal modello un'analisi olistica dei testi fornitigli in input, senza offrirgli indicazioni specifiche a cui rifarsi; rappresenta pertanto la

tipologia di interazione più intuitiva ed immediata che può instaurarsi tra utente e software e rende conto dei criteri impliciti da esso utilizzati per la valutazione.

I risultati ottenuti (consultabili in appendice D), tanto nel primo quanto nel secondo gruppo, si dimostrano abbastanza lontani dai giudizi attribuiti dalla docente.

Per una maggiore precisione, si è scelto di misurare la correlazione esistente tra i giudizi emessi dal modello e quelli della docente, ricorrendo ai coefficienti di correlazione di Pearson e Spearman. La misurazione della correlazione lineare e monotona esistente tra i due gruppi di giudizi coinvolge, nel primo caso, direttamente il valore assoluto delle variabili, mentre nel secondo richiede dapprima il calcolo dei ranghi. Il risultato viene espresso tramite un valore compreso in un intervallo tra -1 e 1, dove 1 indica una perfetta congruenza tra le variabili.

Per quanto riguarda il coefficiente di correlazione di Spearman, il valore è pari a 0,7 per il gruppo A e 0,2 per il gruppo B. Mentre il valore relativo al coefficiente di Pearson è 0,7 per il gruppo A e 0,1 per il gruppo B.

La quasi totale incongruenza in merito ai giudizi del secondo gruppo può trovare una spiegazione nella poca importanza attribuita dal modello alla necessità per gli alunni di doversi attenere ad una traccia, finendo col non penalizzare quanti vi si allontanavano.

Inoltre, se per i testi del gruppo A le valutazioni assegnate dal sistema si discostano in positivo da quelle della docente, consentendo una media dell'8, nei testi del gruppo B si registra la tendenza inversa. Difatti, esclusi i casi più problematici, il modello non attribuisce voti superiori all'8.5, laddove l'insegnante arriva a giudicare svariate produzioni con 9 e 9.5.

La discrepanza col modello potrebbe risultare tanto dal mancato vincolo nei confronti di una rubrica a cui attenersi, quanto dall'assenza di una fase di training precedente lo svolgimento del task.

Questa supposizione potrebbe essere validata dai risultati ottenuti nelle altre casistiche, spesso più vicini a quelli dell'insegnante. Ciò è ricollegabile tanto alla presenza di parametri fissi a cui riferirsi, tanto all'iterazione continuativa del task all'interno della medesima conversazione col software, permettendone una familiarizzazione con le dinamiche di correzione.

Ipotesi di questo tipo verranno successivamente approfondite in §5.3.

5.2.2 Risultati derivanti dal secondo prompt

Il secondo prompt è finalizzato ad attribuire ai testi tanto un giudizio complessivo quanto un punteggio relativo ai singoli parametri, ricorrendo all'utilizzo della griglia di correzione appositamente creata da ChatGPT e consultabile in Appendice B.

Se da un lato si vuole quindi testare l'efficacia del sistema nella sua "autonomia", dall'altro il confronto tra questi risultati e quelli assegnati dalla docente, e più nello specifico un'eventuale corrispondenza tra questi valori, potrebbe decretare la validità della macchina nello svolgimento di task di questo tipo.

Innanzitutto, è bene sottolineare come le categorie individuate dal software ricalchino spontaneamente quelle presenti nella rubrica costruita dall'insegnante, evidenziando la predisposizione della tecnologia a procedere nella correzione servendosi di un approccio molto simile a quello tenuto dai docenti.

Difatti, in entrambe le griglie ritroviamo parametri relativi alla correttezza ortografica e sintattica, alla precisione del linguaggio utilizzato e alla chiarezza dell'esposizione.

ChatGPT dimostra inoltre di riconoscere l'importanza che un fattore quale la coerenza (§1.3) assume all'interno del testo argomentativo, dedicandovi un apposito parametro e restituendone una definizione che rinvia ad una continuità logica e semantica esistente all'interno dei periodi.

La rubrica del chatbot si differenzia ulteriormente da quella della docente per la presenza di due categorie inedite: "Impatto emotivo e persuasione" e "Originalità e profondità". La prima va oltre la mera valutazione della solidità dell'argomentazione, misurando la capacità dello scrivente di emozionare e convincere il lettore a sposare il proprio punto di vista. La seconda considera invece la non convenzionalità delle tesi proposte, combinandola con il rigore adoperato nel discuterle. Considerando la tipologia dei testi analizzati, questi parametri si affermano in quanto particolarmente significativi.

I giudizi relativi ai testi appartenenti al gruppo A (Appendice E), confermano la tendenza generale della macchina ad attribuire voti abbastanza alti se comparati con quelli della docente. Ciò si traduce nell'attribuzione diffusa di valutazioni che spaziano dall'8.5 al 10, voto mai utilizzato dall'insegnante, e nella totale assenza di gravi insufficienze.

Quanto detto non è da riferirsi solo al giudizio complessivo ma anche ai voti assegnati ai singoli parametri, dove il modello non attribuisce mai punteggi inferiori a 1.

Per quanto riguarda i testi del gruppo B (Appendice E), più che rilevare un miglioramento generico delle prestazioni della classe (con il relativo aumento dei voti che esso comporta) le valutazioni attribuite dal modello ricalcano i pattern già presenti nell'analisi del gruppo A, in accordo con quanto riportato nel paragrafo precedente.

Pertanto, la correlazione tra i due feedback, calcolata mediante il ricorso ai coefficienti di Pearson e Spearman, assume il valore di 0,7 in relazione al gruppo A e 0,6 in relazione al gruppo B per quanto riguarda la correlazione lineare (Pearson) e di 0,7 in entrambi i gruppi per quanto riguarda la correlazione monotona (Spearman), confermando una generale congruenza tra le opinioni del modello e quelle della docente.

Vale invece la pena soffermarsi sulla forma personalizzata che assumono le valutazioni. Difatti il software, in qualche modo contravvenendo a quanto esplicitato dal prompt, va oltre la mera attribuzione di un punteggio per i singoli parametri, "giustificandone" l'assegnazione attraverso il riferimento al contenuto del testo. Ciò consente non solo di rilevare eventuali punti deboli, ma anche di offrirne potenziali soluzioni; ne sono un esempio i seguenti giudizi relativi al parametro "Argomentazione", punteggio 1.5:

- "Le argomentazioni sono valide e presentano sia i vantaggi che gli svantaggi delle due modalità di acquisizione della cittadinanza. Tuttavia, potrebbero essere meglio sviluppate e supportate da ulteriori esempi concreti." (Feedback inerente a testo 4, gruppo A)
- "Le argomentazioni presentate sono valide e supportate da esempi concreti, come l'influenza negativa degli influencer sui bambini. Tuttavia, alcune affermazioni potrebbero essere approfondite e ulteriormente sviluppate per una maggiore incisività." (Feedback inerente a testo 17, gruppo B)

5.2.2.1 Confronto tra i due risultati

Un'eventuale corrispondenza tra i risultati ottenuti attraverso il primo ed il secondo prompt, permetterebbe di constatare la robustezza e la rigorosità procedurale del modello. Difatti, i parametri dati in input nella seconda casistica coincidono con quelli tenuti implicitamente in considerazione dal sistema nell'attribuzione del giudizio complessivo risultante dal primo prompt.

Guardando ai testi appartenenti al gruppo A, si registra una corrispondenza di giudizio in 7 casi su 17, con un'oscillazione in positivo del punteggio tra 0,5 e 1,5 per i restanti voti. Per i testi facenti parte del gruppo B invece, i voti coincidono solo in 4 casi su 17; tuttavia, il range relativo alle discrepanze negli altri punteggi si restringe, presentando un'oscillazione in positivo compresa tra 0,5 e 1, che denota un'accuratezza maggiore nelle prestazioni del software.

Pur non essendoci un'aderenza perfetta, il modello dimostra quindi una certa coerenza nell'attribuzione dei punteggi. Ciononostante, questa merita di essere inquadrata all'interno di un'analisi più ampia relativa ad eventuali bias di cui il sistema può potenzialmente essere vittima.

5.2.3 Risultati derivanti dal terzo prompt

Il terzo prompt è ciò che più concretamente ci consente di istituire un parallelismo tra docente e tecnologia, attraverso l'utilizzo da parte del chatbot dello stesso strumento con cui si trova ad operare l'insegnante.

Nella fattispecie, somministrata alla macchina la griglia di valutazione costruita dalla docente, essa procede alla correzione dei testi secondo i parametri indicati. I giudizi sono stati riportati in Appendice F.

Anche qui viene confermata la tendenza del software ad evitare di attribuire giudizi troppo negativi agli elaborati. Infatti, se rapportati ai voti della docente, in alcuni casi ben sotto la soglia della sufficienza, quelli del chatbot risultano più generosi, non assegnando mai voti al di sotto del 4.

Inoltre, anche in caso di scritti non particolarmente brillanti, la principale preoccupazione del sistema rimane quella di fornire un feedback che sia il più costruttivo possibile per lo studente, andando a sottolineare quanto di positivo è stato fatto nel lavoro e citandone i punti deboli solo per spronare l'alunno a migliorarsi in futuro.

Ciò emerge chiaramente nei commenti conclusivi e spontanei (non direttamente richiesti dal prompt) che il sistema decide talvolta di assegnare ai testi, come avviene in: «*, hai presentato un'analisi approfondita e ben strutturata sul ruolo degli influencer, esaminando diverse prospettive e fornendo argomentazioni convincenti. Il tuo testo è chiaro e ben articolato, anche se alcune piccole correzioni di sintassi potrebbero

migliorarne la fluidità. Ottimo lavoro nel fornire una visione completa della questione!» (Giudizio di ChatGPT in relazione a un testo appartenente al gruppo A).

Nella correzione relativa al primo gruppo, l'analisi lascia emergere una scarsa coincidenza tra le valutazioni complessive assegnate dal chatbot e quelle emesse dalla docente, con un'oscillazione in positivo di circa 2 punti per la maggior parte dei giudizi, a conferma della tendenza prima sottolineata.

Non mancano però alcuni casi di perfetta congruenza, come avviene per il testo 16, a cui entrambi i feedback attribuiscono un giudizio pari a 9,5.

Il testo, pur non essendo vincolato ad alcuna traccia, risulta ben strutturato e coerente. Vengono forniti dati statistici e fonti autorevoli a supporto dell'argomentazione, evidenziando le criticità legate al riconoscimento dell'Hate Speech ed avanzando possibili soluzioni per imparare a riconoscere e combattere il problema:

> […] L'Hate Speech non colpisce soltanto uno o due individui, ma colpisce intere fasce di popolazione. Secondo i dati forniti da da Vox insieme all'università La Sapienza di Roma i bersagli più mirati soltanto su Twitter sono le donne (con il 63% dei tweet analizzati), gli omosessuali e migranti(entrambi con il 10% di commenti negativi) e infine diversamente abili (circa il 6%) ed ebrei (per il 2%). Luigi Curini, fondatore di Voices from the Blogs, ha analizzato 80 milioni di tweet nella seconda metà del 2016. I temi che "conquistano" il maggior numero di commenti d'odio sono riguardanti la politica, misoginia, xenofobia e omofobia. […] Il 73% degli intervistati dichiara di non avere mai postato contenuti che potrebbero essere ritenuti Hate Speech, il restante 27% lo ha fatto almeno una volta. […] Tornando ai dati di Alessandro Rosina, c'è un'aria grigia di persone che ancora non conosce l'Hate Speech. va sensibilizzato ed educata il prima possibile, e vale il 32%. C'è un 10% di irriducibili, nei confronti dei quali bisogna lavorare solo in termini di contenimento. Può sembrare che su Internet ci siano solo cose brutte, ma bisogna ricordare che spesso c'è chi prova a rendere il web un ambiente migliore: c'è una maggioranza silenziosa, e un po' stufa che crede in una vita online migliore. A loro è dedicato un manifesto per la comunicazione non ostile nato nel 2017. Il manifesto mira a sensibilizzare le persone dichiarando 10 doveri fondamentali che ognuno di noi dovrebbe prendersi per comunicare con gli altri al meglio, sia online che non. È importante contrastare i discorsi d'odio attraverso manifesti come questo, azioni e politiche che siano giuste per portare l'armonia online, e denunciare i casi di Hate Speech, in quanto si possono descrivere come vera e propria violenza verbale. Ognuno di noi come cittadino ha il diritto di parola, ma il dovere di pensare prima di pubblicare le proprie idee, riflettendo sulle espressioni da usare. Ci sono tante parole, scegliamo quelle giuste.[34]

Osservando i punteggi relativi ai singoli parametri, il software (pur arrotondando per eccesso i voti) dimostra inoltre di riuscire ad individuare gli aspetti in cui un testo risulta più carente. Ad esempio, all'interno del gruppo A, il testo 7 viene segnalato per le imprecisioni ortografiche, mentre il testo 8 per la scarsa qualità dell'argomentazione, così come avviene nella correzione proposta dalla docente.

[34] Estratti dal testo 16, Gruppo A.

Per quanto riguarda invece il secondo gruppo, si nota una maggior congruenza tra voti attribuiti dal modello e voti attribuiti dall'insegnante, sebbene non manchino alcuni evidenti casi di disparità, come avviene in relazione al testo 17.

Se la docente attribuisce allo scritto una grave insufficienza, ChatGPT lo promuove invece a pieni voti, inducendoci ad indagare più a fondo il motivo della discrepanza. Una lettura veloce del testo restituisce immediatamente il senso di indefinitezza che lo caratterizza; l'alunno fatica a mettere fuoco l'argomento principale (il ruolo degli influencer) perdendosi nell'introduzione relativa all'avvento delle nuove tecnologie e costruendo una trattazione che, anche a causa della mancanza di fonti ed esempi concreti, risulta incoerente e difficile da seguire:

> La rivoluzione digitale è pienamente entrata nel nostro patrimonio sociale, culturale e influenza costantemente il nostro stile di vita. Come ogni rivoluzione diviene oggetto di valutazione, sia in senso negativo che in senso positivo, e comunque rimane oggetto di osservazione costante rispetto all'utilizzo che ne fa e alla funzione che ricopre. La rete è certamente un enorme e potente strumento per comunicare ed è in costante evoluzione nelle sue forme utilizzate. Innanzitutto, vale la pena porre l'accento sul significato di "utilizzo", poi che ogni strumento dovrebbe essere considerato come un "mezzo" che viene manovrato dall'uomo e non viceversa. Rainie and Wellman, nella loro analisi sulle tecnologie digitali, compiono una disamina attenta sul cambiamento digitale, ponendo l'attenzione su ciò che le persone fanno con le tecnologie. Malgrado la grande attenzione che viene rivolta ai nuovi gadget, la tecnologia non determina il comportamento umano, sono gli uomini a determinare il modo in cui vengono utilizzate le tecnologie. Di sicuro stiamo assistendo ad una, non consueta, ma singolare modalità di relazione all'interno dei rapporti umani: internet è anche uno strumento di socialità che ha anche assunto una natura "partecipativa" della convivenza sociale. I social network mettono in rapporto il singolo con gruppi sempre più ampi, non solo, ma le relazioni sembrano modificarsi da relazioni stabili e statiche a relazioni rapide, veloci e meno accurate. Pertanto, si tratta di un cambiamento non solo quantitativo, ma anche qualitativo. Gli autori osservano poi come in questa "socialità integrata", le relazioni mutano sperimentando nuove forme in via di evoluzione, ponendo anche l'accento sulla possibilità che esistano maggiori possibilità per ognuno di attivare e arricchire i legami sociali, ma anche allo stesso tempo maggiori responsabilità.[35]

Il giudizio della docente si afferma dunque in quanto più appropriato, permettendoci tuttavia di giustificare l'output del modello ipotizzando che esso venga erroneamente influenzato dal fatto che l'alunno menzioni una serie di concetti chiave senza in realtà approfondirli, arrivando a costruire una coerenza solo apparente che il modello fallisce ad individuare in quanto tale.

Escluso il caso specifico, pur confermando le tendenze di maggiorazione del software osservate fino ad ora, si assiste tuttavia ad una riduzione del range che separa i

[35] Testo 17, Gruppo B.

giudizi da esso emessi da quelli della docente, con un'oscillazione in positivo compresa tra +0,5 e +1 in 9 casi su 17.

Quanto detto finora trova ulteriore conferma ricorrendo ai coefficienti di correlazione di Pearson e Spearman. Tra i due giudizi si registra una correlazione lineare (Pearson) pari a 0,6 per il gruppo A e 0,5 per il gruppo B, mentre una correlazione monotona (Spearman) pari a 0,6 per entrambi i gruppi, confermando le tendenze viste anche nelle altre casistiche.

5.3 Feedback del modello vs. feedback della docente

Confrontando la performance complessiva del modello con le valutazioni emesse dalla docente, siamo in grado di affermare che ChatGPT risponde bene a task di valutazione automatica del testo (*AWE)*, soprattutto considerata l'assenza di una fase di fine tuning.

I giudizi emessi dal software risultano coerenti nei diversi trial, dimostrando la capacità del modello di analizzare testi in maniera consistente.

La discrepanza con le valutazioni della docente, che si traduce quasi sempre nell'attribuzione di un punteggio più alto ai testi da parte di ChatGPT, può essere ricollegata a diverse cause. Da un lato, essa riflette la fallibilità del sistema e le sue limitate competenze tecniche in ambiti specifici, dall'altro risente del venir meno di una conoscenza pregressa dell'alunno, che, assieme ad altre componenti tipicamente umane (prima fra tutti l'empatia) influisce inevitabilmente sul giudizio finale dell'insegnante.

Il modello, al contrario, si focalizza interamente sul testo e in ciò risiede la sua più grande potenzialità. Pur non insistendo a dovere sulle criticità degli elaborati, rischiando di lasciar passare in sordina errori importanti (soprattutto legati a grammatica e sintassi), ne evidenzia aspetti che talvolta sembrano sfuggire alla stessa docente. Ciò si osserva in maniera evidente analizzando i parametri di giudizio alternativi di cui ChatGPT tiene conto all'interno della propria griglia di valutazione, volti a considerare prospettive spesso ignorate nei giudizi tradizionali (quali l'originalità del pensiero e dello stile adottato).

Tuttavia, analizzando l'andamento delle valutazioni, si è riscontrata una maggior sensibilità da parte della docente nel riconoscere le deviazioni dalla traccia (dove tale vincolo era presente) e il focus dell'esposizione, mentre ChatGPT considerava accettabili

anche interpretazioni meno pertinenti (vedi testo 17 §5.2.3). Il modello tendeva inoltre a sovrastimare la ricchezza lessicale presente nei testi, riconoscendo spesso la presenza di un linguaggio tecnico anche laddove non era stato utilizzato.

Le valutazioni del modello si affermano però in quanto fortemente personalizzate, riportando passaggi di testo seguiti da eventuali suggerimenti o modifiche, oltre che da commenti positivi volti a spronare l'alunno.

Nonostante questi elementi siano parte indiscussa anche delle valutazioni dei docenti in aula, includere numerosi dettagli in ogni giudizio si tradurrebbe per loro in sforzi maggiori e soprattutto in un incremento delle tempistiche legate alla correzione.

Viceversa, il numero di fattori tenuti in considerazione dal modello per l'elaborazione delle valutazioni non ha alcuna ripercussione sulla latenza di risposta, rendendolo adatto ad un'analisi tanto approfondita quanto veloce dei testi.

Il principale vantaggio risiede tuttavia nella possibilità per gli alunni di iniziare un dialogo con la macchina e di mettere in discussione eventuali parametri o punteggi, assieme alle spiegazioni che li accompagnano; il tutto senza rubare tempo al docente da un'eventuale lezione. Questo dialogo interattivo permette agli studenti di comprendere meglio le aree di miglioramento e di ricevere un feedback immediato, favorendo un apprendimento più attivo e soprattutto su misura.

Ciononostante, non mancano criticità legate a bias ed imprecisioni della macchina, che corre il rischio di ripetersi, dimenticare la procedura in atto o attribuire troppa importanza a certi fattori piuttosto che ad altri, penalizzando l'alunno col giudizio finale.

Ancora, l'output del modello risulta influenzato dalla non replicabilità che caratterizza ogni interazione con l'utente, rendendo le risposte soggette a variazioni anche quando generate a pochi minuti di distanza. Questo fenomeno è dovuto al grado di casualità incorporato nel sistema al fine di migliorare la creatività delle risposte, a modifiche nei dati di addestramento e al forte ancoramento del software al contesto, al variare anche impercettibile del quale le risposte possono cambiare in modo significativo.

Inoltre, il modello fatica a costruire una visione d'insieme che funga da riferimento per bilanciare le valutazioni attribuite.

Dunque, ChatGPT, e più in generale l'intelligenza artificiale, si afferma nuovamente non in quanto alternativa al docente ma in quanto supporto al lavoro di quest'ultimo.

Se opportunamente guidati, questi sistemi possono infatti apportare un contributo significativo all'innovazione didattica, favorendo l'indirizzamento dell'istruzione verso modelli sempre più personalizzati.

5.4 Alcuni limiti dell'IA generativa

L'indagine lascia trapelare alcuni limiti del modello, legati alla natura stessa dell'architettura su cui si sviluppa.

Il limite principale riguarda l'incapacità del software di trattenere informazioni da altre conversazioni avute precedentemente con lo stesso utente (*Limited Short-Term Memory)*, influenzando la qualità dell'output.

Difatti, riprendendo lo studio di Zhang et al. (2024), potremmo definire l'interazione col modello conversazionale in quanto indirizzata allo svolgimento di un **task**, determinato attraverso il contatto con l'**environment** (rappresentato dall'interlocutore, che fornisce alla macchina i dettagli per lo svolgimento del compito), scandito da conversazioni (**trial**), in cui tecnologia e utente alternano le prese di turno (**step)**.

La memoria di un agente conversazionale si impone quindi come il prodotto di tre diversi fattori: la memoria inerente al trial in svolgimento, quella relativa a trial avvenuti precedentemente ed infine quella legata ad una conoscenza esterna, risultante dal training a cui il modello è stato sottoposto.

Ne consegue che maggiore è la quantità di informazioni a cui il modello riesce ad accedere, superiore sarà la qualità dello scambio. Tuttavia, ad oggi le capacità mnemoniche di ChatGPT risultano alquanto limitate.

Ciò è da imputare principalmente alla natura generativa del sistema e alle dinamiche relative al suo addestramento; quest'ultimo, fondato sull'analisi di quantità massicce di dati, non prevede da parte dell'umano né un controllo capillare della qualità dei testi somministrati né della fattualità degli output emessi dalla macchina.

ChatGPT risulta quindi incapace di individuare eventuali errori, ed anche quando sollecitato dall'utente, provvede a correzioni superficiali che non hanno alcun impatto sulle conoscenze profonde del sistema o sulle sue prestazioni future.

Una soluzione che sembra risolvere apparentemente il problema della memoria a breve termine (adottata anche in alcune fasi di questa indagine) consiste nello scegliere

di condurre l'interazione rimanendo sempre all'interno di un unico trial, utilizzando la medesima conversazione per lo svolgimento di più task.

Nonostante questo approccio agevoli lo scambio soprattutto all'inizio, man mano che il numero di interazioni aumenta, la macchina fatica a mantenere il focus e richiede un'iterazione del prompt, rivelando la fallacia dell'escamotage.

Inoltre, adottare una procedura del genere rischia di generare bias di altro tipo, non potendo escludere conseguenze derivanti dalla familiarizzazione della macchina con il tipo di task che viene reiterato. Nel caso specifico della nostra indagine, si è infatti notato un leggero aumento nella precisione dell'attribuzione dei punteggi proprio a seguito di questa scelta, producendo un vantaggio che tuttavia altera la rappresentazione oggettiva del sistema e che pertanto è stato abbandonato dopo i primi tentativi.

Seppur le prestazioni del modello su un contesto più ampio sembrano migliorare nel passaggio dalla versione 3.5 alla 4.0, o addirittura all'appena rilasciata 4o, l'implementazione della memoria del sistema si impone in quanto sempre più necessaria.

Ad oggi l'unica soluzione al problema è legata alla proposta di OpenAI di aggiungere all'interfaccia di ChatGPT un'impostazione denominata "Memory", che affiancherà la già esistente "Custom Instructions"[36] e che dovrebbe essere rilasciata nel breve periodo. Tale aggiunta permetterà agli utenti di personalizzare il sistema fornendo in prima persona alla macchina elementi che essi reputano imprescindibili per l'interazione, consentendole di ricordare dettagli utili allo svolgimento dei task.

Una soluzione più pervasiva e definitiva della problematica potrebbe invece derivare dalla maggiore estensione di meccanismi di Reinforcement Learning (§2.2 e §2.6.1.1.) durante le fasi di training del modello, garantendo così una diminuzione delle allucinazioni.

5.4.1 Il fenomeno delle allucinazioni

Con "allucinazioni" facciamo riferimento ad output emessi dai LLM, che pur sembrando corretti all'apparenza, si dimostrano errati ad un'analisi più attenta, dando voce a presupposti illogici o infondati.

[36] Tale impostazione consente all'utente di personalizzare la propria interazione col chatbot inserendo alcune informazioni personali (es. i propri hobby) e/o preferenze in merito allo stile di risposta che si preferisce ricevere (ad esempio regolandone registro e tono). Si ha a disposizione un massimo di 1500 caratteri per definire ogni campo.

Lo studio condotto da Zheng et al. nel 2023 si propone di indagare le principali allucinazioni prodotte da ChatGPT nell'ambito del *Question Answering*. Esse vengono suddivise per tipologia, individuandone le cause ed eventualmente proponendone meccanismi di risoluzione a partire dall'analisi di *benchmarks* effettuate da annotatori del settore.

Dall'indagine emergono quattro categorie primarie di errori: *comprehension error*, *factuality error*, *specificity error* ed i*nference error*.

La prima fa riferimento alla difficoltà riscontrata dal software nel comprendere il significato di una richiesta fatta dall'utente, soprattutto in casi di ambiguità lessicali o grammaticali presenti in essa.

La seconda ha invece a che vedere con la mancanza di conoscenza adeguata del modello in merito ad un determinato argomento; ciò gli impedisce di produrre una risposta accurata e spesso porta alla condivisione di un'informazione sbagliata.

La terza è relativa alla non specificità dell'output proposto dal sistema, che si traduce nella generazione di risposte troppo vaghe.

L'ultima categoria è invece da ricondurre a problematiche che emergono nel ragionamento condotto dal modello, il quale, pur possedendo tutti i presupposti necessari, fallisce nell'arrivare alla conclusione corretta.

La ricerca dimostra che la tipologia di errore in cui la macchina incappa più spesso è rappresentata dalla fattualità, ed ha a che vedere con la difficoltà del modello nel recuperare quella che viene definita "conoscenza indispensabile", vale a dire l'informazione che si rende essenziale per soddisfare un determinato quesito.

Difatti, le due abilità necessarie al sistema affinchè possa richiamare alla memoria certe informazioni sono la *Knowledge Memorization* e la *Knowledge Recall*.

La prima è il risultato di un processo per il quale «There exist an appropriate prompt s which, when fed into the model, will result in the essential knowledge p.», mentre la seconda è così descritta: «Given the question q as the prompt, the model is able to output the memorized essential knowledge p.» (Zheng, Huang, & Chang, 2023: 5).

Per implementare entrambe le abilità, gli autori del paper suggeriscono di agire sulla granularità dei dati forniti in input; informazioni più esplicite e dettagliate verranno memorizzate ed eventualmente recuperate dal modello con più facilità.

5.4.2 Etica e tossicità

Ulteriori criticità legate all'utilizzo di ChatGPT (e più in generale dei LLM) rientrano nella sfera dell'etica.

I fruitori del modello risultano particolarmente preoccupati da questioni relative alla gestione della privacy, al diritto d'autore e all'incremento della disinformazione.

Presupposto comune di quanto detto è la capacità del modello di generare contenuti di alta qualità, spesso rendendo complessa (se non impossibile) la disambiguazione con un prodotto dell'intelligenza umana.

Ad oggi, infatti, grazie ai recenti sviluppi dell'Intelligenza Artificiale, modificare testi, immagini e video manipolandone il contenuto non è mai stato così semplice. Ciò alimenta il rischio di creazione e diffusione di fake news, incentivando la disinformazione, la sfiducia nei media e potenzialmente minando la reputazione di singoli individui e di intere organizzazioni.

Ad esempio, si può facilmente dare l'impressione che una figura di spicco del governo abbia tenuto un dibattito divulgando idee molto lontane dalle sue tipiche posizioni, agendo in maniera impercettibile sul labiale e sul doppiaggio del filmato. Inoltre, grazie alla creazione di avatar interamente digitali (riconducibili al "deepfake"), aumenta esponenzialmente il rischio di imbattersi in notizie infondate, divulgate da figure all'apparenza "umane" e pertanto perfettamente credibili.

L'assiduo coinvolgimento dell'IA nella creazione di contenuti genera controversie anche in merito all'attribuzione del diritto d'autore e della proprietà intellettuale, ad esempio nei casi in cui essa viene utilizzata per produrre immagini o contenuti musicali successivamente immessi nel mercato.

Ancora, il fatto che questi modelli siano allenati su grandi quantità di dati li rende esposti ad input potenzialmente nocivi, contenenti idee xenofobe e discriminatorie.

Parliamo quindi di *tossicità* in relazione all'abilità del modello di generare contenuti dannosi ed offensivi.

Seppur nel passaggio da una versione all'altra del software sia stato operato un grande miglioramento nella correzione della maggior parte dei bias di cui inizialmente esso era vittima (osservabili nella system card resa disponibile dalla stessa OpenAI sul proprio sito), lo studio di Zhuo et al. (2023) dimostra che è ancora possibile evadere certe

restrizioni utilizzando tecniche di *jailbreaking*[37], quali il convincere la macchina a calarsi nei panni di una particolare figura, conseguentemente adottandone il linguaggio e gli ideali. Scenari del genere alimentano la preoccupazione dei genitori in merito ad un'eventuale esposizione dei minori a contenuti pericolosi e potenzialmente fuorvianti.

Infine, una delle tematiche più sentite dai fruitori di piattaforme quali ChatGPT, riguarda la gestione delle informazioni reperite e memorizzate dalla macchina attraverso l'interazione con l'utente e la conseguente profilazione a cui quest'ultimo è inconsciamente sottoposto.

Proprio a causa delle perplessità relative all'accesso del modello ai dati sensibili dell'utente, in assenza di una normativa che ne regolasse e tutelasse l'interazione, il 31 marzo 2023 l'Italia è stata il primo paese a bloccare l'utilizzo di ChatGPT su tutto il territorio.

Il blocco è stato poi revocato il 1° maggio 2023, in seguito ad un'implementazione da parte di OpenAI di un sistema di verifica dell'età e di un'esplicita informativa relativa al trattamento dei dati sensibili, in linea con la normativa vigente in Europa.

5.4.2.1 Raccomandazione UNESCO sull'etica dell'IA

Nel novembre del 2021 l'UNESCO rilascia la "Recommendation on the Ethics of Artificial Intelligence", documento volto a promuovere un utilizzo regolamentato e consapevole dell'IA, tanto nell'ambito pubblico quanto nel privato.

Nello specifico, il regolamento evidenzia l'impatto positivo che l'Intelligenza Artificiale potrebbe avere nella risoluzione di problematiche sociali (quali il *gender gap*), ambientali (*climate change*) e persino economiche, rendendone esplicite le potenzialità ma al contempo confermando la necessità di indirizzarne l'utilizzo in senso etico, puntando al miglioramento delle condizioni di vita dell'intera umanità.

In tempi più recenti, e dunque a seguito di un ulteriore sviluppo dell'IA, l'assistente generale dell'educazione per l'UNESCO, Stefania Giannini, ha reso noto un manifesto relativo ad utilizzo e rischi della *GAI* in ambito educativo.

Il documento sottolinea la necessità di ripensare il ruolo degli insegnanti e della didattica tradizionale, adeguando contenuti e strumenti utilizzati nell'istruzione ai

[37] Macrocategoria riferita agli espedienti utilizzati per evadere le restrizioni di un sistema digitale, violandone i termini ed accedendo a funzionalità avanzate.

cambiamenti che avvengono su scala globale anche grazie alla diffusione di questo tipo di tecnologia, il cui avvento ha segnato l'inizio di una nuova era nella rivoluzione digitale.

Si insiste inoltre sulla preoccupazione relativa all'integrazione incondizionata in aula di uno strumento potenzialmente pericoloso per i ragazzi, evidenziando la necessità di un'analisi capillare della suddetta risorsa, volta a limitarne eventuali rischi:

> UNESCO is working with countries to help them develop strategies, plans, and regulations to assure the safe and beneficial use of AI in education. In May 2023, UNESCo organized the first global meeting of Ministers of Education to share knowledge about the impact of generative AI tools on teaching and learning. This meeting has helped UNESCO chart a roadmap to steer the global policy dialogue with governments, as well as academia, civil society and private sector partners. We are not starting from scratch (Giannini, 2023: 5)

Al contempo, si sottolinea quanto sia importante continuare ad investire sulla conoscenza e sulle persone che la promulgano, prima ancora che sulla tecnologia, dando così voce alle paure dei docenti relative ad una loro possibile sostituzione in favore di questa risorsa.

Se è vero che l'IA è in grado di detenere informazioni tecniche e multidisciplinari, è altresì vero che la didattica non può ridursi ad uno sterile trasferimento di nozioni, rendendo più che mai necessaria la collaborazione tra umano e macchina per perseguire il raggiungimento di un'educazione inclusiva, stimolante ed altamente personalizzata, che passa attraverso un utilizzo consapevole delle risorse a nostra disposizione:

> This is perhaps the 'raison d'être' of education: to help us make informed choices of how we want to construct our lives and our societies. The central task for education at this inflection moment is less to incorporate new and largely untested AI applications to advance against the usual targets for formal learning. Rather, it is to help people develop a clearer understanding of when, by whom, and for what reasons this new technology should and should not be used. AI is also giving us impetus to re-examine what we do in education, how we do it, and, most fundamentally, why (Giannini, 2023: 8)

In conclusione, la sinergia tra Intelligenza Artificiale e competenza umana si afferma in quanto elemento imprescindibile per il potenziamento degli ambienti di apprendimento e degli orizzonti educativi.

Conclusioni

Questa tesi ha esplorato le potenzialità dell'intelligenza artificiale (IA) in ambito didattico, focalizzando l'attenzione sui processi di valutazione attraverso la proposta di un caso di studio sperimentale per l'italiano che ha coinvolto studenti e docenti della scuola media. La ricerca si è concentrata in particolare sull'utilizzo del modello generativo del linguaggio ChatGPT per la valutazione automatica di testi argomentativi prodotti dagli studenti, indagandone l'efficacia attraverso un dialogo costante con forme e modalità della didattica tradizionale.

Come discusso nei vari capitoli dell'elaborato, il panorama educativo odierno è caratterizzato da sfide complesse che richiedono soluzioni innovative. L'IA si propone quindi come un potenziale alleato per migliorare l'efficacia, l'equità e l'inclusività del sistema scolastico, offrendo nuovi strumenti per supportare il lavoro dei docenti. In questo contesto, la valutazione automatica di testi scritti rappresenta uno dei campi di applicazione più promettenti dell'IA in ambito didattico.

Lo studio presentato in questa tesi ha cercato di dare un contributo alle ricerche in questa direzione. Un primo risultato significativo è stata la creazione di un corpus a partire dalla raccolta di elaborati prodotti dagli studenti. Questo corpus, composto da 34 testi argomentativi, ha rappresentato la base per la valutazione automatica condotta dal modello generativo del linguaggio ChatGPT, successivamente confrontata con quella della docente.

L'analisi del corpus ha permesso di comprendere meglio le caratteristiche e le difficoltà degli studenti nella produzione di testi argomentativi. Inoltre, ha fornito dati preziosi per valutare le prestazioni di ChatGPT e per identificare i suoi punti di forza e di debolezza.

Sebbene preliminari, i risultati della ricerca dimostrano che ChatGPT può essere un valido strumento per la valutazione automatica di testi argomentativi in italiano. Il modello ha infatti mostrato un'accuratezza generale discreta, fornendo feedback coerenti nei vari test. Si è notato anche che le prestazioni di ChatGPT migliorano significativamente se guidate da parametri espliciti, i quali, quando scelti in autonomia dal modello, ricalcano spontaneamente quelli selezionati dall'insegnante, rivelando la predisposizione della piattaforma allo svolgimento di compiti di questo tipo.

La valutazione automatica avviene quasi istantaneamente e in forma personalizzata, suggerendo il potenziale legato all'adozione di sistemi automatici di valutazione sia per i docenti, riducendo i tempi di esecuzione del task e garantendo maggiore imparzialità, sia per gli studenti, fornendo loro un supporto continuo.

Tuttavia, il software presenta difficoltà nell'attribuire il giusto peso alle diverse tipologie di errori, spesso trascurando quelli meno evidenti a favore di un giudizio positivo, che intende incoraggiare il miglioramento dell'alunno.

Queste tendenze sono in parte attribuibili alla mancanza di familiarità del modello con il tipo di task e di dati trattati, vista l'assenza di un addestramento specifico precedente l'indagine. Ciononostante, questo tipo di ricerca è stata motivata dalla volontà di testare le conoscenze "implicite" dello strumento, misurandone le prestazioni accessibili da qualsiasi utente non esperto e pertanto rappresentative dell'utilizzo che ne possono fare i docenti in aula.

L'estensione limitata del corpus non consente di trarre conclusioni definitive sulle capacità correttive del software, le cui prestazioni potrebbero migliorare ulteriormente con l'applicazione a un campione più ampio o a tipologie differenti di testo. È altresì complicato stimare l'impatto che una maggiore conoscenza della storia linguistica e scolastica degli alunni avrebbe sull'accuratezza delle valutazioni emesse dalla macchina.

Inoltre, sebbene l'introduzione al modello e la partecipazione all'indagine abbiano consentito alla docente di esplorare nuove metodologie didattiche e sviluppare ulteriori competenze, sarebbe interessante indagare anche gli eventuali impatti che il confronto con i giudizi del Chatbot potrebbe avere sull'insegnante, portandola magari a rivedere le sue valutazioni.

Pertanto, la ricerca ha permesso di avviare diverse riflessioni e di individuare possibili direzioni per ricerche future, evidenziando la necessità di focalizzarsi in particolar modo su: ampliamento del corpus, addestramento del modello su dati specifici dell'universo scolastico e valutazione dell'incidenza dei giudizi del chatbot sulle prestazioni di studenti e docenti.

La raccolta di un corpus più ampio di testi rafforzerebbe la validità dei risultati. Un campione limitato potrebbe infatti non rappresentare adeguatamente le caratteristiche stilistiche degli scriventi e le capacità correttive del modello. Viceversa, un corpus esteso

contenente testi appartenenti a tipologie diverse, consentirebbe di testare la tecnologia in diverse condizioni, garantendo una valutazione più accurata delle sue capacità correttive.

Inoltre, l'addestramento del modello attraverso la somministrazione di dati specifici del contesto scolastico quali le informazioni sull'andamento dell'apprendimento degli studenti e sugli obiettivi didattici stabiliti dalle indicazioni ministeriali, consentirebbe al chatbot di fornire valutazioni che tengano conto del progresso individuale degli alunni e degli standard educativi, avvicinandosi così alla completezza della visione d'insieme di un docente esperto.

Ancora, sarebbe interessante valutare in che modo si manifesti l'eventuale incidenza dei giudizi emessi dal modello sulle successive prestazioni di studenti ed insegnanti. Ad esempio, si potrebbe osservare se i feedback automatizzati migliorino o meno la comprensione da parte degli alunni degli errori commessi o se aiutino il docente a considerare elementi prima sfuggiti alla sua attenzione.

In conclusione, l'indagine conferma la validità dei modelli generativi del linguaggio in quanto strumenti di supporto all'insegnamento, mettendone in luce limiti e potenzialità. Un utilizzo consapevole dell'Intelligenza Artificiale può rendere il processo educativo più interattivo e personalizzato, offrendo nuovi spunti di riflessione ai docenti e adattando i contenuti didattici ai bisogni individuali degli studenti.

Tuttavia, la progressiva integrazione di questi modelli nell'istruzione e il loro massimo rendimento richiedono la guida di docenti esperti, capaci di orchestrare le funzionalità offerte dall'IA e di colmare eventuali lacune, dimostrando ancora una volta l'imprescindibilità di queste figure nel processo di apprendimento.

Appendice

Appendice A: Questionario conoscitivo sulle abitudini linguistiche

Questionario conoscitivo sulle abitudini linguistiche

Nome:

Cognome:

In che anno sei nato/a?
☐ 2009 (o prima) ☐ 2010 ☐ 2011 ☐ 2012 (o dopo)

Dove siete nati tu e i tuoi genitori?	tu	madre	padre
• Italia (o Repubblica di San Marino)	☐	☐	☐
• Unione Europea (Austria, Belgio, Bulgaria, Cipro, Croazia, Danimarca, Estonia, Finlandia, Francia, Germania, Grecia, Irlanda, Lettonia, Lituania, Lussemburgo, Malta, Paesi Bassi, Polonia, Portogallo, Regno Unito, Repubblica Ceca, Romania, Slovacchia, Slovenia, Spagna, Svezia, Ungheria)	☐	☐	☐
• Paese europeo non dell'Unione Europea	☐	☐	☐
• Altro	☐	☐	☐

Se tu non sei nato/a in Italia, quanti anni avevi quando sei arrivato/a in Italia? Se sei nato/a in Italia, salta questa domanda e vai direttamente alla domanda successiva. Per rispondere considera gli anni compiuti (ad esempio se avevi tre anni e mezzo, indica la risposta A. fino a tre anni). Barra una sola casella.
☐ Fino a tre anni ☐ Da 4 a 6 anni ☐ Da 7 a 9 anni ☐ Da 10 a 12 anni ☐ Da 13 a 15 anni

A casa, quale lingua parli la maggior parte del tempo?

Metti una crocetta su un solo quadratino.

- ☐ L'italiano
- ☐ Un dialetto (veneto, siciliano, sardo ecc.)
- ☐ Un'altra lingua (inglese, francese, tedesco, rumeno, arabo, cinese, hindi ecc.)

Con gli amici e nel tempo libero, quale lingua parli la maggior parte del tempo?

- ☐ L'italiano
- ☐ Un dialetto (veneto, siciliano, sardo ecc.)
- ☐ Un'altra lingua (francese, tedesco, rumeno, arabo, cinese, hindi ecc.)

Quando hai bisogno di aiuto nei compiti, chi ti aiuta maggiormente?

- ☐ Mi aiuta qualcuno in famiglia (mamma, papà, fratelli, sorelle, nonni...)
- ☐ Mi aiuta qualcun altro (es: doposcuola pomeridiano)

Sei andato alla scuola dell'infanzia (scuola materna)?

Barra una sola casella.

- ☐ No
- ☐ Sì, per un anno o meno di un anno
- ☐ Sì, per più di un anno

Abitualmente con chi vivi?

Metti una crocetta su un solo quadratino

- ☐ Con tutti e due i miei genitori
- ☐ Con uno solo dei miei genitori
- ☐ Un po' da mia madre, un po' da mio padre
- ☐ Non vivo con i miei genitori

Appendice B: Griglie di valutazione

Griglia di valutazione elaborata dalla docente:

Rubrica di valutazione autentica del testo argomentativo					
indicatori/livelli	**avanzato**	**intermedio**	**di base**	**in via di acquisizione**	**voto/10**
focus dell'esposizione	**2**		**0**		
	L'esposizione individua tutti i concetti fondamentali da affrontare e delinea i punti principali da trattare	L'esposizione individua i concetti fondamentali da affrontare	L'esposizione delinea alcuni punti principali da affrontare, ma non individua chiaramente lo scopo	L'esposizione non individua i punti principali da affrontare né lo scopo.	
elementi a sostegno dell'esposizione	**2**	**1,5**	**1**	**0,5**	
	Include 3 o più elementi di prova (fatti, dati, esempi, esperienze di vita reale) che sostengono l'argomentazione.	Include 3 o più elementi di prova (fatti, dati, esempi, esperienze di vita reale) che sostengono l'argomentazione.	Include 2 elementi di prova (fatti, dati, esempi, esperienze di vita reale) che sostengono l'argomentazione.	Include 1 o meno elementi di prova (fatti, dati, esempi, esperienze di vita reale).	
accuratezza ed approfondimenti	**2**	**1,5**	**1**	**0,5**	
	Tutti i fatti e i dati a supporto dell'argomentazione sono riportati in modo accurato ed accuratamente	Quasi tutti i fatti e di dati a supporto dell'argomentazione sono riportati in modo accurato e correttamente	La maggior parte dei fatti e dei dati a supporto della argomentazione sono riportati in modo accurato	La maggior parte dei fatti e dei dati a supporto della argomentazione sono riportati in modo improprio e non	

Adattamento a cura di Giovanna Lazzari da allegato a articolo "Un'esperienza concreta di didattica cooperativa a distanza" di Patrizio Vignola, pubblicato sul sito www.scintille.it nell'aprile 2021.

	approfonditi	approfonditi.	fondamentalmente approfondito	approfondito	
correttezza ortografica e sintattica	**2**	**1,5**	**1**	**0,5**	
	Non sono presenti errori di sintassi o di ortografia che distraggono il lettore dal contenuto		Sono presenti errori di sintassi o di ortografia che distraggono il lettore dal contenuto.		
uso di un linguaggio specifico	**2**	**1,5**	**1**	**0,5**	
	Il linguaggio utilizzato si avvale sempre di espressioni e termini specifici dell'universo scolastico	Il linguaggio utilizzato nella maggior parte dei casi si avvale di espressioni e termini specifici dell'universo scolastico	Il linguaggio utilizzato si avvale perlopiù di espressione e termini riconducibili all'universo scolastico	Il linguaggio utilizzato si avvale di rado di espressione e termini specifici dell'universo scolastico	

Griglia di valutazione elaborata da ChatGPT:

Indicatori	avanzato	intermedio	di base	in via di acquisizione	voto/10
	2	1,5	1	0,5	
Chiarezza e coerenza	Il testo è molto chiaro e completamente coerente. Le idee sono organizzate in modo logico e fluido.	Il testo è chiaro e abbastanza coerente. Le idee sono ben organizzate ma possono esserci alcune lievi incongruenze.	Il testo è parzialmente chiaro ma presenta diverse incoerenze e salti logici. Le idee sono comprensibili ma non ben strutturate.	Il testo è difficile da seguire e manca di coerenza. Le idee sono disorganizzate e confuse.	
Argomentazione	Le argomentazioni sono molto ben sviluppate e supportate da prove solide e pertinenti. Il ragionamento è logico e convincente.	Le argomentazioni sono ben sviluppate e generalmente supportate da prove pertinenti, anche se non sempre in modo completo.	Le argomentazioni sono presenti ma non sono ben sviluppate o supportate adeguatamente.	Mancano argomentazioni solide. Le affermazioni sono deboli e non supportate da prove.	
Originalità e profondità	Il testo è altamente originale e dimostra una profondità eccezionale. Le idee sono sviluppate in modo approfondito e innovativo.	Il testo è originale e dimostra una buona profondità. Le idee sono sviluppate in modo appropriato.	Il testo presenta alcune idee originali, ma la profondità è limitata. Le idee sono sviluppate solo superficialmente.	Il testo manca di originalità e profondità. Le idee sono superficiali e ripetitive.	
Stile ed espressione	Lo stile e l'espressione sono eccellenti. Il linguaggio è ricco, preciso e perfettamente adeguato al contesto.	Lo stile e l'espressione sono appropriati. Il linguaggio è chiaro e adeguato, con poche imprecisioni.	Lo stile e l'espressione sono semplici e possono presentare errori. Il linguaggio è comprensibile ma non sempre adeguato.	Lo stile e l'espressione sono inadeguati. Il linguaggio è povero e non adeguato al contesto.	
Impatto emotivo e persuasione	Il testo ha un forte impatto emotivo e persuade il lettore in modo efficace. Utilizza elementi di coinvolgimento altamente persuasivi.	Il testo ha un buon impatto emotivo e persuade il lettore. Include elementi di coinvolgimento efficaci.	Il testo ha un impatto emotivo limitato e persuade solo in parte. Mancano elementi forti di coinvolgimento.	Il testo non riesce a suscitare emozioni o a persuadere il lettore.	

Appendice C: Giudizi dell'insegnante relativi ai gruppi A e B

Gruppo A

Rubrica di valutazione autentica del testo argomentativo classe 3C _2023/'24							
Studente	Titolo	Focus esposizione	Elementi a sostegno dell'esposizione	Accuratezza e approfondimenti	Correttezza ortografica e sintattica	Uso di un linguaggio specifico	valutazione
1	Il femminismo non è solo "una cosa da femmine"	intermedio 2	di base 1	avanzato 2	intermedio 1,5	avanzato 2	8,5
2	A ogni sport la sua occasione	di base 1	di base 1	intermedio 1,5	intermedio 1,5	intermedio 1,5	6,5
3	Perché chi è nato in Italia [...] non può avere la cittadinanza italiana?	di base 0	di base 1	in via di acquisizione 0,5	di base 1	di base 1	3,5
4	Stereotipi e disparità – un problema sottovalutato	avanzato 2	intermedio 1,5	avanzato 2	intermedio 1,5	avanzato 2	9
5	I social media	di base 0	di base 1	in via di acquisizione 0,5	di base 1	di base 1	3,5
6	Il professore ideale	intermedio 2	intermedio 1,5	intermedio 1,5	avanzato 2	avanzato 2	9
7	Guerra	in via di acquisizione 0	in via di acquisizione 0,5	in via di acquisizione 0,5	in via di acquisizione 0,5	di base 1	2,5
8	?	di base 0	di base 1	in via di acquisizione 0,5	intermedio 1,5	di base 1	4
9	Dis/informazione online, è davvero una fonte sicura per i giovani?	avanzato 2	intermedio 1,5	avanzato 2	intermedio 2	intermedio 1,5	9
10	Perché le stars NBA non giocano in nazionale?	di base 0	di base 1	in via di acquisizione 0,5	intermedio 1,5	intermedio 1,5	4,5
11	Smetti di bullizzare gli artisti K-POP	intermedio 2	in via di acquisizione 0,5	di base 1	di base 1	di base 1	5,5
12	Pallacanestro o calcio; quale è più	intermedio 2	di base 1	di base 1	di base 1	di base 1	6

	intenso?						
13	Ansia e stress scolastico: un problema che riguarda la maggior parte degli studenti italiani	intermedio 2	intermedio 1,5	intermedio 1,5	di base 1	intermedio 1,5	7,5
14	Uguaglianza di genere: a che punto siamo?	di base 1	in via di acquisizione 0,5	in via di acquisizione 0,5	di base 1	di base 1	4
15	Gli aspetti positivi e negativi del divorzio	di base 1	intermedio 1,5	di base 1	di base 1	di base 1	5,5
16	Hate speech – che cos'è il discorso d'odio online?	intermedio 2	avanzato 2	avanzato 2	intermedio 1,5	avanzato 2	9,5
17	La valutazione è necessaria?	di base 1	intermedio 1,5	intermedio 1,5	intermedio 1,5	intermedio 1,5	7

Gruppo B

Rubrica di valutazione autentica del testo argomentativo classe 3C _2023/'24							
Studente	Titolo	Focus esposizione	Elementi a sostegno dell'esposizione	Accuratezza e approfondimenti	Correttezza ortografica e sintattica	Uso di un linguaggio specifico	punteggio
1	Senza titolo	avanzato 2	di base 1	in via di acquisizione 0,5	in via di acquisizione 0,5	di base 1	5
2	Influencer idoli o manipolatori?	avanzato 2	avanzato 2	intermedio 1,5	avanzato 2	avanzato 2	9,5
3	L'e-book: la scelta più comoda per i lettori di oggi	avanzato 2	avanzato 2	avanzato 2	avanzato 2	di base 1	9
4	Influencer: idoli o manipolatori?	avanzato 2	avanzato 2	avanzato 2	avanzato 2	di base 1	9

5	Influencer: sono esempi negativi?	avanzato 2	avanzato 2	avanzato 2	di base 1	di base 1	8
6	Influencer: manipolatori o promotori di un buon stile di vita?	avanzato 2	intermedio 1,5	avanzato 2	avanzato 2	avanzato 2	9,5
7	Il libro cartaceo è meglio dell'e-book?	avanzato 2	avanzato 2	di base 1	avanzato 2	avanzato 2	9
8	Influencer: idoli o manipolatori?	di base 0	di base 1	di base 1	di base 1	avanzato 2	5
9	Vita da influencer *non è facile essere un influencer* È un testo espositivo e va fuori tema	in via di acquisizione 0,5	in via di acquisizione 0,5	in via di acquisizione 0,5	intermedio 1,5	intermedio 1,5	4,5
10	Influencer: modelli della rete o manipolatori? È un testo espositivo!	avanzato 2	in via di acquisizione 0,5	di base 1	avanzato 2	avanzato 2	7,5
11	E-book o cartaceo?	avanzato 2	di base 1	avanzato 2	avanzato 2	avanzato 2	9

12	Gli e-reader, sì o no?	di base 1	in via di acquisizione 0,5	di base 1	di base 1	avanzato 2	5,5
13	Dal virtuale al reale	di base 1	in via di acquisizione 0,5	di base 1	di base 1	avanzato 2	5,5
14	Libro cartaceo o e-book?	avanzato 2	avanzato 2	intermedio 1,5	intermedio 1,5	avanzato 2	9
15	La maschera degli influencer	avanzate 2	avanzate 2	avanzate 2	avanzate 2	di base 1	9
16	I social	in via di acquisizione 0,5	in via di acquisizione 0,5	in via di acquisizione 0,5	di base 1	di base 1	3,5
17	Senza titolo testo copiato!	di base 0	in via di acquisizione 0,5	in via di acquisizione 0,5	di base 1	in via di acquisizione 0,5	2,5

Appendice D: Giudizi ChatGPT Prompt 1

Gruppo A

Testo n°:	Valutazione:
1	8.5
2	7.5
3	7
4	8.5
5	8
6	7.5
7	6.5
8	6
9	8
10	7.5
11	7
12	6.5
13	8.5
14	7
15	7.5
16	8.5
17	8.5

Gruppo B

1	7
2	7.5
3	8
4	8.5
5	7
6	8
7	8
8	7.5
9	8
10	7
11	7.5
12	6.5
13	7.5
14	7.5
15	8
16	7.5
17	8

Appendice E: Giudizi ChatGPT Prompt 2

Gruppo A

Testo n°:	Chiarezza e coerenza:	Argomentazione:	Originalità e profondità:	Stile ed espressione:	Impatto emotivo e persuasione:	Valutazione:
1	2	2	2	2	2	10
2	2	2	1.5	2	1.5	9
3	1.5	1.5	1.5	1.5	1.5	7
4	2	1.5	1.5	2	1.5	8.5
5	2	2	1.5	2	1.5	9
6	1.5	1.5	1.5	1	1.5	7
7	1.5	1	1	1	1	5.5
8	1.5	1	1	1.5	1	6
9	2	2	1.5	1.5	1.5	8.5
10	2	1.5	1.5	1.5	1	7.5
11	1.5	1.5	1.5	1	1.5	7
12	1.5	1.5	1	1.5	1	6.5
13	2	2	1.5	2	2	9.5
14	1.5	1.5	1	1.5	1.5	7
15	1.5	1.5	1.5	1.5	1.5	7.5
16	2	2	2	2	2	10
17	2	2	2	2	2	10

Gruppo B

Testo n°:	Chiarezza e coerenza:	Argomentazione:	Originalità e profondità:	Stile ed espressione:	Impatto emotivo e persuasione:	Valutazione:
1	1.5	1	1	1.5	1	6.5
2	1.5	1.5	2	1.5	1.5	8
3	2	1.5	1.5	2	2	9
4	2	1.5	1.5	2	1.5	8.5
5	1.5	1.5	1	2	1.5	7.5
6	2	1.5	1.5	2	2	9
7	2	1.5	1	2	1.5	8
8	1.5	1	1	1.5	1	6
9	2	1.5	1.5	2	1.5	8.5
10	1.5	1.5	1.5	2	1.5	8
11	2	1.5	1.5	2	1.5	8.5
12	1	1	0.5	1.5	1	5
13	2	1.5	1	1.5	1.5	7.5
14	1.5	1.5	1.5	2	1	7.5
15	1.5	1.5	1.5	2	1.5	8
16	1.5	1.5	1.5	2	1.5	8
17	1.5	1.5	1	2	1.5	7.5

Appendice F: Giudizi ChatGPT Prompt 3

Gruppo A

Testo n°:	Focus:	Elementi a sostegno dell'esposizione:	Accuratezza e approfondimenti:	Correttezza ortografica e sintattica:	Uso di un linguaggio specifico:	Valutazione:
1	2	2	2	2	2	10
2	1.5	2	1.5	1.5	1.5	8
3	1.5	1.5	1.5	1	1.5	7
4	2	2	2	1.5	2	9.5
5	2	2	1.5	1.5	1.5	8.5
6	2	2	1.5	1	1.5	8
7	1	1	1	0.5	1	4.5
8	1	0.5	0.5	1	1	4
9	2	2	1.5	1.5	1.5	8.5
10	1.5	1.5	1.5	1	1	6.5
11	1.5	1	1	1	1	5.5
12	1.5	1	1	1	1.5	6
13	2	1.5	1.5	1.5	2	8.5
14	1.5	1.5	1	1.5	1.5	7
15	1.5	1.5	1.5	1.5	1.5	7.5
16	2	2	2	1.5	2	9.5
17	2	1.5	1.5	1.5	2	8.5

Gruppo B

Testo n°:	Focus:	Elementi a sostegno dell'esposizione:	Accuratezza e approfondimenti:	Correttezza ortografica e sintattica:	Uso di un linguaggio specifico:	Valutazione:
1	1	1	1	2	1	6
2	2	2	2	1.5	1.5	9
3	2	2	2	2	2	10
4	2	2	1.5	2	2	9.5
5	2	2	1.5	1.5	1.5	8.5
6	2	2	2	2	2	10
7	2	2	2	2	2	10
8	1.5	1	1	1.5	1	6
9	2	1.5	1.5	2	1.5	8.5
10	1.5	1	1.5	2	1.5	7.5
11	2	2	1.5	1.5	1.5	8.5
12	1	1	1	1	1	5
13	1.5	1.5	1.5	1.5	1.5	7.5
14	2	2	1.5	1.5	2	9
15	2	2	1.5	2	1.5	9
16	2	2	1.5	2	2	9.5
17	1.5	1.5	1.5	2	1.5	8

Appendice G: Profiling dei testi

Gruppo A

Testo n°:	Numero frasi:	Numero token:	Lunghezza media periodo (in token):	Lunghezza media parola (in caratteri):	Indice Gulpease:	Densità lessicale:	Type/Token Ratio (secondo le prime 100 parole):	"Misura" della lunghezza MEDIA delle relazioni di dipendenza:
1	37	670	18.1	4,7	59,5	0,580	0,780	2.131
2	37	670	18.1	4,7	59,5	0,543	0,720	2.131
3	31	967	31,2	5,2	46,4	0,549	0,690	2,518
4	9	403	44,8	4,8	47,0	0,554	0,580	2,875
5	31	885	28,5	5,2	47,4	0,572	0,670	2,333
6	19	459	24,2	5,5	47,1	0,603	0,640	2,584
7	15	363	24,2	4,9	52,3	0,588	0,630	2,217
8	5	139	27,8	5,0	47,8	0,603	0,660	2,284
9	28	735	26,2	5,3	47,8	0,570	0,670	2,251
10	10	339	33,9	5,0	47,5	0,540	0,650	2,951
11	16	254	15,9	5,0	58,1	0,578	0,670	2,080
12	5	195	39,0	4,6	49,1	0,577	0,660	2,491
13	30	715	23,8	4,8	53,7	0,543	0,670	2,287
14	11	209	19,0	4,9	55,2	0,601	0,700	2,273
15	44	1024	23,3	4,6	57,3	0,544	0,670	2,779
16	37	977	26,4	4,9	52,3	0,559	0,720	2,263
17	55	1123	20,4	4,9	55,9	0,559	0,720	2,262

Gruppo B

Testo n°:	Numero frasi:	Numero token:	Lunghezza media periodo (in token):	Lunghezza media parola (in caratteri):	Indice Gulpease:	Densità lessicale:	Type/Token Ratio (secondo le prime 100 parole):	"Misura" della lunghezza MEDIA delle relazioni di dipendenza:
1	8	121	15,1	5,9	49,8	0,651	0,650	2,418
2	17	460	27,1	4,8	52,0	0,573	0,640	2,302
3	7	194	27.7	5,0	48,4	0,559	0,720	2,350
4	10	323	32,3	5,5	43,4	0,620	0,680	2,282
5	10	321	32,1	4,9	48,0	0,566	0,710	2,230
6	43	634	14,7	5,2	59,7	0,573	0,770	2,002
7	36	716	21,1	4,9	54,1	0,571	0,710	2,109
8	33	569	17,2	4,9	58,4	0,569	0,740	2,652
9	37	692	18,7	4,8	57,7	0,584	0,600	2,393
10	26	428	16,5	5,9	49,1	0,632	0,760	2,107
11	14	345	24,6	4,8	53,4	0,550	0,630	2,510
12	35	541	15,5	5,1	58,2	0,555	0,760	2,002
13	31	597	19,3	5,0	54,8	0,566	0,670	2,326
14	11	353	32,1	5,1	46,5	0,547	0,680	2,185
15	24	376	15.7	5,1	57,9	0,595	0,790	2,069
16	32	454	14,2	4,9	62,1	0,580	0,750	2,059
17	19	428	22,5	5,0	52,2	0,583	0,690	2,710

Bibliografia

Andorno, C. (2003). *La linguistica testuale: un'introduzione.* Roma: Carocci.

Bazzanella, C. (1995). I segnali discorsivi. In L. Renzi, G. Salvi, & A. Cardinaletti, *Grande grammatica italiana di consultazione.* Bologna: Il Mulino.

Bengio, Y., & al., e. (2003). A Neural Probabilistic Language Model. *Journal of Machine Learning Research 3*, 1137-1155.

Bloom, B. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, 11.

Bonneton-Bottè, N., Fleury, S., Girard, N., Le Magadou, M., Cherbonnier, A., Renault, M., . . . Jamet, E. (2020). Can tablet apps support the learning of handwriting? An investigation of learning outcomes in kindergarten classroom. *Computers & Education*.

Bosco, C., Montemagni, S., & Simi, M. (2013). Converting italian treebanks: Towards an italian stanford dependency treebank. *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 61-69.

Brophy, J. E., & Good, T. L. (1970). Teachers' communication of differential expectations for children's classroom performance: Some behavioral data. *Journal of Educational Psychology*, 365-374.

Brunato, D., Cimino, A., Dell'Orletta, F., Montemagni, S., & Venturi, G. (2020). Profiling-UD: a Tool for Linguistic Profiling of Texts. *Proceedings of 12th Edition of International Conference on Language Resources and Evaluation (LREC 2020)*, 11-16.

Chen, Y., Jensen, S., Albert, L., Gupta, S., & Lee, T. (2022). Artifcial Intelligence (AI) Student Assistants in the Classroom:. *Information System Frontiers 25*.

Chu, Z., Ni, S., Wang, Z., Feng, X., Li, C., Hu, X., & al., e. (2024). History, Development, and Principles of Large Language Models-An Introductory Survey. . *arXiv*.

Cignetti, L. (2011a). *L'inciso: Natura linguistica e funzioni testuali.* Alessandria: Edizioni dell'Orso.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*.

Conte, M. E. (1978). Deissi testuale a anafora. *Atti del seminario* (p. 37-54). Firenze: Accademia della Crusca.

Conte, M. E. (1989). *La Lingusitica Testuale.* Milano: Feltrinelli.

Conte, M.-E. (1999). *Condizioni di coerenza. Ricerche di linguistica testuale.* Alessandria: Edizioni dell'Orso.

De Beaugrande, R., & Dressler, W. (1981). *Introduction to Text Linguistics.* London: Routledge.

De Cesare, A. (2010). Deittici. In *Enciclopedia dell'italiano* (p. 345-347).

De Marneffe, M.-C., & al., e. (2021). Universal dependencies. *Computational linguistics 47.2*, (p. 255-308).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics (2019)*.

Eager, B., & Brunton, R. (2023). Prompting Higher Education Towards AI-Augmented Teaching and Learning Practice. *Journal of University Teaching & Learning Practice* .

Ellero, P. (1986). I connettivi. *Quaderni del GISCEL*.

Ferrari, A. (2014). *Linguistica del testo. Principi, fenomeni, strutture.* Roma: Carocci.

Foster, S. (2019). What Barriers do Students Perceive to Engagement with Automated Immediate Formative Feedback. *JOURNAL OF INTERACTIVE MEDIA IN EDUCATION*, 1-5.

Giannini, S. (2023). Generative AI and the future of Education.

Grosan, C., & Abraham, A. (2011). *Intelligent Systems. A modern approach.* Springer.

Halliday, M., & Hasan, R. (1976). *Cohesion in English.* London: Routledge.

Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education*, 551.

Hosseini, M., & al., e. (2023). An exploratory survey about using ChatGPT ineducation, healthcare, and research. *PLOS ONE*, 1-14.

Jeon, J., & Lee, S. (2023). Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*.

Kaplan, R., Grotewold, Hartwick, & Papin. (2023). Generative AI and teacher's Perspectives on Its Implementation in Education. *Journal of Interactive Learning Research*, 313-338.

Kaplan-Rakowski, R., Grotewold, K., Hartwick, P., & Papin, P. (2023). Generative AI and Teachers' Perspectives on Its Implementation in Education. *Journal of Interactive Learning Research*, 313-338.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023). Large Language Models are Zero-Shot Reasoners. *Advances in neural information processing systems*.

Kuhail, M., Elsayary, A., Farooq, S., & Ahlam, A. (2023). Exploring Immersive Learning Experiences: A Survey. *Informatics*.

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open*.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of Linguistics*, 3.

Lenci, A., Montemagni, S., & Pirrelli, V. (2016). *Testo e computer. Elementi di linguistica computazionale.* Carocci.

Lo Cascio, V. (1991). *Grammatica dell'argomentare: strategie e strutture.* Firenze: La nuova Italia.

Lu, X., Li, S. L., & Fujimoto, M. (2020). Automatic Speech Recognition. In Y. Kidawara, E. Sumita, & H. Kawai, *Speech-to-Speech Translation.* Singapore: Springer Singapore.

Massa, S., Usai, G., & Riboni, D. (2023). Monitoring Human Attention with a Portable EEG Sensor and Supervised Machine Learning. *EDBT/ICDT Workshops*, 1-4.

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1955). A proposal for the Dartmouth summer research project on artificial intelligence.

Miaschi, A. (2022). *Tracking Linguistic Abilities in Neural Language Models.* Pisa: Dipartimento di Computer Science, Università degli Studi di Pisa.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, gennaio 16). *Efficient Estimation of Word Representations in Vector Space.* Tratto da arXiv: https://arxiv.org/abs/1301.3781

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*.

Naisimith, B., Mulcaire, P., & Burstein, J. (2023). Automated Evaluation of Written Discourse Coherence Using GPT-4. *18th Workshop on Innovative Use of NLP for Building*, (p. 394-403).

OpenAI. (2023). GPT-4 System Card.

Plevris, V., Papazafeiropoulos, G., & Rios, A. J. (2023). Chatbots put to the test in math and logic problems: A preliminary comparison and assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *arXiv*.

Polverini, G., & Gregoric, B. (2024). How understanding large language models can inform the use of ChatGPT in physics education. *European Journal of Physics 45*, 4.

Renzi, L., & Vanelli, L. (1995). La deissi. In L. Renzi, G. Salvi, & A. Cardinaletti, *Grande grammatica italiana di consultazione* (p. 261-375). Bologna: Il Mulino.

Sabatini, F., & Coletti, V. (1999). *DISC; Dizionario italiano Sabatini Coletti.* Firenze: Giunti.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 210-229.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition.* Oxford: Oxford: Basil Blackwell.

Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT – Exploring the ethical issues of an emerging technology. *International Journal of Information Management* .

Steele, J. L. (s.d.). To GPT or not GPT? Empowering our students to learn with AI. *Computers & Education: Artificial Intelligence*.

Straka, M., Hajič, J., & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 4290-4297.

Su, Y. N., Hsu, C. C., Chen, H. C., Huang, K. K., & Huang, Y. M. (2014). Developing a sensor-based learning concentration detection system. *Engineering Computations*.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 433-460.

UNESCO. (2021). Recommendations on the Ethics of Artificial Intelligence.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention Is All You Need. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems.*

Vylomova, E., Rimell, L., Cohn, T., & Baldwin, T. (2015). Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Berlin: Association for Computational Linguistics.

Werlich, E. (1975). *Typologie der Texte : Entwurf eines textlinguistischen Modells zur Grundlegung einer Textgrammatik.* Heidelberg: Quelle & Meyer.

Yen-Ning, S., Chia-Cheng, H., Hsin-Chin, C., & Yueh-Min, H. (2014). Developing a sensor-based learning concentration detection system. *Engineering Computations*, 216-230.

Zhang, Z., Bo, X., Ma, C., Li, R., Chen, X., Dai, Q., . . . Wen, J.-R. (s.d.). A Survey on the Memory Mechanism of Large Language Model based Agents.

Zheng, S., Huang, J., & Chang, K. C.-C. (2023). Why Does ChatGPT Fall Short in Providing Truthful Answers?

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity.