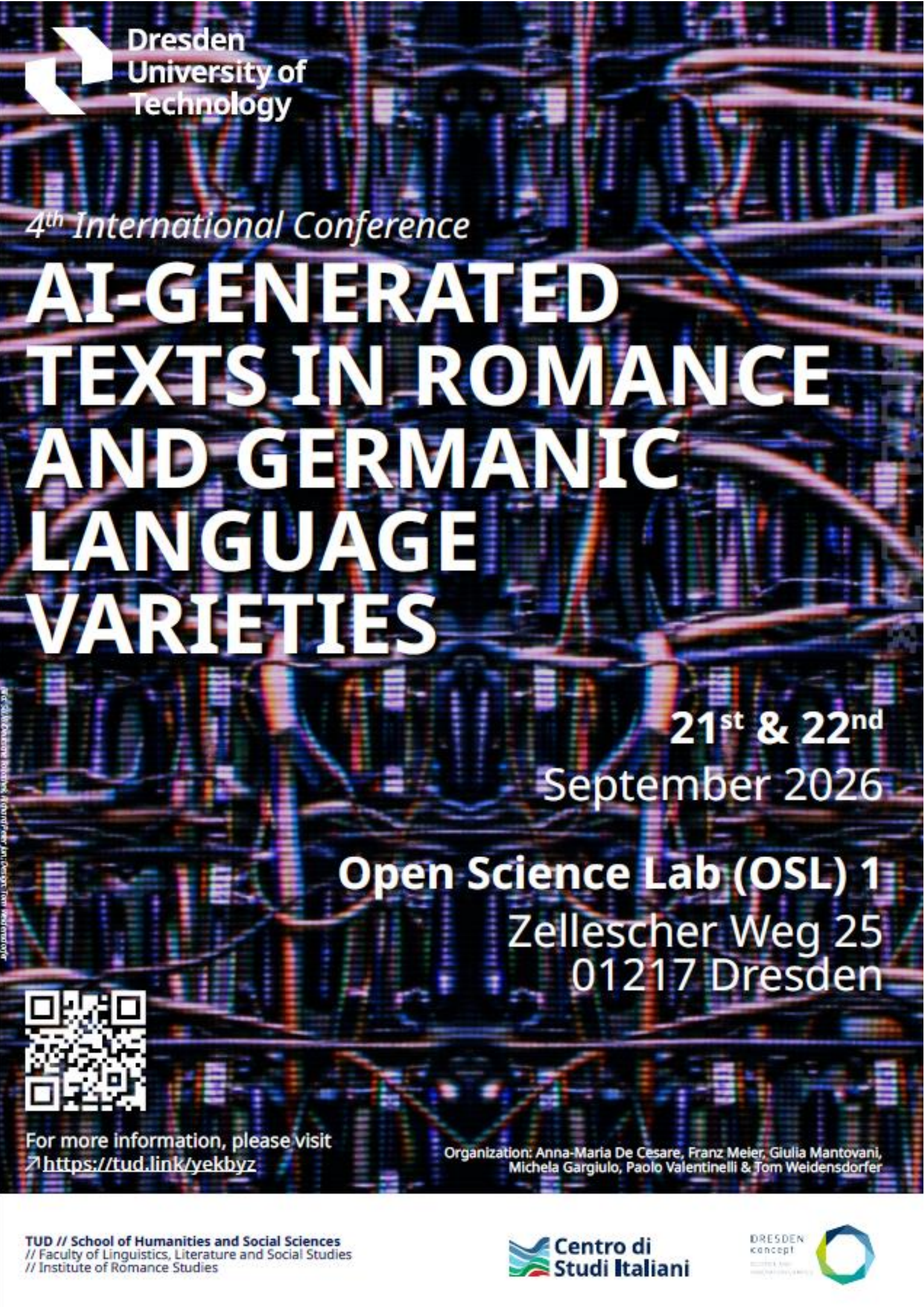

Dresden University of Technology

4th International Conference

AI-GENERATED TEXTS IN ROMANCE AND GERMANIC LANGUAGE VARIETIES

21st & 22nd
September 2026

Open Science Lab (OSL) 1
Zellescher Weg 25
01217 Dresden

For more information, please visit
↗ https://tud.link/yekbyz

Organization: Anna-Maria De Cesare, Franz Meier, Giulia Mantovani, Michela Gargiulo, Paolo Valentinelli & Tom Weidensdorfer

TUD // School of Humanities and Social Sciences
// Faculty of Linguistics, Literature and Social Studies
// Institute of Romance Studies

Centro di Studi Italiani

DRESDEN concept

This conference is funded by the Centro di Studi Italiani (CSI) / Zentrum für Italienstudien (ZI) at Technische Universität Dresden.
The Center for Italian Studies is a scientific institution of the Faculty of Linguistics, Literature and Cultural Sciences at the TU Dresden. The CSI promotes cooperation between the TU Dresden and Italian universities and cultural institutions and is also involved in the conception, realization and support of scientific and cultural events related to Italy.

To stay informed about the Center's activities and upcoming events, we invite you to subscribe to the newsletter using the form available on the Center's website.

Monday 21st September 2026

9:00 - 09:15	Conference opening **Anna-Maria De Cesare & Franz Meier** *(TU Dresden)*
9:15 - 10:15	**Keynote** (Chair: Franz Meier) **Sarah Brommer** *(Universität Bremen)* Sprachliche Vielfalt vs. ‚Supernorm'. KI-gestütztes Schreiben und die Verschiebung von Sprachnormen
10:15 - 10:30	*Coffee break*
10:30 - 11:00	**Barbara Heinisch** *(Eurac Research, Bolzano)* Developing an annotation scheme for terminology evaluation in large language model output for different German language varieties in the university domain
11:00 - 11:30	**Julian Schütz** *(Independent researcher)* Diatopic variation and AI translation: a contrastive analysis of formulaic constructions in regional varieties of Germany and Italy
11:30 - 12:00	**Franz Meier** *(TU Dresden)* Features of Orality in Fictional Dialogues Published in *Il Foglio AI*
12:00 - 12:30	**Francesco Saccà** (*TU Dresden*) Generating onomatopoeia for children's audiobooks: a multilingual comparative study on Multimodal Large Language Models' (MLLMs) sound-symbolic capabilities
12:30 - 14:30	*Lunch break*
14:30 - 15:00	**Paramita Mirza** *(TU Dresden)* Teuken: Building a Multilingual LLM for All of Europe
15:00 - 15:30	**Paolo Valentinelli** *(Università di Trento/TU Dresden)* What's Inside the Black Box? A Novel N-gram-based Approach to Understanding LLM Pre-training Datasets
15:30 - 16:00	**Anna-Maria De Cesare & Giulia Mantovani** (*TU Dresden*) & **Luca Moroni** *(Sapienza Università di Roma)* *Suffissi alterativi* in natural *vs* synthetic fables: an exploratory study
16:00 - 16:30	*Coffee break*
16:30 - 17:30	**Keynote** (Chair: Francesco Saccà) **Gašper Beguš** (*University of California, Berkeley*) Understanding humans, animals, and machines

Tuesday 22nd September 2026

9:30 - 10:30	**Keynote** (Chair: Anna-Maria De Cesare) **Valentina Crestani** *(Università degli Studi di Milano Statale)* Intelligenza artificiale: "persone" semplificate nei testi in lingua facile tedesca e italiana?
10:30 - 11:00	*Coffee break*
11:00 - 11:30	**Claudia Cusano & Alice Suozzi & Sevilnur Bahadir & Gianluca Lebani** *(Università Ca' Foscari Venezia)* Do AI-Generated News Amplify Bias? A Comparative Analysis of Femicide Reporting Across Political Orientations
11:30 - 12:00	**Ilaria Papagno & Vahram Atayan** *(Universität Heidelberg)* Evaluierung großer Sprachmodelle zur automatischen Vereinfachung medizinischer Texte in Einfache Sprache
12:00 - 12:30	**Christopher Haberland** (*University of Washington/Council for Ligurian Linguistic Heritage*) **& Stefano Lusito & Jean Maillard** (*Council for Ligurian Linguistic Heritage*) AI-Generated Texts in a Romance Underrepresented Language: A Pilot Study on Genoese Ligurian translation
12:30 - 14:30	*Lunch break*
14:30 - 15:30	**Keynote** (Chair: Giulia Mantovani) **Andrea Nini** *(University of Manchester)* Large Language Models and Forensic Linguistics: Challenges and Opportunities
15:30 – 16:00	**Ferdinando Longobardi & Laura D'Amodio & Gaia Valeria Gallo** *(Università degli Studi di Napoli L'Orientale)* Human and machine authorship: identifying LLM-Generated narratives through stylometric markers
16:00 - 16:30	**Mirko Tavosanis** *(Università di Pisa)* Discussion/Discussione
16:30 - 16:45	**Anna-Maria De Cesare & Franz Meier** *(TU Dresden)* Closing remarks

KEYNOTE

Sprachliche Vielfalt vs. ‚Supernorm'. KI-gestütztes Schreiben und die Verschiebung von Sprachnormen

Sarah Brommer (Universität Bremen)

KI-basierte Schreibwerkzeuge werden zunehmend in unterschiedlichen sozialen, institutionellen und wissenschaftlichen Kontexten der Textproduktion eingesetzt. Sie prägen damit nicht nur Schreibprozesse, sondern auch sprachliche Formen und die Bedingungen sprachlicher Normbildung. Da diese Systeme auf der Basis großskaliger Sprachdaten operieren und primär wahrscheinliche, konventionalisierte Muster reproduzieren, stellen sie keine neutralen Werkzeuge dar, sondern wirken als indirekte Instanzen sprachlicher Selektion.

Vor dem Hintergrund dieser Entwicklung stellt sich die Frage, wie sich das Spannungsverhältnis zwischen sprachlicher Vielfalt und möglichen Tendenzen zur Vereinheitlichung fassen lässt. Dabei wird davon ausgegangen, dass KI-gestütztes Schreiben weniger zur Entstehung neuer, klar abgrenzbarer Sprachnormen führt als vielmehr bestehende Normen verschiebt, verdichtet und re-konfiguriert.

Die Funktionsweise von KI kann dazu beitragen, Variation zu reduzieren und eine Art ‚Supernorm' zu stabilisieren, die quer zu Registern und Varietäten liegt und sich aus der statistischen Aggregation und Glättung heterogener Sprachgebrauchsmuster speist. Diese Supernorm ist weder mit der Norm im Coseriu'schen Sinne des allgemein üblichen Sprachgebrauchs noch mit der Norm im Sinne institutionell kodifizierter Regeln identisch, entfaltet jedoch normierende Effekte, indem sie sprachliche Variation systematisch reduziert und bestimmte Ausdrucksformen als besonders wahrscheinlich und damit als ‚unmarkiert' stabilisiert.

Der Vortrag diskutiert diese Dynamiken als Verschiebung des Normgefüges im Spannungsfeld von Vielfalt und Vereinheitlichung und zeigt, dass sich aufgrund neuer Relationen zwischen Variation, Kontextsensibilität und normativer Erwartbarkeit das Verhältnis von Norm, Abweichung und Qualität in grundlegender Weise verschiebt.

KEYNOTE

Understanding humans, animals, and machines

Gašper Beguš (University of California, Berkeley)

In this talk, I propose a framework for modeling language as informative imagitation, offering insights into human and non-human communication as well as into how AI models learn. I introduce ciwaGAN, a model that treats language learning as informative imitation and exhibits several linguistically and neurally motivated properties, including communicative intent, learning from raw unsupervised audio, and embodied representations. I also present an interpretability technique (CDEV), which allows us to perform linguistic and neural simulations on the models. I present a case study in which AI interpretability technique (CDEV) enabled the discovery of analogs to human vowels in whale communication. Understanding how generative models learn--and how they resemble or differ from humans--can not only provide insights into human language and cognition but also facilitate scientific discovery.

KEYNOTE

Intelligenza artificiale: "persone" semplificate nei testi in lingua facile tedesca e italiana?

Valentina Crestani (Università degli Studi di Milano Statale)

La produzione e la traduzione di testi scritti in lingua facile tedesca e in lingua facile italiana sono normate da linee guida che, specialmente per quanto riguarda alcuni aspetti, differiscono le une dalle altre. Tra questi rientra l'uso del linguaggio sensibile al genere. Nelle prime linee guida pubblicate per il tedesco, indicazioni su come scrivere le denominazioni di persona non venivano quasi menzionate. La recente norma tecnica DIN SPEC 33429 "Empfehlungen für Deutsche Leichte Sprache" (2025) dedica, invece, all'uso del linguaggio sensibile al genere il dettagliato paragrafo 5.2.13. Nelle linee guida pubblicate per l'italiano, indicazioni sul linguaggio sensibile al genere sono del tutto assenti. Ciò significa che autori e autrici di testi in lingua facile italiana devono risolvere autonomamente il problema di scrittura: a questo si aggiunge che, allo stato attuale, non esistono strumenti di intelligenza artificiale per la traduzione di testi in lingua facile italiana. La traduzione dei testi in lingua facile tedesca può, invece, essere supportata da vari di strumenti di intelligenza artificiale realizzati ad hoc (es. KLAO, SUMM AI).

Differenti linee guida e strumenti di intelligenza artificiale a disposizione per le due lingue facili portano con sé differenti possibilità di ricerca. Ai fini della comparabilità, sia per la lingua facile tedesca sia per la lingua facile italiana, i dati dell'analisi consistono in un corpus di comunicati stampa pubblicati nei siti web di città tedesche ed italiane e tradotti in lingua facile tramite ChatGPT con l'uso di varie strategie di prompting. Per la lingua facile tedesca, si aggiunge un corpus parallelo di comunicati stampa pubblicati nei siti web di città tedesche nella versione originale e nella versione in lingua facile, la cui versione in lingua facile è stata fatta tramite SUMM AI (cfr. Crestani 2025). Le domande della ricerca sulla traduzione delle denominazioni di persona riguardano le strategie di neutralizzazione e di visibilità:

- strategie di neutralizzazione: le strategie di neutralizzazione sono utilizzate maggiormente nelle traduzioni rispetto alle versioni originali? Le forme di neutralizzazione

vengono rese nelle traduzioni in lingua facile tramite forme corrispondenti?

- Strategie di visibilità: le strategie di visibilità vengono utilizzate maggiormente nei testi originali rispetto alle traduzioni? Le espressioni di visibilità vengono sostituite nelle traduzioni in lingua facile da forme di neutralizzazione (es. nomi epiceni) oppure da forme grammaticalmente conformi?

Bibliografia

Crestani, Valentina (2025): Die Deutsche Leichte Sprache zwischen menschlicher Übersetzung und KI-Übersetzung: eine Untersuchung zum Gebrauch geschlechtersensibler Sprache in Pressemitteilungen. In: Di Meola, Claudio/Gerdes, Joachim/Tonelli, Livia (Hg.): Sprachvariation im Deutschen zwischen Theorie und Praxis. Fachsprachlichkeit, Inklusion, Didaktik, Übersetzung, Kontrastivität. Berlin: Frank & Timme, pp. 283-314.

Deutsches Institut für Normung E.V. (2025), „DIN SPEC 33429 Empfehlungen für Deutsche Leichte Sprache“, https://www.din.de/de/mitwirken/normenausschuesse/naerg/e-din-spec-33429-2023-04-empfehlungen-fuer-deutsche-leichte-sprache--901210 (23.03.2026).

KLAO, https://klao.eu/ (23.03.2026).

SUMM AI, https://summ-ai.com/ (23.03.2026).

KEYNOTE

Large Language Models and Forensic Linguistics: Challenges and Opportunities

Andre Nini (University of Manchester)

The Large Language Models (LLMs) revolution will significantly impact the field of Forensic Linguistics, presenting both challenges and opportunities for research as well as influencing casework. In this keynote, I will address these aspects by outlining several new research directions in Forensic Linguistics currently being investigated at the Forensic Language Analysis Lab at the University of Manchester. I will also connect these findings to other related areas of Linguistics.

Firstly, I will report on research into authorship analysis and the potential of LLMs to assist in this crucial forensic linguistic task. Recent evaluations indicate that while LLM-based methods are perform well, they still do not significantly outperform traditional machine-learning methods. Consequently, more transparent methods should be preferred.

Secondly, I will report on findings regarding new research into LLM impersonation of individuals, a novel issue that could affect forensic casework in the near future. Our research demonstrates that existing authorship analysis systems are not deceived if a malicious actor attempts to use an LLM to impersonate someone via prompting alone. This result aligns with existing findings that LLMs are only superficially able to replicate a person's language while failing at a deeper level.

Finally, I will report on new research into the extent to which LLMs possess knowledge of English varieties and how this can be used to profile the authors of anonymous texts. Recent explorations reveal that state-of-the-art models possess this knowledge and perform better than chance at profiling. However, challenges remain in addressing hallucinations and mitigating the influence of content on model decisions.

Developing an annotation scheme for terminology evaluation in large language model output for different German language varieties in the university domain

Barbara Heinisch (Eurac Research, Bolzano)

Widely used large language models (LLMs) are known to not cover all languages or language varieties equally well. This inequality is even more pronounced in specialized communication, such as university texts. Although the European Higher Education Area has introduced a degree of standardization in university terminology, considerable national, regional and institution-specific variation persists. LLMs frequently fail to account for these differences, often defaulting to the dominant variety in generated texts. Addressing the standard varieties of German in Germany, Austria, Switzerland and South Tyrol, this research investigates how LLMs handle language-variety-specific terminology in university texts, including curricula, strategic documents and statutes, with particular attention to translation tasks between German and English.

Positioned at the intersection of variational linguistics, language for specific purposes and system-bound terminology, this study examines "terminology fit" in LLM-generated text outputs. To this end, it develops an annotation scheme that evaluates how well outputs align with terminology conventions across German language varieties. The scheme builds on established frameworks for translation quality assessment, particularly the Multidimensional Quality Metrics (MQM) and the ACTER Terminology Annotation Guidelines. Furthermore, it integrates recent LLM-based evaluation approaches such as GEMBA(-MQM V2).

While these frameworks provide a solid foundation, they do not explicitly capture variation-sensitive terminology, especially system-bound or language-variety-specific usage. In MQM, terminology errors are typically classified as (a) inconsistency with terminology resources, (b) inconsistent use of terminology, or (c) wrong term usage. However, variation across language varieties is only implicitly addressed within these categories. This limitation motivates the development of an extended, variation-aware

annotation scheme that also accounts for hallucinated or non-existing terms, an issue particularly salient in LLM-generated texts.

At its core, the proposed scheme retains MQM's error typology but refines and expands it along additional dimensions inspired by terminology variation typologies, which distinguish dialectal, functional, discursive, interlinguistic and cognitive causes of variation: (1) language variety (e.g. national or regional variety), (2) system-bound terminology (e.g. usage bound to different system levels), (3) text type, purpose and target audience (e.g. communicative function), and (4) term existence (distinguishing terms which are actually in use from fabricated or hallucinated ones). Where relevant, the framework can also be extended through relationship annotation (e.g. relations between concepts) and ontology-supported approaches.

Designed primarily for human annotation, the scheme argues that (terminology) annotation frameworks for LLM-generated texts must explicitly incorporate the variation due to language varieties as a central dimension of evaluation since LLM outputs often achieve general-domain adequacy but fail to meet variety-specific conventions. The proposed annotation scheme addresses this gap by foregrounding terminology variation as a central evaluation dimension, thereby contributing to the practical assessment of LLM performance in multilingual (university) contexts.

References

Freixa, Judit (2006): Causes of denominative variation in terminology. A typology proposal. In TERM 12 (1), pp. 51–77. DOI: 10.1075/term.12.1.04fre.

Heinisch, Barbara (in print): Beyond system-bound terminology: Multi-level variation in German university terminology and its consequences for terminology work.

Inkpen, Diana; Paribakht, T. Sima; Faez, Farahnaz; Amjadian, Ehsan (2016): Term Evaluator: A Tool for Terminology Annotation and Evaluation. In International Journal of Computational Linguistics and Applications 7 (2), pp. 145–165. Available online at https://hdl.handle.net/20.500.14721/14577.

Junczys-Dowmunt, Marcin (2025): GEMBA V2: Ten Judgments Are Better Than One. In Barry Haddow, Tom Kocmi, Philipp Koehn, Christof Monz (Eds.): Proceedings of the Tenth Conference on Machine

Translation. Suzhou, China: Association for Computational Linguistics, pp. 926–933. Available online at https://aclanthology.org/2025.wmt-1.67/.

Kocmi, Tom; Federmann, Christian (2023): Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl et al. (Eds.): Proceedings of the 24th Annual Conference of the European Association for Machine Translation. Tampere, Finland: European Association for Machine Translation, pp. 193–203. Available online at https://aclanthology.org/2023.eamt-1.19/.

Lommel, Arle; Gladkoff, Serge; Melby, Alan; Wright, Sue Ellen; Strandvik, Ingemar; Gasova, Katerina et al. (2024): The Multi-Range Theory of Translation Quality Measurement: MQM scoring models and Statistical Quality Control. In Marianna Martindale, Janice Campbell, Konstantin Savenkov, Shivali Goel (Eds.): Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations). Chicago, USA: Association for Machine Translation in the Americas, pp. 75–94. Available online at https://aclanthology.org/2024.amta-presentations.6/.

Lu, Qingyu; Ding, Liang; Zhang, Kanjian; Zhang, Jinxia; Tao, Dacheng (2025): MQM-APE: Toward High-Quality Error Annotation Predictors with Automatic Post-Editing in LLM Translation Evaluators. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, Steven Schockaert (Eds.): Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE: Association for Computational Linguistics, pp. 5570–5587. Available online at https://aclanthology.org/2025.coling-main.374/

MQM (2024): MQM (Multidimensional Quality Metrics). Available online at https://themqm.org/the-mqm-full-typology/.

Rigouts Terryn, Ayla (2021): ACTER Terminology Annotation Guidelines. Available online at http://hdl.handle.net/1854/LU-8503113.

Diatopic variation and AI translation: a contrastive analysis of formulaic constructions in regional varieties of Germany and Italy

Julian Schütz (Independent researcher)

Construction Grammar, phraseology, and formulaic language all centre on prefabricated linguistic units that are entrenched and frequently used. In line with Construction Grammar, phrasemes can be viewed as conventionalised form-function pairings (cf. Goldberg 2006: 3) that are not assembled compositionally through generative rules but are stored and retrieved as holistic units.

Across many languages, one encounters diatopically marked phrasemes ranging from proverbs, idioms, commonplaces, and mottos to more complex formulaic constructions such as Wellerisms and jokes which frequently cross the traditional syntactic boundary of phraseological units, i.e. the sentence boundary (Burger 2015: 15). Phrasemes are often characterised by a high degree of expressivity (cf. Imperiale/Autelli/Schafroth 2025: 15). Phenomena associated with spoken, informal language tend to be particularly rich in regional and cultural specificity. As such, they frequently serve as identity markers for local communities and reflect distinct values, humour traditions, and attitudes toward everyday life.

Like non-diatopically marked constructions, these units pose significant challenges for foreign language learners due to their idiosyncrasy in both form and meaning. At the same time, AI systems offer valuable support by identifying meanings, typical usage contexts, and semantic as well as functional properties, thereby facilitating the learning process. Nevertheless, the growing reliance on Large Language Models (LLMs) in machine translation raises fundamental questions about the recognition and reproduction of diatopic varieties and the preservation of linguistic diversity.

This contribution presents a contrastive analysis of how LLMs handle translating diatopically marked phrasemes and constructions, including et es wie et es, et kütt wie et kütt (Rhineland), schaffe, schaffe, Häusle baue (Swabia), and Angola (an Gola) kennt'sch misch totsaufm (Saxony), as well as cu nesci, arrinesci (Sicily), ha da passa' 'a nuttata (Naples), and piutost che

nient l'è mej piutost (Milan). First, the examples are translated into the respective standard varieties. In the second step, a bidirectional contrastive analysis is conducted, translating the regional phrasemes both ways: from the German varieties into Standard Italian and regional Italian varieties, and from the Italian varieties into Standard and regional German.

To assess how prompt quality affects translation outcomes, a zero-shot prompt is compared with a more explicit prompt that requests semantic and functional details (idiomatic/literal meaning, usage contexts). The outputs of several LLMs (ChatGPT, Claude, Gemini, Grok) are subsequently evaluated against translations produced by DeepL.

The core of the analysis – situated within current discussions on the quality of automated translation and AI-generated texts (Tavosanis 2024) – focuses on the diatopic markings of these constructions at the phonological, morpho-syntactic, and lexical levels. In addition, the contribution explores the role of formulaic constructions as microtexts, including their positioning within larger texts and their productive and expressive potential, especially when no equivalent exists in the target language (Gerhalter 2024). Do AI systems recognise the specific functional features of these constructions? Do they adequately detect elements of language of immediacy, or are such features levelled out in favour of an artificially generated standard norm?

Finally, the contribution discusses the broader implications of these findings for linguistic diversity, regional cultural identity, and (phraseo-)didactics.

References

Burger, H. (2010). *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt.

Gerhalter, K. (2024), "How do DeepL and ChatGPT process information structure and pragmatics?", *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* 1 (1). (DOI: 10.62408/ai-ling.v1i1.8)

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Imperiale, R./Autelli, E./Schafroth, E. (eds.) (2025), *Manuale di fraseologia italiana*, Alessandria: Edizioni dell'Orso.

Tavosanis, M. (2024), "Valutare la qualità dei testi generati in lingua italiana", *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* 1 (1). (DOI:10.62408/ai-ling.v1i1.14)

Features of Orality in Fictional Dialogues Published in Il Foglio AI

Franz Meier (Technische Universität Dresden)

In March 2025, the Milan-based broadsheet *Il Foglio* launched *Il Foglio AI*, a month-long experiment featuring a daily four-page supplement entirely generated by large language models (LLMs). Following the success of the experiment, the project has continued as a weekly feature since April 2025. Each edition of *Il Foglio AI* contains around 25 articles spanning diverse genres, including fictional face-to-face dialogues between two individuals with opposing political views. These dialogical debates constitute the focus of the present analysis, which explores the extent to which features of conceptual orality characteristic of natural face-to-face discourse are reflected in these dialogues.

We adopt the notion of conceptual orality from Koch and Oesterreicher (2011), who distinguish between the medial realization of a dialogue (spoken or written) and its conceptualization as more or less oral. In face-to-face communication, spoken mode and oral conception typically coincide. In fictional dialogue, by contrast, orality is artificially constructed through writing, as authors reproduce features characteristic of natural spoken interaction (Dufter/Hornsby/Pustka 2020). Such features may be of a universal nature, including phenomena related to metacommunication, turn-taking, topic management and interpersonal involvement. At the same time, conceptual orality may also be manifested through language-specific features of spoken Italian (*italiano parlato*).

With regard to AI-generated fictional dialogues, this raises two questions. First, how successful are LLMs in imitating face-to-face interaction? Which features of conceptual orality are present in the dialogues they generate? Second, what similarities and differences emerge when these generated texts are compared with human-produced forms of fictionalized orality intended for reading rather than performance, such as novels or comics, which potentially constitute an important part of the training data used for LLMs? The present study is based on a corpus of thirty fictional dialogues published in *Il Foglio AI* and examines the most salient manifestations of conceptual orality in AI-generated dialogue.

References

Dufter, Andreas, David Hornsby, and Elissa Pustka. "L'oralité mise en scène dans la littérature: aspects sémiotiques et linguistiques." Zeitschrift für französische Sprache und Literatur, vol. 130, no. 1, 2020, pp. 2–19.

Koch, Peter, and Wulf Oesterreicher. Gesprochene Sprache in der Romania: Französisch, Italienisch, Spanisch. De Gruyter, 2011.

Generating onomatopoeia for children's audiobooks: a multilingual comparative study on Multimodal Large Language Models' (MLLMs) sound-symbolic capabilities

Francesco Saccà (Technische Universität Dresden)

Recent developments demonstrate that Multimodal Large Language Models (MLLMs), perform well in iconicity tests, i.e. they successfully integrate visual or auditory cues into form-meaning association tasks (e.g., Cai et al. 2024, Jeong et al. 2026). Beyond the interpretation of iconic data, MLLMs are also improving in the generation of iconic outputs. For example, GPT-4 by OpenAI is able to generate artificial languages' highly iconic pseudowords, whose meanings are more accurately inferred by humans than those of words in distant natural languages (Marklová et al. 2025).

Such findings suggest that AI-generated iconicity can be employed to simplify communication across various fields, much like in human language. For example, these developments open up possibilities for applying tools typically used in the field of Augmentative and Alternative Communication (AAC). It has been nevertheless proven that text simplification through iconic elements can aid comprehension for recipients with complex communication needs and/or severe disabilities (Lloyd, Fuller & Arvidson 1997).

In this context, particular attention should be given to a specific category of recipients, i.e. children over the age of three, and a specific category of iconic elements, i.e. onomatopoeias. On one hand, the focus is justified by a significant developmental milestone; by the age of three children usually start to develop a better comprehension of spoken language and have enhanced capacities for iconic association (Stephenson 2009). On the other hand, onomatopoeias represent an ideal element for this demographic, since they function as 1:1 mappings, i.e. establishing a direct link between the acoustic signifier and its referent, and rapidly become embedded within the infantile lexicon of a wide range of languages (Dogana 1990).

In line with these premises, the research examines the onomatopoeias generated by three MLLMs, which are intended to

accompany children's audiobooks' narrative texts and to complement access to their propositional content. The inputs are derived from three corpora of children's audiobooks in English, Italian and French, sourced from the Streaker podcast platform, featuring only the narrator's voice and background music. To evaluate the models, two distinct input modalities are utilised: the raw audio files and their corresponding written transcriptions. Furthermore, to ensure linguistic alignment, the models are prompted with a written instruction in the language corresponding to the source material.

The aim of this study is to describe the efficiency of MLLMs' iconic synthesis operations, through the following linguistic analyses.

First, the congruence between the phonological component of the generated onomatopoeias and the semantic dimensions evoked by the audiobook text segments is evaluated. This is done by comparing the outputs with scientific research on sound-symbolic realities that are considered to be nearly universal, rather than language-specific (Sidhu, Vigliocco & Pexman 2022).

Secondly, the language-specific phonological form of the generated outputs is assessed through a comparison with three corpora of mimetic words in English (Jeong et al. 2026), French (Jeong et al. 2026) and Italian (De Mauro). The analysis aims to identify recurring generative strategies (e.g. composition, reduplication, creation of new forms, etc.) and to determine if there are in generated Italian and French outputs any signs of onomatopoeic standardisation towards lexical forms influenced by English phonology and morphology.

Finally, based on research on AAC for children (e.g., Romski & Sevcik 2005; Costantino 2011), the results of the descriptive analysis should assess the potential benefits of using MLLMs in the field of inclusive communication and content accessibility.

References

Cai, Zhenguang & Duan, Xufeng & Haslett, David & Wang, Shuqi & Pickering, Martin. 2024. Do large language models resemble humans in language use? In Kuribayashi, Tatsuki & Rambelli, Giulia & Takmaz, Ece & Wicke, Philipp & Oseki, Yohei (ed.), *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics.* 37–56. Bangkok: Association for Computational Linguistics.

Costantino, Maria Antonella. 2011. *Costruire libri e storie con la CAA. Gli IN-book per l'intervento precoce e l'inclusione*. Roma-Trento: Erickson.

Dogana, Fernando. 1990. *Le parole dell'incanto. Esplorazioni dell'iconismo linguistico*. Roma: FrancoAngeli.

Jeong, Jinhong & Lee, Sunghyun & Lee, Jaeyoung & Han, Seonah & Yu, Youngjae. 2026. Do Language Models associate sound with meaning? A multimodal study of sound symbolism. *Proceedings of the AAAI Conference on Artificial Intelligence* 40(37). 31247–31255. https://doi.org/10.1609/aaai.v40i37.40387 (last accessed on 30.04.2026).

Lloyd, Lyle L. & Fuller, Donald R. & Arvidson, Helen, H. 1997. Introduction and overview. In Lloyd, Lyle L. & Fuller, Donald R. & Arvidson, Helen, H (ed.), *Augmentative and Alternative* Communication. A handbook of principles and practices. 1–17. Boston: Allyn and Bacon.

Marklová, Anna & Milička, Jiří Ryvkin, Leonid & Lacková Bennet, L'udmila & Kormaníková, Libuše. 2025. Iconicity in large language models, *Digital Scholarship in the Humanities* 40(4). 1203–1224. https://doi.org/10.1093/llc/fqaf095 (last accessed on 30.04.2026).

Romski, MaryAnn & Sevcik, Rose A. 2005. Augmentative communication and early intervention: Myths and realities. *Infants & Young Children* 18. 174–185. https://doi.org/10.1097/00001163-200507000-00002 (last accessed on 30.04.2026).

Sidhu, David. M., Vigliocco, Gabriella, & Pexman, Penny M. 2022. Higher order factors of sound symbolism. *Journal of Memory and Language* 125. 1–14. https://doi.org/10.1016/j.jml.2022.104323 (last accessed on 30.04.2026).

Stephenson, J. (2009). Iconicity in the development of picture skills: typical development and implications for individuals with severe intellectual disabilities. *Augmentative and Alternative Communication* 25(3). 187–201. https://doi.org/10.1080/07434610903031133 (last accessed on 30.04.2026).

Teuken: Building a Multilingual LLM for All of Europe

Paramita Mirza (Technische Universität Dresden)

Most large language models speak English first and everything else second. Teuken challenges that assumption. It is a 7-billion-parameter model trained from scratch to serve all 24 official languages of the European Union, including the full family of Romance languages. In this talk we trace the multilingual thread that runs through every stage of the project: why language equity matters for European AI, how we sourced and filtered data across dozens of languages, how tokenizer design can inadvertently punish entire language families, and what instruction tuning across languages actually requires. We close with evaluation results and some honest reflections on what we still do not get right.

References

Ali, Mehdi & Brack, Manuel & Lübbering, Max & Wendt, Elias & Goher Khan, Abbas. 2025. Judging quality across languages: a multilingual approach to pretraining data filtering with Language Models. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 8859–8898. https://aclanthology.org/2025.emnlp-main.449.pdf (last accessed on 05/06/2026).

Ali, Mehdi et al. 2024. Tokenizer choice for LLM training: negligible or crucial? *Findings of the Association for Computational Linguistics: NAACL 2024*, 3907–3924. https://aclanthology.org/2024.findings-naacl.247.pdf (last accessed on 05/06/2026).

Ali, Mehdi et al. 2025. Teuken-7B-Base & Teuken-7B-Instruct: towards European LLMs. *arXiv*. https://doi.org/10.48550/arXiv.2410.03730

Arno Weber, Alexander & Thellmann, Klaudia & Ebert, Jan & Flores-Herr, Nicolas. 2024.Investigating multilingual instruction-tuning: do polyglot models demand for multilingual instructions? *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 20829–20855. https://aclanthology.org/2024.emnlp-main.1159.pdf (last accessed on 05/06/2026).

Mirza, Paramita & Weber, Lucas & Küch, Fabian. 2025. LASER: Stratified selective sampling for instruction tuning with dedicated scoring strategy. *Findings of the Association for Computational Linguistics:*

EMNLP 2025, 19949–19974. https://aclanthology.org/2025.findings-emnlp.1086.pdf (last accessed on 05/06/2026).

Thellmann, Klaudia & Fromm, Michael & Stadler, Bernhard & Schulze Buschhoff, Jasper. 2024. Towards Multilingual LLM Evaluation for European Languages. *arXiv*. https://doi.org/10.48550/arXiv.2410.08928

What's Inside the Black Box? A Novel N-gram-based Approach to Understanding LLM Pre-training Datasets

Paolo Valentinelli (Università di Trento/Technische Universität Dresden)

The statistical supremacy of English in the pre-training data of Large Language Models (LLMs) raises critical questions regarding the homogenisation of AI-generated texts. With most LLMs relying heavily on English (Brown et al. 2020; Johnson et al. 2026), widely spoken languages such as German and Italian (among many others) are unexpectedly relegated to a few-resource status, whereas minority languages are categorised as low-resource. Consequently, this imbalance forces LLMs to adapt target-language outputs to English linguistic patterns. Addressing the concern of whether AI systems reinforce dominant linguistic varieties, this study investigates two fundamental research questions: (i) to what extent do different LLMs produce uniform linguistic outputs, and (ii) what does this uniformity reveal about the composition of their pre-training datasets? This convergence potentially induces a simplification of discourse and, by extension, of language itself.

To assess potential convergence, the research analyses the textual outputs of ten state-of-the-art LLMs in German and Italian. To construct a comparable multilingual corpus, three distinct text genres categorised under three text types were chosen (Sabatini 1999; De Cesare 2021): (i) decrees for highly binding administrative texts; (ii) news reports for semi-binding informative texts; and (iii) short stories for loosely binding literary texts. Text generation was performed via Python API scripts using zero-shot natural language prompting with a dual-layer architecture (comprising both a system prompt and a user prompt). These prompts were tested across a spectrum of temperature profiles, ranging from deterministic (0.0) and standard (0.7–1.0) configurations to levels characterised by high stochasticity (1.0–2.0). The quantitative analysis employed Python libraries, such as SpaCy and Textstat, to process over 41,000 German sentences and 47,000 Italian sentences in 1,800 texts. Lastly, the study extracted and evaluated frequent n-grams (Manning & Schütze 1999) to identify domain-

specific occurrences and compare generative convergence across the different LLMs.

Preliminary results demonstrate that the linguistic convergence of LLMs strictly depends on the structural constraints of the requested text type. In highly binding administrative texts, LLMs exhibit maximum convergence, relying on identical formulaic conventions and lexical placeholders absorbed during their pre-training phases. In semi-binding informative texts, convergence remains high for unigrams but dissipates rapidly for bigrams. Conversely, in loosely binding literary texts, LLMs display maximum divergence and substantial generative autonomy. Crucially, trigram analysis reveals that structural convergence drops to near zero across all contexts. While LLMs share a fundamental vocabulary and certain recurring

structures, they construct syntactic sequences in an autonomous manner which is a result consistent with the fundamental characteristics of generative AI.

Crucially, these findings provide a direct lens into the architectural composition of pre-training datasets. The patterns of linguistic uniformity detected across various LLMs serve as a representative indicator of their foundation, implying that these models rely on a substantial volume of overlapping training material. Ultimately, this n-gram-based approach offers preliminary insights into the thematic and structural configurations within these datasets, representing a significant step toward opening what have traditionally been regarded as black boxes.

References

Brown, Tom B. & Mann, Benjamin & Ryder, Nick & Subbiah, Melanie & Kaplan, Jared & Dhariwal, Prafulla & Neelakantan, Arvind et al. 2020. Language models are few-shot learners. *arXiv*. https://doi.org/10.48550/arXiv.2005.14165

De Cesare, Anna-Maria. 2021. Tipologie testuali e modelli. In Antonelli, Giuseppe & Motolese, Matteo & Tomasi, Lorenzo. *Storia dell'italiano scritto V. Testualità*. Roma: Carocci, 57–86.

Johnson, Rebecca & Dias Duran, Leslye Denisse & Panai, Enrico & Menéndez González, Natalia & Pistilli, Giada & Kalpokienė, Julija & Bertulfo, Donald Jay. 2026. The ghost in the machine speaks with an American accent: cultural value drift in early GPT-3 and the case

for pluralist evaluation of generative AI. *AI Ethics* 6:212. https://doi.org/10.1007/s43681-026-01038-x

Manning, Christopher D. & Schütze, Hinrich. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge, Massachusetts: MIT Press.

Sabatini, Francesco. 1999. Rigidità-esplicitezza vs elasticità-implicitezza: possibili parametri massimi per una tipologia dei testi. In Coletti, Vittorio & Coluccia, Rosario & D'Achille, Paolo & De Blasi, Nicola & Proietti, Domenico (eds). 2011. *L'italiano nel mondo moderno. Saggi scelti dal 1968 al 2009.* Vol. II, Napoli: Liguori, 183–216.

Suffissi alterativi *in natural fables vs synthetic fables: an exploratory analysis of the impact of training data and LLM architectures*

Anna-Maria De Cesare[1] & Giulia Mantovani[1] & Luca Moroni[2]
([1]Technische Universität Dresden, [2]Sapienza Università di Roma)

The advent of large language models (LLMs), which have been available since late 2022, has introduced a new paradigm in the generation of written texts. Thanks to highly sophisticated computational techniques, based on transformers and attention mechanisms (for further details, see Vaswani et al. 2017), the output of language models has become very natural, that is, similar to texts produced by humans. Texts generated by LLMs, however, differ from natural ones at various levels (lexicon, punctuation, morphology, syntax). The discrepancies observed can be explained primarily by the training data, particularly the overrepresentation of English-language data in the corpora used to train LLMs. Several studies have highlighted a category of 'discrepancies' that can be termed algorithmic imprints of English (De Cesare 2026), found in texts generated in Italian by various models (see De Cesare 2023, 2024, 2026; Cicero 2023: 739; Tavosanis 2024: 17–19; Mantovani in press).

This paper empirically evaluates the naturalness of AI-generated texts compared to those written by humans by comparing a corpus of 'natural' fables (97,449 words) with a corpus of 'synthetic' fables generated by five LLMs (102,057 words): Llama3.1-8B, Mistral-7B-v0.1, OLMo-2, Minerva-7B-v1.0, and Minerva-7B-4bit. The models were selected because they represent different linguistic proportions of Italian and English in the training data and varying degrees of transparency: Minerva-7B and OLMo-2 are open-source models, characterized by the availability of documentation regarding the data and training configurations. OLMo-2 is trained primarily on English-language texts, while Minerva-7B (both v1.0 and 4bit) has a balanced distribution, with approximately 50% of the data in English and the remainder in Italian. In contrast, Mistral-7B-v0.1 and Llama3.1-8B are closed-source models, for which such information is not made public. To study the text generated in Italian, adapted versions of

Mistral-7B-v0.1 and Llama3.1-8B were used via the vocabulary adaptation technique (see Moroni et al. 2025), which allows for the use of tokenizers different from the model's original ones; specifically, a tokenizer optimized for the Italian language was adopted.

The study evaluates the naturalness of the fables generated by the five LLMs through an analysis of alterative suffixes. The case study of alteration, as a distinctive feature of the Italian language (see Simone 2010), allows for the identification of possible traces of English, which, as an analytic language, is less productive in the use of synthetic diminutives (see Dressler/Merlini Barbaresi 1994: 112; Schneider 2003: 10). Through careful model selection, this work examines how the number of tokens and the languages used during the pre-training phase influence the output. By using models adapted to Italian, it will also be possible to explore the impact of the tokenizer's fertility (i.e., the number of tokens per word) on the generation of fables.

References

Cicero, Francesco. 2025. Esercizi di stile: La narrativa per l'infanzia delle intelligenze artificiali. *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* 2(2). 1–21. https://doi:10.62408/ai-ling.v2i2.19

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. Romance Corpora and Linguistic studies* 10, 179-210. https://revistas.uam.es/chimera/article/view/17979 (last accessed on 17/04/2026).

De Cesare, Anna-Maria. 2024. Nuove dinamiche di contatto linguistico: Le "impronte digitali" dell'inglese nell'italiano generato da LLM. *LId'o. Lingua Italiana d'Oggi* XXI. 67–91.

De Cesare, Anna-Maria. 2026. *L'italiano sintetico dell'intelligenza artificiale generativa*. Firenze, Cesati.

Dressler, Wolfgang & Merlini Barbaresi, Lavinia. 1994. *Morphopragmatics. Diminutives and Intensifiers in Italian, German, and Other Languages.* Berlin/New York, De Gruyter.

Mantovani, Giulia. In stampa. Forme sintetiche e forme analitiche per i diminutivi dei nomi: un confronto tra favole naturali e favole generate dall'intelligenza artificiale. In Lala, Letizia & Castro, Enrico

& Garzonio, Jacopo (a cura di), *Il nome in italiano. Morfologia, sintassi, testualità*. Firenze: Cesati.

Moroni, Luca & Puccetti, Giovanni & Huguet Cabot, Pere-Lluís & Bejgu, Andrei Stefan & Miaschi, Alessio & Barba, Edoardo & Dell'Orletta, Felice & Esuli, Andrea & Navigli, Roberto. 2025. Optimizing LLMs for Italian: Reducing token fertility and enhancing efficiency through vocabulary adaptation. *Findings of the Association for Computational Linguistics: NAACL 2025*. 6661–6675. Albuquerque, New Mexico. Association for Computational Linguistics.

Schneider, Klaus P. 2003. *Diminutives in English*. Tübingen, Max Niemeyer.

Simone, Raffaele. 2010. *Lingue romanze e italiano*, Enciclopedia dell'italiano. https://www.treccani.it/enciclopedia/lingue-romanze-e-italiano_(Enciclopedia-dell'Italiano)/ (last accessed on 17/04/2026).

Tavosanis, Mirko. 2024. Valutare la qualità dei testi generati in lingua italiana. *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses* 1(1). 1–24. https://doi:10.62408/ai-ling.v1i1.14

Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan N. & Kaiser, Lukasz & Polosukhin, Illia. 2017. Attention is all you need. *arXiv*. https://doi.org/10.48550/arXiv.1706.03762

Do AI-Generated News Amplify Bias? A Comparative Analysis of Femicide Reporting Across Political Orientations

Claudia Cusano & Alice Suozzi & Sevilnur Bahadir & Gianluca Lebani (Università Ca' Foscari Venezia)

The massive use of Generative AI and Large Language Models (LLMs) raises questions about reliability and bias of AI-generated texts. Data used to train LLMs contain inherent biases, which are perpetuated or even amplified in AI-generated data [1], warranting a systematic comparison of human-written and AI-generated texts. While several studies focused on this issue [2], [3], [4], [5] and on bias detection in political news [6], [7], [8], fewer works explored implicit biases in non political news [9], [10], [11], especially comparing different political standpoints.

This study aims to address this gap by comparing human-written and AI-generated non-political news articles on femicide cases from three political leanings (i.e. right-wing, left-wing, centrist). We focused on femicides because of their high susceptibility to gendered narratives. Despite international efforts [12], news discourse often reproduces biased gender stereotypes. Femicide reporting represents a critical testing ground for examining how human- and AI-generated texts, across different political orientations, encode these biases.

We selected eight high-profile femicide cases that occurred in the US. For each case, we collected three articles from *The Guardian*, *NewsNation* and *Fox News*, well-known outlets representing diverse ideological and stylistic profiles, which were characterized by GPT-5.4 [13]: for each outlet, a short and a longer description (summarizing vs. detailing political stance, writing style, and typical framing of social issues) were created, used to guide article generation. For each case, ChatGPT adopted three distinct personas, corresponding to journalists writing for left-leaning, centrist, and right-leaning newspapers. Two prompting strategies were employed: (i) original titles of the corresponding human-written articles as input; (ii) neutral summary of the event produced by ChatGPT as input.

For each case, we employed a fully crossed 2×2×3 design, manipulating prompting input (title vs. summary), outlet

description length (short vs. long), and persona ideology (left-leaning, centrist, right-leaning). The final dataset contains 120 articles.

Articles were analyzed from linguistic and narrative perspectives. The linguistic measures are lexical density, distribution of grammatical categories, sentence length, average depth of the syntactic tree, subordinates and average depth of subordinates [4], [14], [15].

Articles were then manually annotated for 15 features (Table 1), reflecting higher or lower adherence to the Istanbul Convention guidelines [12] and typically associated with left-leaning (higher adherence) and right-leaning (lower adherence) narratives. Each article was scored by subtracting lower-adherence features from higher-adherence ones.

Linguistic measures distinguish AI-generated from human-written articles: the former are more lexically dense, with longer sentences and greater overall syntactic complexity. The latter exhibit a higher frequency of subordinate clauses. No differences emerge across ideological orientations.

The narrative feature-based score (Figure 1) primarily distinguishes AI-generated from human-written articles, irrespective of ideological orientation. Significant variation emerges by ideology only within AI-generated texts: left-leaning texts display higher adherence than centrist and right-leaning ones, placing greater emphasis on structural frames and on the victim. Notably, such features are not observed in the corresponding human-written articles, indicating that AI-generated texts may reproduce or amplify ideologically patterned framing even when these elements are not explicit in the human-written counterparts.

Narrative Features	Feature Type	Description
euphemisms_elusion	Lower-Adherence	Use of terms like 'tragedy', 'disappearance' or 'domestic dispute' that downplay the act.

implicit_victim_blaming	Lower-Adherence	Subtle suggestions that the victim may bear partial responsibility.
narrative_focus_defense	Lower-Adherence	Focus on the killer's defense strategy.
narrative_focus_prosecution	Higher-Adherence	Focus on the victim's perspective and accusations.
narrative_focus_victim	Higher-Adherence	Primary emphasis on the victim's story.
perpetrator_emotional_state	Lower-Adherence	Descriptions of the perpetrator's emotions (e.g., jealousy, distress).
perpetrator_kindness_portrayal	Lower-Adherence	Portrayal of the perpetrator as a "good man" at heart.
perpetrator_social_status	Lower-Adherence	Information about the perpetrator's socioeconomic status or reputation.
relational_conflict_focus	Lower-Adherence	Framing the event as a couple conflict or relationship issue.
responsibility_perpetrator	Higher-Adherence	Attribution of responsibility primarily to the perpetrator.
responsibility_victim	Lower-Adherence	Attribution of responsibility partially or fully to the victim.
structural_frames	Higher-Adherence	References to broader systemic causes (e.g., misogyny, gender-based violence).
victim_extramarital_relations	Lower-Adherence	Mentions of alleged or real extramarital relationships of the victim.
victim_physical_appearance	Lower-Adherence	Descriptions of the victim's physical appearance.

violence_sensationalization	Lower-Adherence	Use of dramatic or sensational language to describe the violence.

Table 1: The set of narrative features used to annotate newspaper articles.

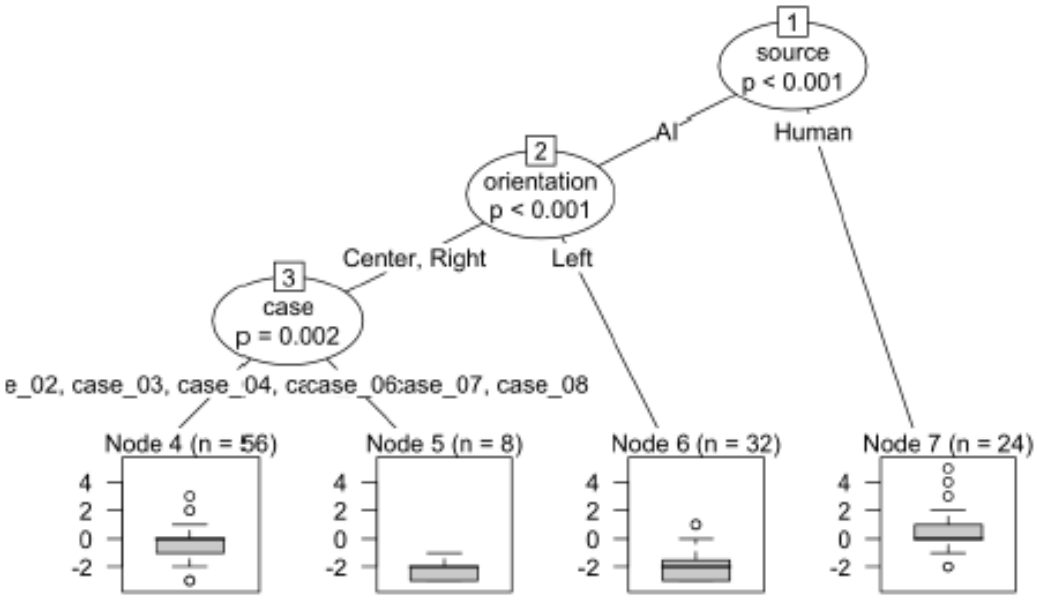

Figure 1. Conditional inference tree. Dependent variable, feature-based score; independent variable: source, orientation, case. The boxplots represent the score (higher score = lower adherence; lower score = higher adherence).

References

[1] De Cesare, Anna-Maria (2026). *L'italiano sintetico dell'intelligenza artificiale generativa*. Franco Cesati.

[2] Rosenfeld, A., & Lazebnik, T. (2024). Whose LLM is it Anyway? Linguistic Comparison and LLM Attribution for GPT-3.5, GPT-4 and Bard. *arXiv.Org*, arXiv.org, 2024-02.

[3] Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., … Wu, Y. (2023). *How Close Is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection*.

[4] Gersh, V. (2025) *Syntactic and Emotional Properties of AI-Generated Texts: a Corpus-Based Comparison* (Master's Thesis). Università Ca' Foscari Venezia. https://hdl.handle.net/20.500.14247/26177

[5] Blevins, T., Schmalwieser, S., & Roth, B. (2025). Do language models accommodate their users? A study of linguistic convergence. *arXiv.Org*, arXiv.org, 2025-08.

[6] Ozbagriacik, U., & Dubossarsky, H. (2026). *Language matters: Target-language supervision for political bias detection in Turkish news*. In *Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026)* (pp. 72–81). Association for Computational Linguistics.

[7] Lim, S., Jatowt, A., Färber, M., & Yoshikawa M. (2020). Annotating and Analyzing Biased Sentences in News Articles using Crowdsourcing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 1478–1484), Marseille, France. European Language Resources Association.

[8] Raza, S., Rahman, M., & Ghughe, S. (2024). Dataset Annotation and Model Building for Identifying Biases in News Narratives. *CEUR Workshop Proceedings*, *3671*, 5–15.

[9] Lee, S., Kim, J., Kim, D., Kim, K. J., & Park, E. (2023). Computational approaches to developing the implicit media bias dataset: Assessing political orientations of nonpolitical news articles. *Applied Mathematics and Computation*, *458*, 128219.

[10] Valentinelli, P. (2025). *Between Human and Artificial Language: Comparing the Syntax of Large Language Models and German Journalistic Texts* [Conference abstract]. Automated Texts in the Romance and Germanic Languages Conference.

[11] Meier, F. (2025). *Attitude and subjectivity in human-written and AI-generated editorials published in Il Foglio* [Conference abstract]. Automated Texts in the Romance and Germanic Languages Conference.

[12] Council of Europe. (2011). *Explanatory Report to the Council of Europe Convention on preventing and combating violence against women and domestic violence*. https://rm.coe.int/1680a48903

[13] OpenAI. (2026). *ChatGPT (March 2026 version)* [Large language model]. https://chat.openai.com

[14] Brunato, D., & Venturi, G. (2022). Why is this language complex? Cherry-pick the optimal set of features in multilingual treebanks. *Linguistics Vanguard*, *9*(s1), 59-72. https://doi.org/10.1515/lingvan-2021-0078

[15] Lu, X. (2011), A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers' Language Development. *TESOL Quarterly*, *45* (1): 36-62. https://doi.org/10.5054/tq.2011.240859

Evaluierung großer Sprachmodelle zur automatischen Vereinfachung medizinischer Texte in Einfache Sprache

Ilaria Papagno & Vahram Atayan (Universität Heidelberg)

Das Verständnis medizinischer Informationen ist eine wesentliche Voraussetzung für die Selbstbestimmung von Patientinnen und Patienten. Gesteigerte Verständlichkeit, reduzierte Missverständnisse und eine verbesserte Arzt-Patienten-Kommunikation tragen nachweislich zur Entlastung des Gesundheitssystems bei und gelten als zentrale Ziele der Gesundheitskompetenzförderung [1]. Der Einsatz von Einfacher Sprache im medizinischen Kontext stellt dabei eine etablierte Strategie dar, um die Zugänglichkeit gesundheitsrelevanter Informationen für eine breite Bevölkerungsgruppe zu verbessern [2][3]. Da die manuelle Textvereinfachung jedoch ressourcenintensiv ist, rücken Large Language Models (LLMs) zunehmend als potenzielle Automatisierungswerkzeuge in den Fokus der Forschung [4]. Der vorliegende Beitrag präsentiert eine systematische Evaluation von vier LLMs anhand eines Korpus von ca. 160 automatisch vereinfachten Texten aus der ApothekenUmschau, einer der meistgelesenen deutschsprachigen Gesundheitswebseite. Die ausgewählten Texte entstammen verschiedenen medizinischen Themenbereichen und für jeden Text liegt eine menschlich erfasste Version in Einfacher Sprache, die zum Goldstandard für diese Arbeit dient. Die Quelltexte werden mit identischem Prompt durch alle Modelle vereinfacht, um eine direkte Vergleichbarkeit zu gewährleisten. Ziel ist es, modellspezifische Merkmale der generierten Texte zu identifizieren und zu beschreiben, und gleichzeitig zur Systematisierung von Einfacher Sprache im medizinischen Bereich beizutragen. Die Analyse wird auf drei Ebene erfolgen. Auf der lexikalischen Eben werden Operationen wie die Ersetzung medizinischer Fachtermini durch allgemeinsprachliche Entsprechungen und das Frequenzprofil des generierten Wortschatzes untersucht. Auf der syntaktischen Ebene stehen Satzlänge, Subordinationsgrad, Passivkonstruktionen und Kohäsionsmittel im Vordergrund. Die konzeptuelle Ebene widmet sich die Frage, wie die Modelle mit fachspezifischen Konzepten umgehen, d.h., ob die Inhalte adäquat reformuliert, vereinfacht

oder ausgelassen werden und ob die Hinzufügung nicht im Ausgangstext vorhandener Informationen stattfindet [5]. Anhand der Ergebnisse werden Empfehlungen für gezielte Post-Editing-Interventionen formuliert. Die vereinfachten Texte werden mithilfe automatische Metriken ebenfalls evaluiert: Etablierte Indizes wie der Flesch-Reading-Ease [6] werden zur Messung der Lesbarkeit herangezogen, während BLEU [7] und SARI [8] werden einen systematischen Vergleich mit dem Goldstandard ermöglichen. Die Arbeit versteht sich nicht nur als Beitrag zur LLM-Evaluationsforschung, sondern auch als praktisch orientierte Ressource für den Einsatz von KI-gestützter Textvereinfachung im professionellen Kontext. Die Ergebnisse sollen zudem Aufschluss darüber geben, inwiefern unterschiedliche Modelle als komplementäre Werkzeuge eingesetzt werden können.

Literatur

1. Schaeffer, D., Hurrelmann, K., Bauer, U. und Kolpatzik, K. (Hrsg.): Nationaler Aktionsplan Gesundheitskompetenz. Die Gesundheitskompetenz in Deutschland stärken. Berlin: KomPart 2018.
2. Maaß, C. (2015). *Leichte Sprache*. Das Regelbuch. LIT Verlag.
3. Bock, B. M. (2019). *„Leichte Sprache" – kein Regelwerk*. Frank & Timme.
4. Deilen, S. et al. (2024) Towards AI-supported Health Communication in Plain Language: Evaluating Intralingual Machine Translation of Medical Texts. In *Proceedings of the First Workshop on Patient-Oriented Language Processing* (CL4Health) @ LREC-COLING 2024, pages 44-53, Torino, Italia. ELRA and ICCL.
5. Stodden, R. (2026). *Automatic German Text Simplification: Data, Evaluation, and Models*. Frank & Timme.
6. Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221-233.
7. Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of ACL 2002*, 311-318.
8. Xu, W., Napoles, C., Pavlick, E., Chen, Q., & Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4, 401-415.

AI-Generated Texts in a Romance Underrepresented Language: A Pilot Study on Genoese Ligurian Translation

Christopher Haberland[1,2] & Stefano Lusito[2] & Jean Maillard[2]
([1]University of Washington, [2]Council for Ligurian Linguistic Heritage)

Computer-assisted translation has existed for several decades; however, the languages involved are almost always major national languages, and only rarely regional or minority varieties. This also applies to the regional languages of Italy, which are notably distinct from Italian in terms of phonetics, morphology, syntax, and vocabulary. While this situation has begun to improve in recent years with the emergence of freely accessible machine translation tools (such as Google Translate and DeepL), significant challenges remain in achieving satisfactory results.

One such case is Genoese Ligurian, an Italian regional variety with a substantial body of written texts and a well-established literary and orthographic tradition (Toso 2009). In recent years, several machine translation systems have been developed for this language, some of which have raised legitimate concerns regarding their effectiveness (Ferrante 2024). Against this background, we analyse the output of various state-of-the-art AI models when tasked with translating Italian text to Ligurian. First, we evaluate the performance of these systems on well-established automatic machine translation benchmarks (ZenaMT, BOUQuET, FLORES+) and find that the results show substantial gaps in performance, with most models achieving scores that indicate low adequacy and substantial grammatical degradation. Second, we follow this up with an in-depth manual evaluation of system outputs on texts from a diverse set of domains and with varying degrees of syntactic complexity; this analysis confirms the presence of significant and systematic weaknesses across systems.

Table 1: Performance of various AI models on standard Italian-Ligurian translation tasks, measured in terms of chrF++ metric (higher is better).

Model	ZenaMT	BOUQuET	FLORES+
Grammatist	**74.9**	**51.1**	47.2
NLLB-3.3B	53.6	38.5	40.6
Google Translate	52.0	48.6	**47.3**
ChatGPT-5.4	50.6	39.7	40.1

Finally, building on our analysis of the gaps in current state-of-the-art AI, we describe a new experimental machine translation system, which we call Grammatist. Rather than constituting a standalone model, Grammatist augments state-of-the-art AI systems through the targeted use of a new open-access Genoese Ligurian lexicographical resource (DEIZE) based on scientific criteria (several of which are drawn from and partly adapted from Lusito 2025). The system relies on a selective and context-aware retrieval of lexical information, ensuring that only the most relevant entries are leveraged at inference time. We show that this approach outperforms other systems, particularly in domains that are rich in proper nouns – such as toponyms – and in cases requiring accurate handling of domain-specific lexical material that is typically underrepresented or absent from standard training datasets. Crucially, the approach is modular: it can be integrated into different large language models via tool-calling mechanisms, allowing it to remain compatible with, and benefit directly from, ongoing advances in underlying AI systems without requiring retraining. This same modularity also makes it readily adaptable to other languages.

References

BOUQuET = Andrews, Pierre et al. (2025). "BOUQuET: dataset, Benchmark and Open initiative for Universal Quality Evaluation in Translation" in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27515-27535, Association for Computational Linguistics.

DEIZE = Maillard, Jean et al. (2021-2026). *Diçionäio italian-zeneise*, online at https://conseggio-ligure.org/it/dizionario/deize.

Ferrante, Edoardo (2024). "Google non conosce tutti i termini genovesi, per questo la sua traduzione è imperfetta" in *Il Secolo XIX*, p.18, interview by A. Acquarone, 2024-07-18.

FLORES+ = Burchell, Laurie et al. (2024). "Findings of the WMT 2024 Shared Task of the Open Language Data Initiative" in *Proceedings of the Ninth Conference on Machine Translation*, pp. 110-117, Association for Computational Linguistics.

Lusito, Stefano (2025). *Aspetti teorici e pratici della redazione di un dizionario genovese-italiano della lingua contemporanea: metalessicografia di una varietà romanza di koinè*. Alessandria: Edizioni dell'Orso.

Toso, Fiorenzo (2009). *La letteratura ligure in genovese e nei dialetti locali: profilo storico e antologia*. Recco: Le Mani. 7 vols.

ZenaMT = Haberland, Christopher et al. (2024). "Italian-Ligurian Machine Translation in its Cultural Context", in *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages*, pp. 168-176, ELRA Language Resource Association.

Human and machine authorship: identifying LLM-Generated narratives through stylometric markers

Ferdinando Longobardi & Laura D'amodio & Gaia Valeria Gallo
(Università degli Studi di Napoli L'Orientale)

The researchers' interest in many projects concerning AI systems and language reiterates the much older desire of man to build an intelligent automaton, a thinking, reasoning machine that could replicate human cognitive abilities.

Philosophers, logicians, and linguists have been tracing a tortuous path between dreams towards mechanization, periodically amplified by technological improvements, and concerns with formalization and perfection, which sustained attention to natural languages consistently scales back.

Information technology has rendered such aspirations strikingly tangible, yet it has also generated new, and at times even basilar challenges: for instance, the construction of an intelligent machine is a central objective of artificial intelligence, and natural language constitutes one of the most essential components of human intelligence.

A closely related issue is increasingly raised by contemporary studies regarding generative AI systems and their linguistic identities (Wu, Yang, Zhan, Yuan, Sam Chao, Fai Wong 2025).

Drawing on the hypothesis of "linguistic fingerprints" – an expression, though, not always considered appropriate (Olsson, 2004) – or of "linguistic DNA", insofar as the analytical method is analogous for DNA and language, both studied in terms of sequences of character strings (McMenamin, 2002), the stylometric identification of the author of a text – whether human or artificial – would undergo both quantitative and qualitative methods (Przystalski, Argasiński, Grabska-Gradzińska, Ochab 2026). Statistical methods (Gudjonsson, Haward 1983, Cabras 1996, Chaski 2001), for instance, allow to measure stylistic features.

This paper aims to investigate whether texts generated by large language models (LLMs) display recognizable stylistic patterns, and how they differ from human-made

ones. The analysis is based on a balanced corpus of forty short novels in Italian, twenty of which are written by Italian

novelists active during the second half of the twentieth century, while the remaining half is artificially generated using ChatGPT-5.3 by OpenAI and Claude by Anthropic.

The study's methodology integrates stylometric feature extraction and classification within a Python environment, adopting a Support Vector Machine as a reference model. Specifically, lexical, syntactic, and stylistic metrics are considered, with the aimed result of identifying the features that most significantly contribute to distinguishing between human-authored and artificially generated texts; the evaluation uses a cross-validation technique due to the limited size of the dataset.

Additionally, a pilot study on Bestie by Andrea Damasco, the first Italian novel written with the aid of an AI system, is carried out, so to verify whether the stylometric features observed in the corpus are also detected in a hybrid text.

The previously introduced ambition of many of building an intelligent machine is closely intertwined with the even older, longstanding pursuit of uncovering the mystery of creation and knowledge. Behind the construction of a human-like machine lies the need for a deeper understanding of the human nature, including both its strengths and its flaws, in order to address and provide an improvement for the latter. The linguistic and statistical analysis presented in this study should therefore be considered as an intermediary framework between theoretical aims and practical outcomes, one of the bridges connecting human reasoning and generative AI's reprocessing abilities.

References

Cabras, Cristina. 1996. Analisi del contenuto e stilometria: un metodo per l'esame documentale. In Cabras, Cristina (a cura di) Psicologia della prova, 89-121, Milano: Giuffrè Editore

Chaski, Carole. 2001, Empirical evaluations of language-based author identification techniques. Forensic Linguistics 8 (1). 1-65.

Gudjonsson, Gísli, Haward, Lionel Richard. 1983. Psychological analysis of confession statements. Journal of Forensic Science Society 23. 113-120.

McMenamin, Gerald. 2002. Forensic linguistics: advances in forensic linguistics. Boca Raton: CRC Press.

Olsson, John. 2004. Forensic Linguistics. Londra-New York: Continuum.

Przystalski, Karol, Argasiński, Jan, Grabska-Gradzińska, Iwona, Ochab, Jeremi. 2026. Stylometry recognizes human and LLM-generated texts in short samples. Expert Systems with Applications 296.

Wu, Junchao, Yang, Shu, Zhan, Runzhe, Yuan. Yulin, Sam Chao, Lidia, Fai Wong, Derek. 2025. A survey on LLM-generated text detection: necessity, methods, and future directions. Computational Linguistics. 1-64.