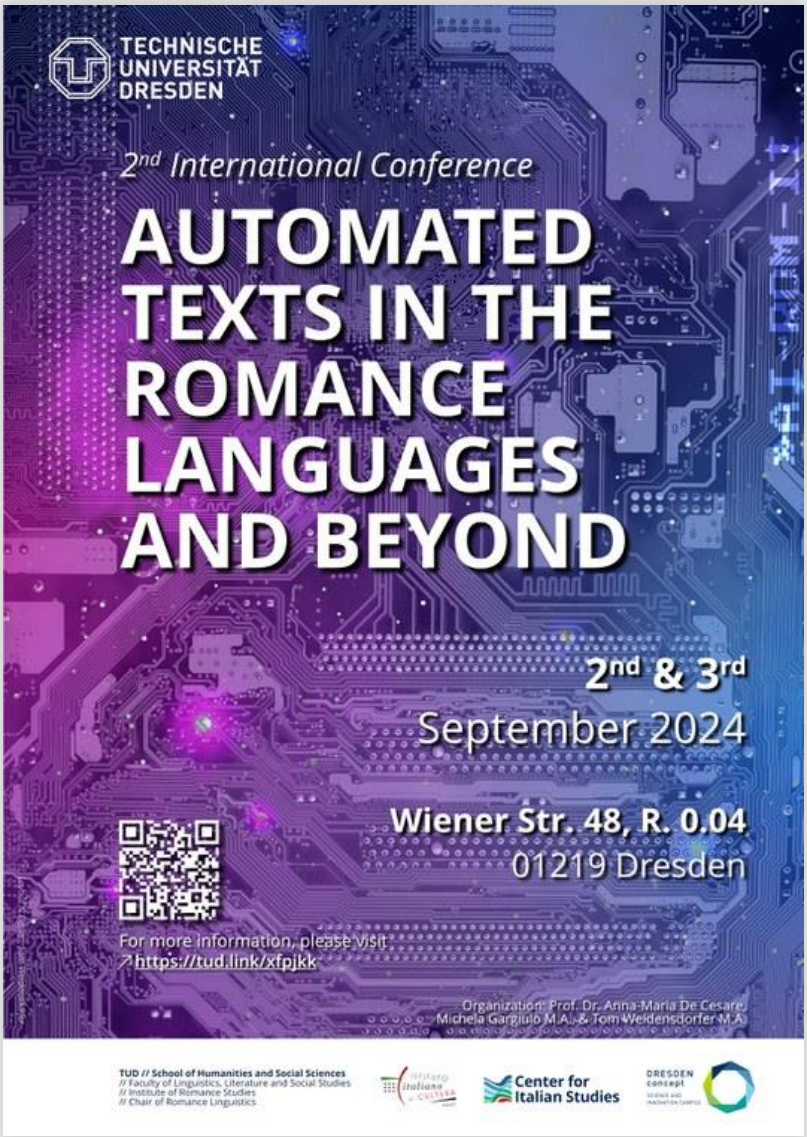
AUTOMATED TEXTS IN THE ROMANCE LANGUAGES AND BEYOND

Book of abstracts

2nd International Conference

Organization: Anna-Maria De Cesare, Michela Gargiulo, Tom Weidensdorfer

Dresden, 2nd & 3rd September 2024



https://tud.link/xfpjkk

Pragmatic abilities from humans to LLMs

Valentina Bambini (Istituto Universitario di Studi Superiori IUSS Pavia)

Pragmatic abilities include a wide range of expressive and receptive skills, from providing the adequate amount of information in conversation to understanding the implicit and non-literal nuances of meanings (Domaneschi & Bambini, 2020). Because of their complexity, pragmatic skills represent a late achievement in development (Tonini et al., 2023) and are vulnerable in numerous pathological conditions, from schizophrenia to neurodegenerative diseases (Bambini, Arcara, Bechi et al., 2016; Bambini, Arcara, Martinelli et al., 2016). How do Large Language Models (LLMs) perform from the pragmatic point of view and how effective and context-sensitive is their communication?

In the talk, I will start from illustrating how pragmatic skills are typically tested in clinical groups, focusing especially on the APACS test (Arcara & Bambini, 2016). Then, I will present the results of a study where ChatGPT (vers. 3.5, March 2023) was tested like a patient in a clinical setting, i.e., administering the APACS items and questions via zero-shot prompting (Barattieri di San Pietro et al., 2023). Overall, ChatGPT performed remarkably well, although it exhibited some weakness in discourse informativity and non-literal understanding compared to normative human data. Qualitatively, these weaknesses do not match those of people with pragmatic disorders, and are indicative of different processing strategies, especially when it comes to metaphor understanding (Carenini et al., 2023). Yet, LLMs can simulate pragmatic disorder, when adequately trained. On this, I will present preliminary evidence of pragmatic behavior of digital twin patients, i.e., a LLM fine-tuned to mimic the linguistic fingerprint of people with schizophrenia, which opens interesting applications for clinical purposes (Barattieri di San Pietro et al., 2024).

References

Arcara G, Bambini V (2016) A test for the Assessment of Pragmatic Abilities and Cognitive Substrates (APACS): Normative data and psychometric properties. *Frontiers in Psychology* 7, 70.

Bambini V, Arcara G, Bechi M, Buonocore M, Cavallaro R, Bosia M (2016) The communicative impairment as a core feature of schizophrenia: Frequency of pragmatic deficit, cognitive substrates, and relation with quality of life. *Comprehensive Psychiatry* 71, 106-120.

Bambini V, Arcara G, Martinelli I, Bernini S, Alvisi E, Moro A, Cappa SF, Ceroni M (2016) Communication and pragmatic breakdowns in amyotrophic lateral sclerosis patients. *Brain and Language* 153-154, 1-12.

Barattieri di San Pietro C, Frau F, Mangiaterra V, Bambini V (2023) The pragmatic profile of ChatGPT: Assessing the communicative skills of a conversational agent. *Sistemi Intelligenti* 35(2), 379-399.

Barattieri di San Pietro C, Scalingi B, Frau F, Agostoni G, Bechi M, Cavallaro R, Bosia M, Bianchini N, Bertini F, Bambini V (2024) Distributional Semantics, NLP, and Machine Learning: a Combined Approach to Language Analysis in Schizophrenia. *2nd DISCOURSE Satellite Meeting; 9 April, 2024; Pavia (I)*.

Carenini G, Bodot L, Bischetti L, Schaeken W, Bambini V (2023), Large Language Models Behave (Almost) As Rational Speech Actors: Insights From Metaphor Understanding, *NeurIPS 2023 workshop: Information-Theoretic Principles in Cognitive Systems*

Domaneschi F, Bambini V (2020) Pragmatic competence. In Fridland E, Pavese C (eds), *The Routledge Handbook of Philosophy of Skill and Expertise*, London, Routledge, 419-430.

Tonini E, Bischetti L, Del Sette P, Lecce S, Bambini V (2023) The relationship between metaphor skills and Theory of Mind in middle childhood: Task and developmental effects. *Cognition*, 238, 105504.

Esercizi di stile. La narrativa per l'infanzia delle intelligenze artificiali

Francesco Cicero (Università degli Studi di Milano)

La narrativa per l'infanzia non è solamente uno dei settori più floridi dell'industria editoriale contemporanea, ma è anche uno dei primi ambiti della scrittura creativa in cui l'intelligenza artificiale sta trovando una diffusa applicazione. Racconti scritti e illustrati – per intero o in parte – da intelligenze artificiali sono, infatti, ormai disponibili in varie lingue nelle principali librerie online. Inoltre, esistono siti basati sulla nuova tecnologia (come, ad esempio, childbook.ai, createbookAI.com) che propongono strumenti semplici e intuitivi, destinati a rendere la creazione di libri per bambini veloce e accessibile a tutti. Che l'intelligenza artificiale generativa abbia facilmente attecchito nel campo della letteratura dell'infanzia non è motivo di sorpresa. In genere, i testi composti dall'IA sono caratterizzati da un ventaglio di soluzioni linguistiche e stilistiche ridotto e a tratti ripetitivo (cfr. per l'italiano, Cicero, 2023). Un aspetto che, allo stato attuale, limita le possibilità di applicazione ai generi letterari più ambiziosi, ma non a quelli generalmente definiti come paraletteratura, e che proprio nella prevedibilità trovano una delle loro cifre stilistiche (cfr. Ricci, 2013). Ma quali sono le caratteristiche linguistiche e stilistiche proprie della letteratura per l'infanzia composta dalle intelligenze artificiali? E con quale efficacia questi testi riescono ad approssimarsi agli scritti di autori umani?

L'intervento intende rispondere a queste domande presentando i risultati dell'esame di un corpus di racconti composti in italiano dalle intelligenze artificiali. A otto testi generati appositamente, facendo uso delle intelligenze artificiali più diffuse e accessibili in Italia (ChatGPT-3.5, Copilot e Gemini), saranno affiancate due opere di recentissima pubblicazione composte dall'IA: *Viaggio oltre l'ignoto* (Il Castoro, 2024), a cura di Pierdomenico Baccalario, Marco Magnone e Davide Morosinotto, e *La volpe e il futuro* (2024), del collettivo Roy Ming. Nello specifico, in primo luogo, l'analisi dei racconti si soffermerà sull'individuazione di elementi costanti dal punto di vista sintattico, testuale (deissi, segnali discorsivi…) e soprattutto lessicale e fraseologico (lessemi colloquiali, alterati, similitudini, espressioni idiomatiche, interiezioni…): aspetti che di solito determinano il «garbo stilistico» caratteristico dell'italiano dell'infanzia (Ricci, 2009). In secondo luogo, l'attenzione sarà dedicata alla misurazione della distanza linguistica che separa le parti narrative da quelle dialogiche, provando a evidenziare la sensibilità delle intelligenze artificiali nel riprodurre le movenze del parlato-scritto e nel differenziarlo dallo scritto-scritto (cfr. le categorie proposte in Nencioni, 1976).

Bibliografia citata

Cicero 2023: Francesco C., *L'italiano delle intelligenze artificiali generative*, in «Italiano LinguaDue», vol. 15, n. 2, pp. 733-761.

Nencioni 1976: Giovanni N., *Parlato-parlato*, *parlato-scritto*, *parlato-recitato*, in «Strumenti critici» 10, pp. 1-56 (poi in Id., *Di scritto e parlato*, *Discorsi linguistici*, Zanichelli, Bologna, 1983, pp. 126-179).

Ricci 2009: Laura R., *L'italiano per l'infanzia*, in *Lingua e identità. Una storia sociale dell'italiano*, a cura di Pietro Trifone, Carocci, Roma, pp. 323-350.

Ricci 2013: Laura R., *Paraletteratura. Lingue e stile dei generi di consumo*, Carocci Editore, Roma.

Vocabolari d'autore e Intelligenza Artificiale: è possibile creare il vocabolario della lingua di Boccaccio con ChatGPT?

Claudia Palmieri (Università per Stranieri di Siena)

Nel campo della ricerca linguistica, la sinergia tra lessicografia digitale e Intelligenza Artificiale (IA) rappresenta una frontiera di immense potenzialità e sfide. Questa convergenza sta rivoluzionando il nostro approccio all'analisi del linguaggio, ampliando gli orizzonti del modo in cui comprendiamo e interagiamo con le parole in un contesto digitale. La lessicografia digitale, che un tempo si occupava principalmente della digitalizzazione delle risorse testuali, ora abbraccia le capacità dell'IA per trasformare il modo in cui le informazioni lessicali vengono curate, analizzate e diffuse (cfr. de Schryver 2023).

Questo contributo si propone di esplorare l'interazione tra lessicografia e IA, ancorando la sua analisi al contesto del progetto VocaBO (*Vocabolario della lingua di Boccaccio Online*). Questa iniziativa, guidata dalla professoressa Giovanna Frosini, è frutto della collaborazione tra l'Università per Stranieri di Siena e l'Ente Nazionale Giovanni Boccaccio, in collaborazione con Cnr-Istituto di Linguistica Computazionale "Antonio Zampolli" (CNR-ILC), capofila nelle Digital Humanities, e l'Accademia della Crusca. Il lessico di Boccaccio, in particolare quello del *Decameron*, si distingue per la sua *varietas*, termine che indica un'ampia gamma semantica e una grande ricchezza terminologica (cfr. Alfano/Fiorilla/Quondam 2017). Questa ricchezza linguistica costituisce un punto di partenza ideale per esplorare le potenzialità e i confini dell'innovazione lessicografica attraverso l'IA.

L'analisi qui proposta prende in esame, in particolare, le relazioni che possono intercorrere tra lo studio della lingua e del lessico di Boccaccio e l'ausilio di ChatGPT. Per valutare se e come l'IA generativa possa costituire uno strumento utile alla ricerca in ambito storicolinguistico e alla redazione di un vocabolario d'autore, saranno illustrati i risultati ottenuti da una serie di indagini, condotte attraverso l'uso di ChatGPT per la compilazione di voci lessicografiche. Per la redazione delle voci, a ChatGPT verrà fornito il testo del *Decameron*, al fine di osservare come il modello gestisce un testo dell'italiano antico. Una particolare attenzione verrà dedicata anche all'elaborazione di prompt efficaci per generare risultati quanto più possibile soddisfacenti.

In un discorso tenuto in apertura della conferenza ASIALEX 2023, Michael Rundell si domanda se ChatGPT possa generare buoni dizionari con un input umano minimo e la risposta è negativa. Gli esperimenti condotti suggeriscono che ChatGPT può produrre dizionari accettabili, almeno per quanto riguarda i lemmi

più semplici. Ma un esame più attento rivela quasi sempre dei problemi di omissione di informazioni, di invenzione e, quindi, di inattendibilità (cfr. Rundell 2023). Parafrasando il quesito di Rundell, ci chiediamo: «ChatGPT è in grado di realizzare un vocabolario della lingua di Boccaccio?». Dare una risposta alla domanda è lo scopo di questo contributo.

Bibliografia

Alfano/Fiorilla/Quondam 2017 = Giovanni Boccaccio, *Decameron*, a cura di Giancarlo Alfano,
Maurizio Fiorilla, Amedeo Quondam, Milano, Rizzoli BUR, 2017.

de Schryver, Gilles-Maurice 2023 *Generative AI and Lexicography. The Current State of the Art Using ChatGPT*, in «International Journal of Lexicography», 36, 355-387.

Rundell, Michael 2023, *Automating the Creation of Dictionaries: Are We Nearly There?*, *Lexicography, Artificial Intelligence, and Dictionary Users* (Seoul, Yonsei University, 22-24 giugno 2023), 1-9.

Large Language models and regional variety of standard: the case of Italian.

Angelapia Massaro (Università del Salento) &
Giuseppe Samo (Beijing Language and Culture University)

Recent studies have observed that Large Language Models (henceforth LLMs) and Neural Networks (NNs) can easily perform syntactic tasks (Linzen & Baroni 2021 *inter alia*). An interesting result is the ability of marking between harder to parse and easier to parse structures, as well as grammatical and ungrammatical clauses (see the overview in Merlo & Samo 2024). In our presentation, we tackle a problem lying in-between with respect to data from Standard Italian and regional varieties of Italian: sentences that are grammatical and fully productive in one variety of a standard language and marked/ungrammatical in another variety, as the example of new information focus structures given in (1) for Sicilian Italian.

(1) - Chi è?
Who is it?
- Sono Salvo (Standard Italian)
am Salvo
- Salvo sono (Italian - Sicily)
Salvo am (Cruschina 2006: 369)

From a syntactic point of view, regional Italians tend to differ with respect to the availability of movement to the left periphery, triggered by discourse-related properties such as Focus or surprise (Benincà and Poletto 2004, Cruschina 2006), the realization of D elements as Topic markers (Ledgeway 2011), or also the in situ position of adverbial elements. As is well known, this level of syntactic variation is due to the presence of the local languages (the *dialects* - although this notion is not to be confused with the anglophone notion of *dialect*, which designates a variety of the standard language).

We perform two computational studies inspired by the works in psycholinguistics (see Wilcox et al. 2023), (i) grammaticality judgments-like task and (ii) a questionnaire-like task. As for (i) we retrieve priority scores of masked language models testing minimal pairs of devised ex-novo sentences representative of the structures (in the spirit of Samo & Merlo 2023). With respect to (ii), we run a series of “questionnaire-like” interactions with conversational AIs based on LLMs (Massaro & Samo 2023; see also Merlo 2023) and investigate the output as a corpus (Mikros et al.).

References

Benincà, P. and Poletto C. (2004). Topic, focus and V2: defining the CP sublayers. In L. Rizzi (E.) *The Structure of IP and CP. The Cartography of Syntactic Structures*, Vol. 2, 52–75. Oxford: Oxford University Press.

Cruschina, S. (2006). Informational focus in Sicilian and the left periphery. In M. Frascarelli (Ed.), *Phases of Interpretation*, 363-386. Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110197723.5.363

Ledgeway, A. (2011). Subject licensing in CP. *Mapping the Left Periphery: The Cartography of Syntactic Structures, Volume 5*, 257.

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, 7, 195-212.

Massaro, A., & Samo, G. (2023). Prompting Metalinguistic Awareness in Large Language Models: ChatGPT and Bias effects on the Grammar of Italian and Italian Varieties. *Verbum*, *14*, 4-4.

Merlo, P. (2023) Blackbird language matrices (BLM), a new task for rule-like generalization in neural networks: Can Large Language Models pass the test?. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 8119–8152, Singapore. Association for Computational Linguistics.

Samo, G., Merlo, P., 2023. Distributed computational models of intervention effects: a study on cleft structures in French. In C. Bonan & A. Ledgeway (Eds.), It-clefts: Empirical and Theoretical Surveys and Advances (pp. 157–180). Berlin, Boston: De Gruyter.

Generati, rigenerati, manipolati: uno studio sulle abilità di ChatGPT e Gemini nel trattamento di testi scritti scientifici di ambito linguistico

Alessandro Puglisi (Università per Stranieri di Siena)

La "varietà" di italiano scritto prodotta attraverso le tecnologie digitali e di Rete rappresenta un oggetto di studio assai frequentato dai linguisti. I diversi ambienti (chat, messaggistica istantanea, forum, e-mail, social media, blog, ecc.) nei quali essa si manifesta danno luogo a produzioni linguistiche in verità assai diversificate, come è stato notato, fra gli altri, da Antonelli (2016), Pistolesi (2014), Fiorentino (2016), Prada (2015). Ciò rende complessa, nonché scivolosa, una classificazione uniforme e coerente di tali produzioni, riconsegnandoci alla frammentarietà e, di fatto, impedendoci di parlare di una singola "varietà".

Assai meno studiata, al momento, è invece la lingua italiana scritta che viene prodotta e "riprodotta" da strumenti basati sull'intelligenza artificiale come i Large Language Models (LLM), i quali, com'è noto, solo di recente hanno conosciuto un tumultuoso sviluppo tecnologico, nonché grande visibilità nel dibattito pubblico.

Il contributo intende indagare le abilità linguistiche, in termini di trattamento dei testi, di due LLM. In particolare, si presenta uno studio "testa a testa" effettuato fra ChatGPT 3.5 (OpenAI) e Gemini 1.0 Pro (Google), entrambi gratuiti e utilizzabili da Web e applicazioni per dispositivi mobili. I due LLM vengono sollecitati a eseguire una serie di compiti su quattro estratti da testi scritti scientifici di ambito linguistico in italiano. Nello specifico, viene richiesto ai due sistemi, mediante tre differenti *prompt* per ogni compito, di 1) riassumere; 2) individuare i concetti chiave; 3) operare una semplificazione lessicale e sintattica, partendo da brevi estratti da due testi di Tullio De Mauro (2008 [1963]; 2014) e due di Luca Serianni (1987; 2003). La scelta specifica dei due studiosi italiani deriva dalla volontà di offrire al LLM una lingua densa, accorta ma, allo stesso tempo, attenta alla chiarezza dell'argomentazione e alla limpidezza del dettato.

I testi prodotti da ChatGPT 3.5 e Gemini 1.0 Pro vengono sottoposti ad analisi quantitativa e qualitativa. La prima si sostanzia nella rilevazione dell'estensione (in parole), della densità lessicale e nel calcolo dell'indice di leggibilità (con due diverse metodiche). Da un punto di vista qualitativo, invece, si verifica se il compito è stato effettuato e, in caso affermativo, quanto il risultato possa dirsi soddisfacente in relazione alla richiesta, e quali differenze lessicali e sintattiche emergono da testo a testo, a seconda del *prompt* utilizzato.

Una valutazione delle *performance* dei due LLM può contribuire in maniera significativa alla riflessione sul loro ruolo nella didattica dell'italiano (L1/L2/LS), in quanto i compiti proposti riguardano le cosiddette abilità di studio, il cui sviluppo e mantenimento è cruciale per studentesse e studenti, tanto nella scuola quanto in ambito universitario.

Bibliografia

Antonelli G. (2016), *L'e-taliano tra storia e leggende*, in Lubello S. (a cura di), *L'e-taliano. Scriventi e scritture nell'era digitale*, Firenze, Franco Cesati, pp. 11-28.

De Mauro T. (2014), *Storia linguistica dell'Italia repubblicana*, Roma, Laterza.

De Mauro T. (2008), *Storia linguistica dell'Italia unita*, Roma, Laterza, ed. or. 1963.

Fiorentino G. (2016), *Scrittori per caso: scritture spontanee sul web*, in Lubello S. (a cura di), *L'etaliano. Scriventi e scritture nell'era digitale*, Firenze, Franco Cesati, pp. 53-72.

Pistolesi E. (2014), *Scritture digitali*, in Antonelli G., Motolese L., Tomasin L. (a cura di), *Storia dell'italiano scritto, vol. III: Italiano dell'uso*, Roma, Carocci, pp. 349-375.

Prada M. (2015), *L'italiano in rete: usi e generi della comunicazione mediata tecnicamente*, Firenze, FrancoAngeli.

Serianni L. (2003), *Italiani scritti*, Bologna, Il Mulino.

Serianni L. (1987), *Scripta manent*, «Italiano & Oltre», 4, pp. 182-186.

Grammatica generata: accettabilità e inaccettabilità di costruzioni prodotte dai sistemi di generazione di testo

Mirko Tavosanis (Università di Pisa)

La diffusione dei Large Language Models (LLM) si è rapidamente accompagnata a un esame approfondito dei tratti caratterizzanti delle loro produzioni linguistiche. Questo esame, per quanto ancora ben lontano dall'essere considerabile definitivo e condiviso, ha evidenziato che tali sistemi, nonostante possiedano una notevolissima capacità di imitare la scrittura umana, a volte se ne allontanano in tratti che sembrano riconducibili a differenze strutturali nella gestione delle informazioni linguistiche. In particolare, le divergenze tra il comportamento umano e quello degli LLM risultano evidenti nei casi in cui sintassi e semantica sono intrecciate in modo più stretto. Per la lingua inglese sono in effetti già stati esaminati diversi contesti di questo tipo, soprattutto in rapporto alla gestione degli argomenti di verbi e sostantivi: per esempio, sono state documentate le difficoltà degli LLM nel gestire le Argument Structure Conjunctions (Weissweiler, Köksal e Schütze 2024) e nel generalizzare le strutture (Wilson, Petty e Frank 2023). Per l'italiano, sono stati mostrate divergenze nell'espressione della progressione tematica e nell'uso dei connettivi (De Cesare 2023).

In continuità con questa impostazione, il contributo prenderà in esame alcuni esempi relativi all'italiano, che, come altre lingue romanze, rimane ancora relativamente meno studiato (Cicero 2023). La prospettiva di riferimento è quella della testualità, in riferimento soprattutto alla coesione (Ferrari 2014) e con particolare attenzione alla gestione delle reggenze plurime (Serianni 1990). Queste ultime, in effetti, rappresentano un ambito in cui il comportamento degli LLM e quello degli esseri umani mostrano sia divergenze sia punti di contatto. Per esempio, le costruzioni in cui una completiva esplicita e una completiva implicita vengono collocate in dipendenza dallo stesso verbo e coordinate, sebbene evitate nello scritto formale, sono relativamente frequenti nei testi di chi apprende l'italiano scritto e nelle produzioni testuali degli LLM. Pertanto, è utile anche confrontare questi fenomeni con la variazione dei giudizi di accettabilità presso diversi tipi di pubblico, secondo un approccio empirico che in altri contesti ha portato a diverse conclusioni interessanti (Grandi 2018).

Bibliografia

Cicero, Francesco. 2023. "L'italiano delle intelligenze artificiali generative", *Italiano LinguaDue*, 2, pp. 731-751. https://riviste.unimi.it/index.php/promoitals/article/view/21990

De Cesare, Anna-Maria. 2023. "Assessing the quality of ChatGPT's generated output in light of human-written texts. A corpus study based on textual parameters", *CHIMERA*, 10, pp. 179-210.

Ferrari, Angela. 2014. *Linguistica del testo. Principi, fenomeni, strutture*. Roma, Carocci.

Grandi, Nicola. 2018. "Sulla penetrazione di tratti neo-standard nell'italiano degli studenti universitari. Primi risultati di un'indagine empirica", *Griseldaonline* 17, 1, pp. 1-24.

Serianni, Luca. 1990. *Prima lezione di grammatica*. Roma-Bari, Laterza.

Weissweiler, Leonie, Abdullatif Köksal e Hinrich Schütze. 2024. "Hybrid Human-LLM Corpus Construction and LLM Evaluation for Rare Linguistic Phenomena", *arXiv preprint arXiv:2403.06965*.

Michael Wilson, Jackson Petty, Robert Frank. 2023. "How Abstract Is Linguistic Generalization in Large Language Models? Experiments with Argument Structure", *Transactions of the Association for Computational Linguistics*, 11, pp. 1377–1395. https://doi.org/10.1162/tacl_a_00608

TECHNISCHE UNIVERSITÄT DRESDEN

Center for Italian Studies

What are you talking about? Generative AI and Language Understanding

Alessandro Lenci (Università di Pisa)

Large Language Models (LLMs) reveal outstanding linguistic abilities acquired through a simple, general-purpose word-prediction learning objective that allows them to extract huge amounts of knowledge encoded in distributional statistics. These so-called “emerging abilities” apparently include important aspects of pragmatic competence as well, such as for instance mastering indirect speech acts or interpreting figurative expressions. These phenomena are particularly important because they are usually considered to be the hallmark of human language understanding and creativity, and to presuppose sophisticated “theory of mind” capabilities. In this talk, I will present evidence from recent research investigating the performance of both standard and multi-agent LLMs in pragmatic tasks. Results will be discussed within the general question of the scope and nature of language understanding in Generative AI.

Validazione e confronto tra semplificazione automatica e semplificazione manuale di testi in italiano istituzionale ai fini dell'efficacia comunicativa

Giuliana Fiorentino (Università degli Studi del Molise)

In questo contributo gli autori sviluppano una ulteriore fase di un progetto che va avanti da circa un anno e mezzo. Obiettivo finale del progetto è costruire - con un modello di intelligenza artificiale ispirato a chat GPT - un ATS (denominato SEMPL-IT, applicativo risultante dal progetto PRIN VerbAcxSS) addestrato su un corpus di testi raccolto all'interno del progetto (denominato ItaIst), per semplificare testi di italiano istituzionale al fine di mettere a disposizione della Pubblica Amministrazione e degli utenti finali, i cittadini, uno strumento di lavoro facile da usare (in open access).

Gli autori – dopo aver semplificato automaticamente l'intero corpus ItaIst, valutando il risultato rispetto al testo di origine – hanno aperto una seconda fase di lavoro in cui è stata semplificata manualmente una porzione del corpus ItaIst. Questa seconda fase di lavoro ha permesso a) di confrontare e valutare quantitativamente il modo di semplificare del modello di intelligenza artificiale con la semplificazione manuale utilizzando diverse formule di semplificazione come Gulpease Index, Flesch Reading Ease; Flesch Kincaid Grade, Automated Readability oltre che considerando l'incremento del Vocabolario di Base (sui parametri e criteri da usare nella semplificazione di testi giuridici si vedano anche Brunato, Venturi 2014; Dell'Orletta et alii 2011); b) di studiare le diverse regole di semplificazione applicate spontaneamente dal software confrontandole con quelle applicate da due studiosi; c) di valutare che anche nella semplificazione manuale si possono applicare regole di semplificazione in modo non omogeneo (differenze tra i due revisori umani); d) di raggruppare e descrivere regole di semplificazione anche diverse da quelle normalmente suggerite dalla manualistica italiana corrente (Cortelazzo et alii 199; Cortelazzo, 2021). A questo proposito gli obiettivi che ci proponiamo sono i) valutare il diverso peso che parametri quantitativi (ad esempio lunghezza delle frasi o delle parole) e parametri qualitativi (ad esempio il soggetto esplicito preferibilmente animato) hanno nel raggiungere l'efficacia comunicativa ed inoltre ii) integrare maggiormente il ricorso a parametri qualitativi nel nostro applicativo (*SEMPL-IT*).

Presenteremo dunque i risultati di una terza fase del lavoro che consiste in test di valutazione da somministrare a due campioni di utenti di madrelingua italiana. Infatti a partire da 4 insiemi di testi paralleli che sono: 1. testi amministrativi originali; 2. testi semplificati dal modello di intelligenza artificiale; 3. testi semplificati manualmente da un revisore 4. testi semplificati manualmente da un

secondo revisore; si sono costruiti due test. Il primo consiste nel far valutare a un campione di utenti la qualità della semplificazione (campione di esperti, linguisti, studenti universitari); il secondo consiste nel valutare il diverso impatto dei 4 testi paralleli mediante domande di comprensione che consentiranno di verificare l'efficacia dei metodi di semplificazione utilizzati nelle 3 tipologie di testi.

L'articolo discuterà criticamente i risultati dei due test.

Bibliografia

Brunato D., Venturi G. 2014. "Le tecnologie linguistico-computazionali nella misura della leggibilità di testi giuridici". In *Informatica e diritto*, XL: 23, 2014, pp. 111-142.

Cortelazzo, M.A. 2021. *Il linguaggio amministrativo. Principi e pratiche di modernizzazione*. Roma, Carocci.

Cortelazzo, M.A., Pellegrino, F., Viale, M. Padova, 1999. *Semplificazione del linguaggio amministrativo. Esempi di scrittura per le comunicazioni ai cittadini* Comune di Padova.

Dell'Orletta F., Montemagni S., Venturi G. 2011. "*READ-IT: assessing readability of Italian texts with a view to text simplification*". In: SLPAT '11 – SLPAT '11 Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (Edimburgo, UK, 30 Luglio 2011). Proceedings of Association for Computational Linguistics Stroudsburg, 2011, pp. 73 – 83.

Fiorentino G., Ganfi, V. (in corso di stampa). Parametri per semplificare l'italiano istituzionale: revisione della letteratura. *Italiano LinguaDue*

Assessing the effectiveness of ChatGPT-3.5 and ChatGPT-4 in simplifying Italian institutional and administrative texts: first results of human evaluation

Mariachiara Pascucci (Università di Pisa) &
Claudia Gigliotti (Università degli Studi di Firenze)

Text simplification aims to enhance the accessibility of documents by reducing linguistic complexity. In the field of NLP studies, text simplification is a common task where the text is automatically rewritten to make it easier to understand. Such a process could have fruitful applications in the context of institutional and administrative communication. However, data on the effectiveness of the operation are not yet sufficient to guarantee the quality of the generated texts. This research aims to evaluate the performance of ChatGPT-3.5 and ChatGPT-4 in simplifying Italian institutional and administrative texts through a structured procedure encompassing data collection, automatic simplification, analyses, and human evaluation.

Specifically, the research employs a subset of regulatory texts, such as ministerial guidelines, which provide indications regarding central aspects for citizens' participation in public life. These texts exhibit a distinct bureaucratic style commonly found in Italian administrative documentation. Sourced from an expanding self-assembled corpus of Italian institutional and administrative documents, they serve as representative samples for analysis.

Simplification processes are conducted manually and via ChatGPT, with all resulting texts subject to evaluation in order to obtain comparative data on different simplifications. Manual simplification involves the careful revision and restructuring of texts by human annotators. Automated simplification is realized through ChatGPT-3.5 and ChatGPT-4, using specific prompts developed for the task to systematically generate simplified versions of the original documents, ensuring clarity and coherence while retaining essential information.
The study employs a blend of analyses and evaluations to verify the effectiveness of the simplification procedure. In particular, quantitative analysis entails the extraction and examination of key linguistic features from both original and simplified texts, providing quantitative insights into linguistic complexity and readability. The human evaluation, scheduled for May 2024, aims to provide a comprehensive assessment of the simplification outcomes, ensuring validity and reliability in the assessment process. Given the limitations of conventional evaluation metrics for neural network models, human evaluation assumes paramount importance. The evaluation is designed to discern the quality of generated texts, comparing them against human-authored counterparts. Italian

university students, specifically trained in text evaluation methodologies, have been recruited as evaluators. A rigorous evaluation procedure has been established encompassing intrinsic and extrinsic criteria to assess the quality of simplified texts.

Future developments include using the results of this survey to prepare the field for a study that will involve the use of eye-tracking methodologies. The aim is to provide useful information for setting up pre-tests for the evaluation of experimental stimuli in eye-tracking studies on readability. As Large Language Models continue to evolve, understanding their strengths and limitations in facilitating linguistic accessibility is crucial for informing the processes of the practical application of these tools in different areas. By identifying areas of improvement and best practices in simplification techniques, this research prepares the ground for advancing the state of the art in this research area. Using both evaluation and analysis, it seeks to contribute to the evolving discourse of text simplification methodologies and the role of LLMs in this field.

AImplicit – Reading between the lines with LLMs

Alessandro Panunzi (Università di Firenze)

Understanding human communication requires the comprehension of the implicit contents conveyed by the utterances. This involves not just decoding the explicit words, but also interpreting the underlying meanings, intentions, and contextual nuances that give those words their true significance.

In recent years, Large Language Models (LLMs) have demonstrated to be able to understand, generate and translate texts, performing exceptionally well in tasks that primarily focus on syntactic processing and semantic understanding. However, their capacity to grasp meaning from a pragmatic perspective is still under investigation, and scholars have not reached a consensus on the results (e.g. Bojic *et al.* 2024; Lee *et al.* 2024; Qiu et al. 2023; Sravanthi *et al.* 2024).

The talk mainly focuses on the ability of LLM to process pragmatic implicatures (implicit content crucial for the success of communication; Grice 1975), starting from a survey of very recent works on this topic. From a critical analysis of the literature, two main aspects emerge.

On one hand, the prompts used to obtain the desired output have a significant impact on the quality of the answers, highlighting the importance of how questions are framed (Kim *et al.* 2023). Prompt engineering has become a critical area of research, with scholars experimenting with different ways to phrase questions and commands to elicit the most accurate and relevant responses from LLMs (Vatsal & Dubey 2024). Changes in prompting techniques can lead to vastly different outputs, demonstrating the sensitivity of these models to input variations.

On the other hand, the performance of language models is significantly enhanced through fine-tuning strategies, involving the training of a pre-existing model on a specific dataset that is tailored to the desired application. (Ruis *et al.* 2023; Sravanthi *et al.* 2024).

Moreover, data on processing Italian contents in different language models will be discussed, underlining the importance of building a high-quality dataset specifically designed for this task. Creating high-quality datasets for pragmatic tasks requires a careful selection and annotation of textual data, but is essential for advancing the capabilities of LLMs in handling different languages and improving their overall performance (Sravanthi *et al.* 2024).

References

Bojic, L., Kovacevic, P., & Cabarkapa, M. (2023). GPT-4 Surpassing Human Performance in Linguistic Pragmatics". https://doi.org/10.48550/arXiv.2312.09545

Lee, B. J., & Cook, D.R. (2024). Exploring the Potential of AI for Pragmatics Instruction. http://dx.doi.org/10.2139/ssrn.4810301

Grice, H. P. (1975). Logic and Conversation. In P. Cole, & J. L. Morgan. (Eds.*), Syntax and Semantics, Vol. 3, Speech Acts* (pp. 41-58). New York: Academic Press, 41-58.

Kim, Z. M., Taylor, D. E., & Kang, D. (2023). "Is the Pope Catholic?" Applying Chain-of-Thought Reasoning to Understanding Conversational Implicatures.
https://doi.org/10.48550/arXiv.2305.13826

Qiu, Z., Duan, X., & Cai, Z. G. (2023). Does ChatGPT Resemble Humans in Processing Implicatures? In *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop (NALOMA23)*. Association for Computational Linguistics, 25-34.

Ruis, L., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., & Grefenstette, E. (2023). The Goldilocks of Pragmatic Understanding: Fine-Tuning Strategy Matters for Implicature Resolution by LLMs". In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds), *Advances in Neural Information Processing Systems 36 (NeurIPS2023)*. Curran Associates, 20827-20905.

Sravanthi, S.L., Doshi, M., Kalyan, P.T., Rudra Murthy, R., Bhattacharyya, P., & Dabre, R. (2024). PUB: A Pragmatics Understanding Benchmark for Assessing LLMs' Pragmatics Capabilities. https://doi.org/10.48550/arXiv.2401.07078

Vatsal, S, & Dubey, H. (2024). A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks. https://doi.org/10.48550/arXiv.2401.07078

Can Large Language Models process implicatures and presuppositions?
Engineering and testing prompting techniques to detect and classify implicit content in Italian political discourse.

Walter Paci (Università degli Studi di Firenze)

This work presents a study on various Large Language Models (LLMs) and their capabilities to detect and classify implicit content, namely Implicatures and Presupposition, in Italian political discourse.

Pragmatic proficiencies of LLMs have already been investigated at various levels and with different depths and results: previous experiences focused primarily on zero-shot, few-shot and Chain-of-Thought prompting to investigate to which extent LLMs can understand implicatures (Wei et al., 2022; Ortega-Martín et al., 2023; Qiu et al., 2023; Kim et al., 2023). In this work we test some of the state-of-the-art open-source and closed-source models currently available with various prompting techniques (PTs).

For our analysis, we extracted a test and a train datasets from the recent IMPAQTS corpus, an Italian spoken corpus of political discourse tagged for implicit contents (Cominetti et al., 2022). IMPAQTS theoretical framework is based on Sbisà (2015) and its later development conducted by Lombardi Vallauri and Masia (2014), Lombardi Vallauri et al. (2020), Cominetti et al. (2023). This framework includes four categories of linguistic implicitness: *presupposition, implicature, vagueness, and topicalization*. The scope of this study focuses on the first two categories of implicit content, i.e. presupposition and implicatures. Presuppositions involve taking information for granted or assuming that information as common ground in a conversation (Stalnaker, 2002). On the other hand, implicatures, that can be Conversational, Conventional and Generalized, are propositions implied but not explicitly stated in utterances, as defined by Grice (1975).

For our experiment, we are using state-of-the-art models (ChatGPTs 3.5-Turbo and Mistral, among others) to investigate if they can detect and classify implicit phenomena. To this aim, we are using foundational PTs, i.e., zero-shot and few-shot prompting, and advanced PTs like Chain-of-Thought, Tree-of-Thought (Yao et al., 2024) and Sociodemographic (Beck et al. 2024) prompting.

Two tasks are so faced:

1. A binary detection task, where the LLM must provide a Yes/No answer to a question asking if the input sentence has some implicit meaning or content.

2. A classification task where the LLM is asked to recognize if implicit contents are conveyed through implicatures or presuppositions.

For our tests, we randomly extracted 100 samples for each kind of implicit content, and 400 samples without implicit content from the IMPAQTS corpus, totalling 800 utterances. This whole dataset is used to test the former task, while a sub-dataset containing just 200 samples of utterances, 100 with a presupposition and 100 with an implicature, is used to test the latter.

References

Beck, T., Schuff, H., Lauscher, A., & Gurevych, I. (2024). Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting. *In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Cominetti, F., Cimmino, D., Coppola, C., Mannaioli, G., & Masia, V. (2023). Manipulative impact of implicit communication:: A comparative analysis of French, Italian and German political speeches. *Linguistik online*, *120*(2), 41-64.

Cominetti, F., Cimmino, D., Coppola, C., Mannaioli, G., & Masia, V. (2023). Manipulative impact of implicit communication:: A comparative analysis of French, Italian and German political speeches. *Linguistik online*, *120*(2), 41-64.

Cominetti, F., Gregori, L., Lombardi Vallauri, E., & Panunzi, A. (2022). IMPAQTS: un corpus di discorsi politici italiani annotato per gli impliciti linguistici. In *Corpora e Studi linguistici. Atti del LIV Congresso della Società di Linguistica Italiana (Online, 8–10 settembre 2021), a cura di Emanuela Cresti e Massimo Moneglia. Milano, Officinaventuno* (pp. 151-164).

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41-58). Brill.

Kim, Z. M., Taylor, D. E., & Kang, D. (2023). " Is the Pope Catholic?" Applying Chain-of-Thought Reasoning to Understanding Conversational Implicatures. *arXiv e-prints*, arXiv-2305.

Ortega-Martín, M., García-Sierra, Ó., Ardoiz, A., Álvarez, J., Armenteros, J. C., & Alonso, A. (2023). Linguistic ambiguity analysis in chatgpt. *arXiv preprint arXiv:2302.06426*.

Qiu, Z., Duan, X., & Cai, Z. G. (2023). Pragmatic implicature processing in ChatGPT.

Sbisà, M. (2015). *Detto non detto: le forme della comunicazione implicita*. Gius. Laterza & Figli Spa.

Stalnaker, R. (2002). Common ground. *Linguistics and philosophy*, *25*(5/6), 701-721.

Vallauri, E. L., & Masia, V. (2014). Implicitness impact: measuring texts. *Journal of Pragmatics*, *61*, 161-184.

Vallauri, E. L., & Masia, V. (2020). La comunicazione implicita come dimensione di variazione tra tipi testuali. In *Linguaggi settoriali e specialistici. Sincronia,*

diacronia, traduzione, variazione (Proceedings of the International SILFI Conference 2018) (pp. 113-120). Cesati Firenze.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2024). Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems, 36.

Female Anthroponyms and Occupation Names in LLM Outputs: Insights from the BioFem Corpus

Anna-Maria De Cesare (Technische Universität Dresden)

This talk explores the linguistic representation of prominent women in biographies generated by three LLMs from the GPT family and two models from the Mistral family. The analysis focuses on how these women are referred to, the lexical forms used to describe their professions, and the syntax of agreement. Four key aspects are examined: (i) the form and frequency of the anthroponyms used to refer to the women in the biographies (e.g., *Lina Bo Bardi, (la) Bo Bardi, Lina*); (ii) the frequency of feminine and masculine forms describing their professions, with particular attention to 12 influential female architects; (iii) the syntax of agreement between the noun *architetto/architetta* and its targets in the noun phrase (NP), as well as the agreement between the subject and nominal predicate; and (iv) a comparison of these findings with those observed in human-authored biographies. This case study aims to illuminate the linguistic portrayal of women in LLM-generated texts and consider the potential implications these representations may have on the future norms of the Italian language.

References

De Cesare, A.-M. 2023. Assessing the quality of ChatGPT's generated output in light of human-written texts. A corpus study based on textual parameters. *CHIMERA. Romance Corpora and Linguistic studies* 10, 179-210. https://revistas.uam.es/chimera/article/view/17979

De Cesare, A.-M. 2023. *Giorgia Meloni, Meloni* o *la Meloni?* La codifica degli antroponimi femminili in biografie generate da ChatGPT e pubblicate su *Wikipedia*. *Lingue e Culture dei Media* 7(1-2): 1–20, DOI: https://doi.org/10.54103/2532-1803/22388