TECHNISCHE UNIVERSITÄT DRESDEN

3rd International Conference

AUTOMATED TEXTS IN THE ROMANCE AND GERMANIC LANGUAGES

4th & 5th
September 2025

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

For registration *(on site or online)*
and more information, please visit
↗https://tud.link/hapvju

Organization: Prof. Dr. Anna-Maria De Cesare,
Michela Gargiulo M.A., & Tom Weidensdorfer M.A.

Layout & Design: Tom Weidensdorfer

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

Center for Italian Studies

DRESDEN concept
SCIENCE AND INNOVATION CAMPUS

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.

TECHNISCHE UNIVERSITÄT DRESDEN

»AI-ROM-III

»Thursday, 4.9.25

Open Science Lab, Room OSL 1

»09.00–09.15	*Conference Opening* **Anna-Maria De Cesare** (Technische Universität Dresden)
»09.15–10.15	*Keynote* **Noah Bubenhofer** (Universität Zürich): *Vectors and Neural Learning: TextgenAI from a Linguistic Perspective*
»10.15–10.30	*Coffee Break*
»10.30–11.00	**Robert Cornelis Schuppe** (Technische Universität Dresden): *"Write a bedtime story for my four year old daughter!" Features and implications of ChatGPT-generated narrations*
»11.00–11.30	**Paolo Valentinelli** (Università di Trento): *Between Human and Artificial Language: Comparing the Syntax of Large Language Models and German Journalistic Texts*
»11.30–12.00	**Franz Meier** (Universität Augsburg): *Attitude and subjectivity in human-written and AI-generated editorials published in* Il Foglio
»12.00–12.30	**Iris Ferrazzo** (Universität Bonn): *Reimagining linguistic data collection: The role of Large Language Models (LLMs) as crowd workers*
»12.30–14.00	*Lunch Break*
»14.00–14.30	**Veronica Mangiaterra, Hamad Al-Azary, Chiara Barattieri di San Pietro & Valentina Bambini** (Istituto Universitario di Studi Superiori IUSS Pavia): *GPT as a rater: systematic evaluation of machine-generated norms for English and Italian metaphors*
»14.30–15.00	**Marie Dewulf** (Universiteit Gent): *Evaluating Syntactic Generalization in Neural Language Models: A Case Study in French*
»15.00–15.30	**Ginevra Martinelli, Chiara Barattieri di San Pietro, Maddalena Bressler, Veronica Mangiaterra & Valentina Bambini** (Istituto Universitario di Studi Superiori IUSS Pavia): *MetaMap – Mapping Metaphors Across Languages and Cultures: the Cases of Love and Anger*
»15.30–16.00	*Coffee Break*
»16.00–17.00	*Keynote* **Rachele Raus** (Alma Mater Studiorum - Università di Bologna): *Quelques réflexions sur les dispositifs d'IA en traduction des langues romanes*

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

Center for Italian Studies

DRESDEN concept
SCIENCE AND INNOVATION CAMPUS

TECHNISCHE UNIVERSITÄT DRESDEN

»AI-ROM-III

»Friday, 5.9.25

Open Science Lab, Room OSL 1

»09.30–10.30	*Keynote* **Francesca Chiusaroli** (Università di Macerata): *Overcoming Language Barriers with an Emoji-Based Pivot Language: Localization Challenges in Automated Translation Experiments Using LLMs*
»10.30–11.00	*Coffee Break*
»11.00–11.30	**Maria Margherita Mattioda & Vincenzo Lambertini** (Università di Torino): *L'oral au prisme de la traduction automatique neuronale et de l'IA générative: retour d'expériences (français et italien) et perspectives didactiques*
»11.30–12.00	**Aurora Trapella** (Università di Torino): *Linguistic features and ethical implications of ChatGPT-generated Italian translations of news: a case study on hate speech*
»12.00–12.30	**Giuliana Fiorentino** (Università degli Studi del Molise): *Integrating AI into everyday document workflows: a pilot study*
»12.30–14.00	*Lunch Break*
»14.00–14.30	**Maria Laura Ferroglio** (Università di Torino): *"I need more practice!". Exploring Italian teachers' perceptions on the use of ChatGPT to support their English lessons: preliminary results from a pilot study*
»14.30–15.00	**Anne-Marie Lachmund** (Universität Potsdam): *Fostering intercultural competences in the FLE-classroom: Materials development with the help of LLM*
»15.00–15.30	**Mariachiara Pascucci** (Università di Pisa/Universität Basel) **& Angela Ferrari** (Universität Basel): *Migliorare la chiarezza dei testi amministrativi con ChatGPT: analisi qualitativa degli interventi sintattici e degli effetti testuali e comunicativi*
»15.30–16.00	**Mirko Tavosanis** (Università di Pisa): *Gli errori grammaticali degli LLM: diversità tra sistemi e caratteristiche generali*
»16.00–16.15	*Closing Remarks*

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

Center for Italian Studies

DRESDEN concept
SCIENCE AND INNOVATION CAMPUS

KEYNOTE

Vectors and Neural Learning: TextgenAI from a Linguistic Perspective

Noah Bubenhofer (Universität Zürich)

What can be said about the possibilities and limitations of generative AI from a linguistic perspective? For the first time, linguistic theories based on language use are being empirically tested on a large scale. Will these ideas, such as the distributional hypothesis, prove themselves? This is certainly the case, but are there other concepts from linguistics that are important for generative AI in order to anticipate its possibilities and limitations? Concepts such as body, practices, and multilingualism are among them, as will be explained in the presentation.

KEYNOTE

Overcoming Language Barriers with an Emoji-Based Pivot Language: Localization Challenges in Automated Translation Experiments Using LLMs.

Francesca Chiusaroli (Università di Macerata)

The talk will focus on experiments in translating verbal languages into emoji, conducted by both humans and AI tools, as part of a broader emoji-based interlingua project (*Emojitaliano*, *Emojilingo*). The project aims to develop an automated emoji-based code to support language simplification, enhance accessibility, and promote international communication.

Using examples from both Italian and English, the presentation will explore the expressive potential of emoji in conveying meaning, resolving semantic ambiguities, and simplifying complex linguistic structures.

Special emphasis will be placed on the challenges of localization - specifically, some limitations of conveying meaning through visual symbols across diverse linguistic and cultural contexts. Preliminary results demonstrate LLMs' remarkable ability to interpret and translate both contemporary and literary vocabulary, while also highlighting key challenges, particularly the role of cultural specificity in emoji interpretation.

Dedicated to the memory of Federico Sangati.

Emojilingo: https://emojilingo.org/

KEYNOTE

Quelques réflexions sur les dispositifs d'IA en traduction des langues romanes

Rachele Raus (Università di Bologna)

L'intégration de l'intelligence artificielle (IA) dans la traduction automatique (TA) via la technologie des réseaux neuronaux – connue sous le nom de traduction automatique neuronale (TAN) – connaît une diffusion massive. Ce développement est aujourd'hui possible grâce à l'introduction des réseaux neuronaux « convolutionnels » (Le Cun 2019), une architecture qui apprend directement à partir des données, et du modèle d'apprentissage automatique *Bidirectional Encoder Representations from Transformers* (BERT), introduit par Google en 2019 (Devlin *et al.* 2019).

La TAN est souvent présentée comme une ressource précieuse pour la société et les institutions (Torres-Hostench 2022). Cependant, cette technologie soulève également de nombreuses questions sur l'implication humaine dans l'apprentissage automatique, notamment si les systèmes de TAN doivent être entraînés de manière supervisée (Cerquitelli, Raus, Molino en cours de presse) et si la qualité des traductions doit être évaluée par des méthodes informatiques et/ou par l'humain (voir aussi Langlais 2023). La prolifération récente des grands modèles de langage (LLM), encouragée par le succès de ChatGPT et d'autres outils d'IA générative, accentue l'urgence de ces questions, car ces modèles ne nécessitent pas vraiment de supervision humaine et sont devenus la norme pour la génération de texte.

Des préoccupations ont également été soulevées concernant l'adoption de l'anglais comme langue pivot pour l'apprentissage profond (Moorkens 2022; Vetere 2023), étant donné qu'environ 93 % de la formation de ChatGPT-3 a été réalisée en anglais (Brown *et al.* 2020:14). De plus, les implications sociales et culturelles des mégadonnées (*big data*) ont fait l'objet de nombreux débats (Sheng *et al.* 2021; Ghosh, Caliskan 2023, etc.), ainsi que la disponibilité et/ou les limites des grands corpus multilingues (Raus, Tonti 2025). L'apprentissage automatique pourrait conduire à « une érosion de la diversité linguistique [car] ces outils ont tendance à (…) homogénéiser les expressions »

(Larsonneur 2021) ou pourrait propager des biais socioculturels en raison de données d'entrée biaisées (European Human Agency for Fundamental Rights 2022).

Dans ce contexte, nous aimerions discuter des implications de la TAN pour la diversité linguistique des langues romanes, en donnant des exemples variés, tirés de l'italien et du français.

À partir des réflexions menées, nous proposerons des pistes de recherche possibles, notamment:

1. L'introduction de nouveaux observables qui permettent d'encadrer les faits linguistiques par rapport aux intelligences artificielles qui les produisent (Humbley, Zollo 2022 ; De Cesare en cours de publication);
2. La nécessité d'utiliser les dispositifs d'IA de manière suffisamment critique (Raus 2023, De Cesare en cours de publication);
3. Le rôle que la supervision de l'humain peut encore jouer pour éviter certains problèmes d'ordre linguistique (Crosthwaite, Baisa 2023).

References

Brown T., Mann B., Ryder N. Subbiah, M. Kaplan J. D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A. *et al.* (2020), "Language models are few-shot learners", *Advances in Neural Information Processing Systems*, n° 33. https://arxiv.org/pdf/2005.14165

Cerquitelli T., Raus R., Molino A. (en cours de presse), "Artificial Intelligence and Neural Machine Translation", dans S. Baumgartner, M. Tieber (eds.), *Handbook of Translation Technology and Society*, New York / Londres : Routledge.

Crosthwaite P., Baisa V. (2023), "Generative AI and the end of corpus-assisted data-driven learning? Not so fast!", *Applied Corpus Linguistics*, n.3 (3), https://www.sciencedirect.com/science/article/pii/S2666799123000266

European Human Agency for Fundamental Rights (2022), *Bias in Algorithms. Artificial Intelligence and Discrimination*. https://fra.europa.eu/en/publication/2022/bias-algorithm#publication-tab-0

De Cesare, A.-M., "L'intelligenza artificiale generativa al servizio della parità di genere: uno studio esplorativo sugli annunci di lavoro

della Confederazione svizzera", dans AA. VV., *Inclusione ed elaborazione del linguaggio naturale nell'era dell'intelligenza artificiale generativa*, Milan: Ledizioni, 59-84.

Devlin J., Chang M.-W., Lee K., Toutanova K. (2019), "BERT: Pre-training of deep bidirectional transformers for language understanding", dans *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis.

Gosh S., Caliskan A. (2023), "ChatGPT perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across Bengali and five others low resources languages", Conference on AI, Ethics, and Society 2023, https://arxiv.org/pdf/2305.10510

Humbley J., Zollo S. D. (2021) "Réflexions et études de cas à l'aune de l'intelligence artificielle. Vers de nouveaux observables linguistiques ?" dans R. Raus, A. M. Silletti, S. D. Zollo, J. Humbley (eds), *Multilinguisme et variétés linguistiques en Europe à l'aune de l'intelligence artificielle*, Milan: Ledizioni, 35-44. https://www.collane.unito.it/oa/items/show/132#?c=0&m=0&s=0&cv=0

Langlais P. (2023) "For a common European framework for evaluating AI based translation technologies", dans R. Raus (éd.), *How Artificial Intelligence Can Further European Multilingualism,* Milan: Ledizioni, 93-96.

Larsonneur C. (2021), "Intelligence artificielle ET/OU diversit. linguistique: les paradoxes du traitement automatique des langues", *Hybrid*, n°7, https://journals.openedition.org/hybrid/650

Le Cun Y. (2019), *Quand la machine apprend: la révolution des neurones artificiels et de l'apprentissage profond*, Paris: Éditions Odile Jacob.

Moorkens, J. (2022), "Ethics and machine translation", dans D. Kenny (éd.) *Machine Translation for Everyone. Empowering Users in the Age of Artificial Intelligence.* Berlin: Language Science Press, 121-140.

Raus R. (ed.) (2023), *How Artificial Intelligence Can Further European Multilingualism,* Milan: Ledizioni.

Raus R., Tonti M. (eds), "Intelligence artificielle, corpus et diversité linguistique: enjeux et perspectives. Introduction". *Langages* n°23, 7-20.

Sheng E., Chang K.-W., Natarajan P., Peng, N. (2021), "Societal biases in language generation: Progress and challenges", dans *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural*

Language Processing, https://aclanthology.org/2021.acl-long.330.pdf

Torres-Hostench, O. (2022), "Europe, multilingualism and machine translation", dans D. Kenny (éd) *Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence.* Berlin: Language Science Press, 1-21.

Vetere G. (2022), "Elaborazione automatica dei linguaggi diversi dall'inglese: introduzione, stato dell'arte e prospettive, dans R. Raus, A. M. Silletti, S. D. Zollo, J. Humbley (eds), *Multilinguisme et variétés linguistiques en Europe à l'aune de l'intelligence artificielle*, Milan: Ledizioni, 69-87. https://www.collane.unito.it/oa/items/show/132#?c=0&m=0&s=0&cv=0Introduction/Introduzione/Introduction

Evaluating Syntactic Generalization in Neural Language Models: A Case Study in French

Marie Dewulf (Universiteit Gent)

In computational linguistics, recent advances in multilingual and monolingual large language models (LLMs) offer new opportunities for investigating how morphological and syntactic knowledge is represented and learned. Morphologically rich languages such as French present a compelling case study for testing generalization, i.e. the ability to generalize over structured grammatical rules such as agreement. This study adapts targeted syntactic evaluation techniques to assess phenomena in French, building on methodologies from Linzen et al. (2016), Marvin and Linzen (2018), Hu et al. (2020), and Pérez-Mayos et al. (2021).

This research aims to answer the following questions: (1) To what extent do multilingual and French-specific language models acquire French syntactic representations in an Englishdominant training context? (2) How consistent are these representations across syntactic features? (3) Can surprisal-based predictions uncover representational deficits not captured by perplexity alone?

Language models are commonly evaluated by their *perplexity*, which measures how well it predicts words in context. As Marvin and Linzen (2018) note, this metric conflates various factors contributing to next-word prediction, including collocations, semantics, pragmatics and syntax (Marvin & Linzen 2018). The grammaticality of the predictions of a language model can be more precisely assessed on the model's ability to make human-like generalizations for specific syntactic phenomena. Targeted syntactic evaluation (Marvin & Linzen 2018) involves constructing minimal pairs of grammatical and ungrammatical sentences targeting specific syntactic features. For example, to test subject-verb agreement, verbs are varied to match or mismatch with the subject in person and/or number (e.g., *Tu travailles* vs. **Tu travaillez*). Each test item appears in controlled conditions that isolate syntactic dependencies.

We use the measure of surprisal to infer the LLMs' syntactic knowledge of French. Surprisal quantifies how unexpected a word is in its context, based on the model's prediction. Lower surprisal

indicates higher model expectation for that word; higher surprisal signals lower expectation. The syntactic sensitivity of the model is quantified by the extent to which predicted surprisal differences match theoretical expectations: grammatical continuations should be less surprising than ungrammatical ones. Preliminary results suggest that LLMs often capture basic agreement patterns in French but falter on more complex constructions, especially those requiring long-distance dependencies or interaction of multiple morphosyntactic features. Varying results in models trained on different corpora sizes raise questions on the impact of data scale and typological distance between English and French.

This approach contributes to linguistic research by adapting a principled framework for probing grammatical knowledge in LLMs. This framework enables model comparison across architectures and training regimes and offers insights into the developmental trajectory of non-English syntactic competence in models with constrained exposure to other languages. Furthermore, the test suite design is extensible to other more typologically diverse languages, supporting broader efforts in cross-linguistic evaluation.

References

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. (2020). A Systematic Assessment of Syntactic Generalization in Neural Language Models. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1725–1744). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.158

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4, 521–535. https://doi.org/10.1162/tacl_a_00115

Marvin, R., & Linzen, T. (2018). Targeted Syntactic Evaluation of Language Models. In E.Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 1192–1202). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1151

Pérez-Mayos, L., Táboas García, A., Mille, S., & Wanner, L. (2021). Assessing the Syntactic Capabilities of Transformer-based

Multilingual Language Models. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 3799–3812). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.333

Reimagining linguistic data collection: The role of Large Language Models (LLMs) as crowd workers

Iris Ferrazzo (Universität Bonn)

Data elicitation from human participants plays a central role in both NLP and empirical linguistics. Such studies range in scale from small, controlled participant groups to extensive crowdsourcing approaches via platforms like Prolific[1] or Amazon Mechanical Turk[2]. Yet, despite their reach, these methods come with challenges, including limited control over participant attention, ethical concerns about remuneration (Van Zoonen, 2024), and time-consuming experimental designs.

This contribution explores whether Large Language Models (LLMs) can help mitigate these issues and serve as proxies for human crowd workers in empirical linguistic research. Recent work in NLP has shown that LLMs can match or exceed human performance in many tasks (Törnberg, 2023; He et al., 2023), and that human crowd workers are relying on LLMs to improve their practice in crowdsourcing studies, often without disclosure (Veselovsky et al., 2023). To examine whether these trends apply to linguistics, we replicate two prior studies using OpenAI's models GPT-4o-mini, a smaller version of GPT-4o, and o4-mini, a reasoning model, as crowd workers: Cruz (2023) on gender assignment in Spanish–English code-switched speech, and Lakhzoum et al. (2021) on semantic similarity ratings of French word pairs. Since LLMs, like humans, are sensitive to prompt wording and response format, often displaying systematic biases (Tjuatja, 2024; Lu et al., 2025), we reproduce two tasks that differ in both linguistic phenomena (bilingual morphosyntax vs. lexical semantics) and response types (binary answer generation vs. 7-point Likert scale) to assess model behavior across multiple linguistic and task dimensions. We evaluate model performance through a data elicitation pipeline that mirrors the original study setups, benchmarking model responses against those of human participants. Cruz (2023) is replicated using zero-shot prompting, while Lakhzoum et al. (2021) is replicated using few-shot prompting conditions. One goal of this contribution is to provide an open-source framework that can

[1] https://www.prolific.com/
[2] https://www.mturk.com/

serve as an introductory guide for prompting LLMs in a Python coding environment.

Results show that GPT-4o-mini outperforms both o4-mini (in 81% of tested conditions) and human participants (in 87.5%) in terms of expected gender assignment congruency in Cruz's (2023) replication. However, GPT-4o-mini's apparent "over-performance" raises important questions, given that the task concerns naturally occurring data used to study a linguistic phenomenon, rather than eliciting objectively right or wrong answers. In Lakhzoum et al. (2021)'s replication, both models achieve similar performance, supporting the view that few-shot prompting helps stabilize LLMs' performance under varied scoring formats, countering the claims of instability in previous works (cf. Tsvilodub et al. 2024). While the differences between model and human similarity ratings are not insignificant (fewer than 10% are exact matches, 60% fall within a 1-point difference, and around 30% within 2 points), they reflect a general alignment between model and human judgments. Follow-up experiments will further explore the extent of this model-human alignment in semantic similarity ratings.

Our findings suggest that, with careful application, LLMs may offer viable support for data elicitation in empirical linguistics. However, ethical considerations, the need for additional research, and the importance of human-in-the-loop pipelines remain paramount.

References

Cruz, A. (2023). Linguistic factors modulating gender assignment in Spanish–English bilingual speech. *Bilingualism: Language and Cognition*, *26*(3), 580–591. https://doi.org/10.1017/S1366728922000839

He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., Jiao, J., Yiu, S. M., Duan, N., & Chen, W. (2023). *AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2303.16854

Lakhzoum, D., Izaute, M., & Ferrand, L. (2021). Semantic similarity and associated abstractness norms for 630 French word pairs. *Behav Res*, *53*, 1166–1178. https://doi.org/10.3758/s13428-020-01488-z

Lu, Y.-L., Zhang, C., & Wang, W. (2025). *Systematic Bias in Large Language Models: Discrepant Response Patterns in Binary vs. Continuous*

Judgment Tasks (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2504.19445

Tjuatja, L., Chen, V., Wu, T., Talwalkwar, A., & Neubig, G. (2024). LLMs Exhibit Human-like Response Biases? A Case Study in Survey Design. *Transactions of the Association for Computational Linguistics*, *12*, 1011–1026.

Törnberg, P. (2023). *ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2304.06588

Tsvilodub, P., Wang, H., Grosch, S., & Franke, M. (2024). *Predictions from language models for multiple-choice tasks are not robust under variation of scoring methods* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2403.00998

Van Zoonen, W., Sivunen, A. E., & Treem, J. W. (2024). Algorithmic management of crowdworkers: Implications for workers' identity, belonging, and meaningfulness of work. *Computers in Human Behavior*, *152*, 108089. https://doi.org/10.1016/j.chb.2023.108089

Veselovsky, V., Ribeiro, M. H., & West, R. (2023). *Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2306.07899

"I need more practice!". Exploring Italian teachers' perceptions on the use of ChatGPT to support their English lessons: preliminary results from a pilot study

Maria Laura Ferroglio (Università di Torino)

The appearance of Generative AI (GenAI) in education and teaching has attracted the interest of the research community, and publications have soared in the past years (Crompton et al., 2024; Law, 2024). In investigating its potential, researchers have identified possibilities ranging from the reduction of teachers' workload to an increase in learner autonomy (Amonova et al., 2023; Barrot, 2024). These opportunities are counterbalanced by issues such as overreliance and content bias (Barrot, 2024; Binu, 2024; Cong-Lem et al., 2024; Ghafouri et al., 2024), and one aspect that has been repeatedly highlighted in current research is how training and professional development promote teachers' eNective and responsible integration of GenAI in their practice (Allehyani & Algamdi, 2023; Crompton et al., 2024; Cong-Lem et al., 2024). While studies have theorised potential use of AI in diNerent educational contexts (Dehghani & Mashhadi, 2024), limited empirical research has explored its hands-on use by English Language teachers to create teaching materials or plan lessons (Dornburg & Davin, 2024; Evmenova et al., 2024; Ghafouri et al., 2024).

As it often occurs when new technologies are introduced in education, there seems to be an expectation that tools will be seamlessly integrated once practitioners are trained to master them, with little consideration for educators' existing subject knowledge and educational expertise (Mishra & Koehler, 2006); Generative AI may be no exception. This is why a useful background model to ground Generative AI usage in is the TPACK framework (Shulman, 1986, Mishra & Koehler, 2006), which seeks to integrate Technological, Pedagogical and Content Knowledge by focusing on how teachers can use technology to convey content information and find alternative approaches to enhance learning outcomes. The present work is grounded in this perspective.

This contribution summarises the author's PhD research project in its current phase, where a scoping literature review and a pilot study have been conducted. The project aims at exploring

what Italian secondary school teachers of English know and understand about ChatGPT, and how they develop further knowledge using the GenAI tool to design materials for developing writing skills. More specifically, after highlighting empirical evidence from a scoping review, findings from the pilot study will be framed within the current English Language Teaching educational scenario, oNering insights into a responsible, technologyintegrated teaching practice.

References

Allehyani, S. H., & Algamdi, M. A. (2023). Digital competences: Early childhood teachers' beliefs and perceptions of ChatGPT application in teaching English as a second language

(ESL). *International Journal of Learning, Teaching and Educational Research, 22*(11), 343– 363. https://doi.org/10.26803/ijlter.22.11.18

Amonova, S., Juraeva, G., & Khidoyatov, M. (2023). Harnessing the potential of artificial intelligence in language learning: Is AI threat or opportunity? In *Proceedings of the ACM International Conference* (pp. 292–297).

Barrot, J. S. (2024). ChatGPT as a language learning tool: An emerging technology report. *Technology, Knowledge and Learning, 29*(2), 1151–1156.

Binu, P. M. (2024). ANordances and challenges of integrating artificial intelligence into English language education: A critical analysis. *English Scholarship Beyond Borders, 10*(1), 34–51.

Cong-Lem, N., Tran, T. N., & Nguyen, T. T. (2024). Academic integrity in the age of generative AI: Perceptions and responses of Vietnamese EFL teachers. *Teaching English with Technology, 24*(1), 28–47.

Crompton, H., Edmett, A., Ichaporia, N., & Burke, D. (2024). AI and English language teaching: ANordances and challenges. *British Journal of Educational Technology, 55*(6), 2503–2529. https://doi.org/10.1111/bjet.13489

Dehghani, H., & Mashhadi, A. (2024). Exploring Iranian English as a foreign language teachers' acceptance of ChatGPT in English language teaching: Extending the technology acceptance model. *Education and Information Technologies, 29*(15), 19813–19834. https://doi.org/10.1007/s10639-024-12345-6 *(placeholder DOI—please check your source)*

Dornburg, A., & Davin, K. J. (2024). ChatGPT in foreign language lesson plan creation:

Trends, variability, and historical biases. *ReCALL*. Advance online publication. https://doi.org/10.1017/S0958344024000035

Evmenova, A. S., Borup, J., & Shin, J. K. (2024). Harnessing the power of generative AI to support all learners. *TechTrends, 68*(4), 820–831. https://doi.org/10.1007/s11528-02400940-0

Ghafouri, M., Hassaskhah, J., & Mahdavi-Zafarghandi, A. (2024). From virtual assistant to writing mentor: Exploring the impact of a ChatGPT-based writing instruction protocol on EFL teachers' self-eNicacy and learners' writing skill. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/13621688241229883

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open, 5*, 100174. https://doi.org/10.1016/j.caeo.2024.100174

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14. https://doi.org/10.3102/0013189X015002004

Integrating AI into everyday document workflows: a pilot study

Giuliana Fiorentino (Università degli Studi del Molise)

Over the past three years, I have led the development of *SEMPL-IT*, an AI-driven tool designed to simplify and modernize institutional Italian through a chain-of-prompt approach. The project emerges from a longstanding body of linguistic research highlighting the structural and lexical complexity of administrative Italian—a variety often inaccessible to citizens unfamiliar with bureaucratic registers or lacking the necessary linguistic competence. Foundational studies (e.g., Cortelazzo & Pellegrino 2003; Piemontese 2023; Fiorentino & Ganfi 2024) have shown how this complexity hinders transparency and inclusivity in public communication.

Building on recent international work validating the ability of Large Language Models (LLMs) to enhance text readability (Feng et al. 2023; Guo et al. 2023; North et al. 2023), and incorporating Italian-specific insights (Tavosanis 2018, 2019), SEMPL-IT applies advanced Natural Language Processing techniques to perform syntactic, lexical, and textual simplification. The system was trained on a newly created corpus: *ItaIst*, a 208-document dataset of authentic Italian institutional texts, and *ItaIst-Sempl*, its simplified counterpart generated via GPT-3.5/4.0. These datasets were used to fine-tune several LLMs (mT5, umT5, GPT-2 ITA).

This presentation introduces the finalized SEMPL-IT software, its user-friendly interface, and the results of a validation process. In fact, I will report on an ongoing pilot involving public sector employees, who are now integrating SEMPL-IT into their everyday document workflows. Their usage data and feedback offer a unique lens through which to assess the practical impact of AI-assisted linguistic simplification in the public administration.

Through this contribution, I aim to demonstrate the feasibility and social value of integrating domain-specific LLMs into real-world governance contexts, where clarity and accessibility are central to democratic participation.

References

Cortelazzo, M.A., Pellegrino F. (2003). *Guida alla scrittura istituzionale*. Roma-Bari: Laterza.

Feng, Y., Qiang, J., Li, Y., Yuan, Y., & Zhu, Y. (2023). Sentence simplification via large language models. arXiv preprint arXiv:2302.11957.

Fiorentino, G., Ganfi, V. (2024) "Parametri per semplificare l'italiano istituzionale: revisione della letteratura". *Italiano LinguaDue* 16,1. 220-237.

Guo, B. et al. "How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection", arXiv e-prints. doi:10.48550/arXiv.2301.07597.

North, K., Ranasinghe, T., Shardlow, M., & Zampieri, M. (2023). Deep Learning Approaches to Lexical Simplification: A Survey. arXiv preprint arXiv:2305.12000.

Piemontese, M. E. (2023). È ancora «fatica gittata osar d'ingentilire» la lingua delle nostre leggi e della nostra burocrazia? In: Piemontese (a cura di): 19-36.

Shardlow, M. (2014). A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1), 58-70.

Tavosanis, M. (2018). *Lingue e intelligenza artificiale.* Roma: Carocci.

Tavosanis, M. (2019). *Valutazione umana di Google Traduttore e DeepL per le traduzioni di testi giornalistici dall'inglese verso l'italiano*. In *CLiCit 2019 – Proceedings of the Sixth Italian Conference on Computational Linguistics*, a cura di Raffaella Bernardi, Roberto Navigli e Giovanni Semeraro, CEUR Workshop Proceedings, Aachen University, pp. 1-7.

Fostering intercultural competences in the FLE-classroom: Materials development with the help of LLM

Anne-Marie Lachmund (Universität Potsdam)

One of the main objectives of foreign language teaching is to promote intercultural and transcultural competences (KMK 2023, 7). As important as the demand for the consistent integration of this transversal competence is, the learning objectives, levels and competence standards to be achieved in the foreign language classroom are vague (Byram 2021). As a result, there is often a lack of adequate teaching materials that reflect the current convictions and discussions surrounding culturally sensitive foreign language teaching, facing the danger of an outdated, stereotyped, homogeneous reproduction of "culture" (Holliday et al. 2010; Udah 2019). The question now is whether AI models for creating teaching material are able to fill a gap and support teachers in providing good material for use (Tomlinson 2011), or whether they achieve less useful results.

The contribution presents the results of a didactic master's seminar for prospective teachers who used various LLMs to create teaching materials to promote intercultural and transcultural skills for students of different years of learning French as a foreign language in Germany. The programmes used were among others to-teach.ai, magicschool.ai, eduaide.ai and chatgpt.com. The prompts were created in the context of the seminar and were adapted to different learning years (from first to third year of learning *FLE*). The AI-generated outcome was analyzed and evaluated with the help of a self-created criteria grid (based on current cultural theories relevant for foreign language learning in a German context, inspired by Usener 2016; Tomlinson 2011). The presentation will focus on the AI-generated outcome, which should be appropriate in terms of language and content.

First preliminary results demonstrate the variety of tasks and activities suggested by the LLM beyond QA-formats and prove a certain amount of creativity and learner orientation. The situations created had an intercultural background and focused on everyday culture with a strong accent on cultural differences.

Regarding the language used we can see e.g. incoherent language switches (here French & German) and that the adaption of the language learning level does not always correspond to the anticipated level of the French learners in Germany, i.e., using inversion questions in an A1-text. The analysis does not only shed a light on the quality of AI-generated teaching materials but moreover it trains future teachers to evaluate the outcome critically and to adapt it according to their learners' needs (Hockly 2023).

References

Byram, M. (2021). *Teaching and Assessing Intercultural Communicative Competence: Revisited.* Multilingual Matters. https://doi.org/10.21832/9781800410251

Hockly, N. (2023). Artificial Intelligence in English Language Teaching: The Good, the Bad and the Ugly. *RELC Journal, 54*(2), 445-451. https://doi.org/10.1177/00336882231168504

Holliday, A., Hyde, M., & Kullman, J. (2010). *Intercultural communication: An advanced resource book for students*. Routledge.

KMK – Ständige Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (Ed.) (2023). *Bildungsstandards für die erste Fremdsprache (Englisch/ Französisch) für den Ersten Schulabschluss und den Mittleren Schulabschluss.*

Beschluss der Kultusministerkonferenz vom 15.10.2004 und vom 04.12.2003 i. d. F. vom 22.06.2023. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2023/2023_06_22-Bista-ESA-MSA-ErsteFremdsprache.pdf

Tomlinson, B. (2011). „Introduction: principles and procedures of materials development", in B. Tomlinson (Ed.): *Materials development in language teaching.* Cambridge University Press, 1-31.

Udah, H. (2019). Searching for a place to belong in a time of othering. *Social Science, 8*(11), 1-16. https://doi.org/10.3390/socsci8110297

Usener, J. (2016). *Lehrwerke und interkulturelle Kompetenz im Spanischunterricht. Analyse und Perspektiven*. Diss. https://nbn-resolving.org/urn:nbn:de:gbv:3:4-16923; http://d-nb.info/1116950545/34

GPT as a rater: systematic evaluation of machine-generated norms for English and Italian metaphors

Veronica Mangiaterra, Hamad Al-Azary, Chiara Barattieri di San Pietro & Valentina Bambini (Istituto Universitario di Studi Superiori IUSS Pavia)

Behavioral and electrophysiological studies on metaphors highlighted the deep impact of metaphor features, such as familiarity or aptness, on their processing [1-2]. To account for this variability, a fundamental step in current research is to collect human ratings on the psycholinguistic properties of linguistic items to include in the analysis. This task is resource-consuming and attempts to make it more efficient spanned from crowdsourcing [3] to computational augmentation of existing datasets [4].

With the advent of Large Language Models (LLMs), many linguistic annotation tasks have been delegated to these models [5], enabling the construction of large, annotated resources. This line of research has also extended to psycholinguistic norms, with a growing interest in augmenting human-rated datasets with ratings from LLMs [6-7]. Yet, systematic evaluations of these approaches for complex items beyond single words, i.e., metaphors, and for language other than English are needed.

Here, we assessed the validity and test-retest reliability of machine-generated ratings of metaphor familiarity and comprehensibility. To elicit ratings, we prompted three GPT models with the same instructions provided to human participants in previous studies on metaphor processing in English and Italian [8-14]. To test the validity of GPT-generated ratings, we measured 1) their correlation with human-generated ratings and 2) their ability to predict behavioral (reaction times – RTs) and neurophysiological responses (N400 component). To test the reliability of GPT ratings, we compared the output obtained with the same model in two separate sessions.

Machine-generated ratings showed moderate to strong correlations with human-generated ratings (range for ratings generated by GPT4o: 0.59-0.80). Similar levels of correlation are reported, for ratings generated by GPT4o and GPT4o-mini, for English and Italian metaphors, while for Italian metaphors GPT3.5-generated ratings show lower correlation with human ratings

compared to English ones. Interestingly, body-related metaphors and metaphors referring to physical characteristics showed lower alignment with human judgments (range for ratings generated by GPT4o: 0.54-0.64) compared to object-related and mental metaphors (range for ratings generated by GPT40: 0.69-0.74). Also, like human-generated ratings, machine-generated familiarity predicted RTs (p <.001), with higher familiarity associated with shorter RTs, and N400 responses in centro-parietal electrodes (p <.05), with higher familiarity associated with a reduced negativity. Finally, the correlations between GPT-generated ratings elicited in different sessions ranged from 0.89 to 0.99. Larger models (GPT4o-mini, GPT4o) showed both higher validity and reliability compared to smaller ones (GPT3.5-turbo).

Our results indicate that ratings obtained from GPT can achieve high to very high validity and excellent reliability both for English and Italian metaphors, supporting the use of machine-generated ratings in metaphor research to model behavioral and neurophysiological effects. Also, our results indirectly speak about the capacity of LLMs to process metaphorical expressions: the large amount of linguistic input that the models received during training might partly support the processes needed to understand metaphors, with reduced adherence to human behavior when the stimuli require the integration of multimodal aspects of meaning. Overall, these data can contribute to setting guidelines to better navigate the integration of AI tools in psycholinguistic research in this rapidly evolving phase.

References

1. Lai, V. T., Curran, T., & Menn, L. (2009). Comprehending conventional and novel metaphors: An ERP study. *Brain research*, *1284*, 145-155.
2. Mcquire, M., Mccollum, L., & Chatterjee, A. (2017). Aptness and beauty in metaphor. *Language and Cognition*, *9*(2), 316-331.
3. Keuleers, E., & Balota, D. A. (2015). Megastudies, crowdsourcing, and large datasets in psycholinguistics: An overview of recent developments. *Quarterly Journal of Experimental Psychology*, *68*(8), 1457-1468.
4. Ljubešić, N., Fišer, D., & Peti-Stantić, A. (2018). Predicting concreteness and imageability of words within and across languages via word embeddings. *arXiv preprint arXiv:1807.02903*.

5. Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), e2305016120.
6. Trott, S. (2024). Can large language models help augment English psycholinguistic datasets?. *Behavior Research Methods*, 1-19.
7. Brysbaert, M., Martínez, G., & Reviriego, P. (2025). Moving beyond word frequency based on tally counting: AI-generated familiarity estimates of words and phrases are an interesting additional index of language knowledge. *Behavior Research Methods*, 57(1), 1-15.
8. Al-Azary, H., & Buchanan, L. (2017). Novel metaphor comprehension: Semantic neighbourhood density interacts with concreteness. *Memory & Cognition*, *45*, 296-307.
9. Campbell, S. J., & Raney, G. E. (2016). A 25-year replication of Katz et al.'s (1988) metaphor norms. *Behavior research methods*, *48*, 330-340.
10. Cardillo, E. R., Watson, C., & Chatterjee, A. (2017). Stimulus needs are a moving target: 240 additional matched literal and metaphorical sentences for testing neural hypotheses about metaphor. *Behavior research methods*, *49*, 471-483.
11. Bambini, V., Ghio, M., Moro, A., & Schumacher, P. B. (2013). Differentiating among pragmatic uses of words through timed sensicality judgments. *Frontiers in psychology*, *4*, 938.
12. Bambini, V., Resta, D., & Grimaldi, M. (2014). A dataset of metaphors from the Italian literature: Exploring psycholinguistic variables and the role of context. *PloS one*, 9(9), e105634.
13. Bambini, V., Ranieri, G., Bischetti, L., Scalingi, B., Bertini, C., Ricci, I., Schaeken, W., & Canal, P. (2024). The costs of multimodal metaphors: comparing ERPs to figurative expressions in verbal and verbo-pictorial formats. *Discourse Processes*, *61*(1-2), 44-68.
14. Canal, P., Bischetti, L., Bertini, C., Ricci, I., Lecce, S., & Bambini, V. (2022). N400 differences between physical and mental metaphors: The role of Theories of Mind. *Brain and Cognition*, *161*, 105879

MetaMap – Mapping Metaphors Across Languages and Cultures: the Cases of Love and Anger

Ginevra Martinelli, Chiara Barattieri di San Pietro, Maddalena Bressler, Veronica Mangiaterra & Valentina Bambini (Istituto Universitario di Studi Superiori IUSS Pavia)

Is love a journey or a fire? Metaphors are core to human cognition, contributing to our understanding of new experiences through familiar ones [1]. However, cultural factors shape mappings among concepts, leading to a lack of a one-to-one correspondence between metaphorical target (e.g., LOVE) and source (e.g., JOURNEY) domains across languages [2,3]. Most studies have examined metaphor variation within a few conceptual domains and limited language sets [4,5], leaving large-scale, cross-linguistic analyses underexplored. Recently, Large Language Models (LLMs), such as transformer-based models like GPT [6], have demonstrated state-of-the-art performances in figurative language detection and processing [7], enabling to perform quantitative studies on large multilingual corpora.

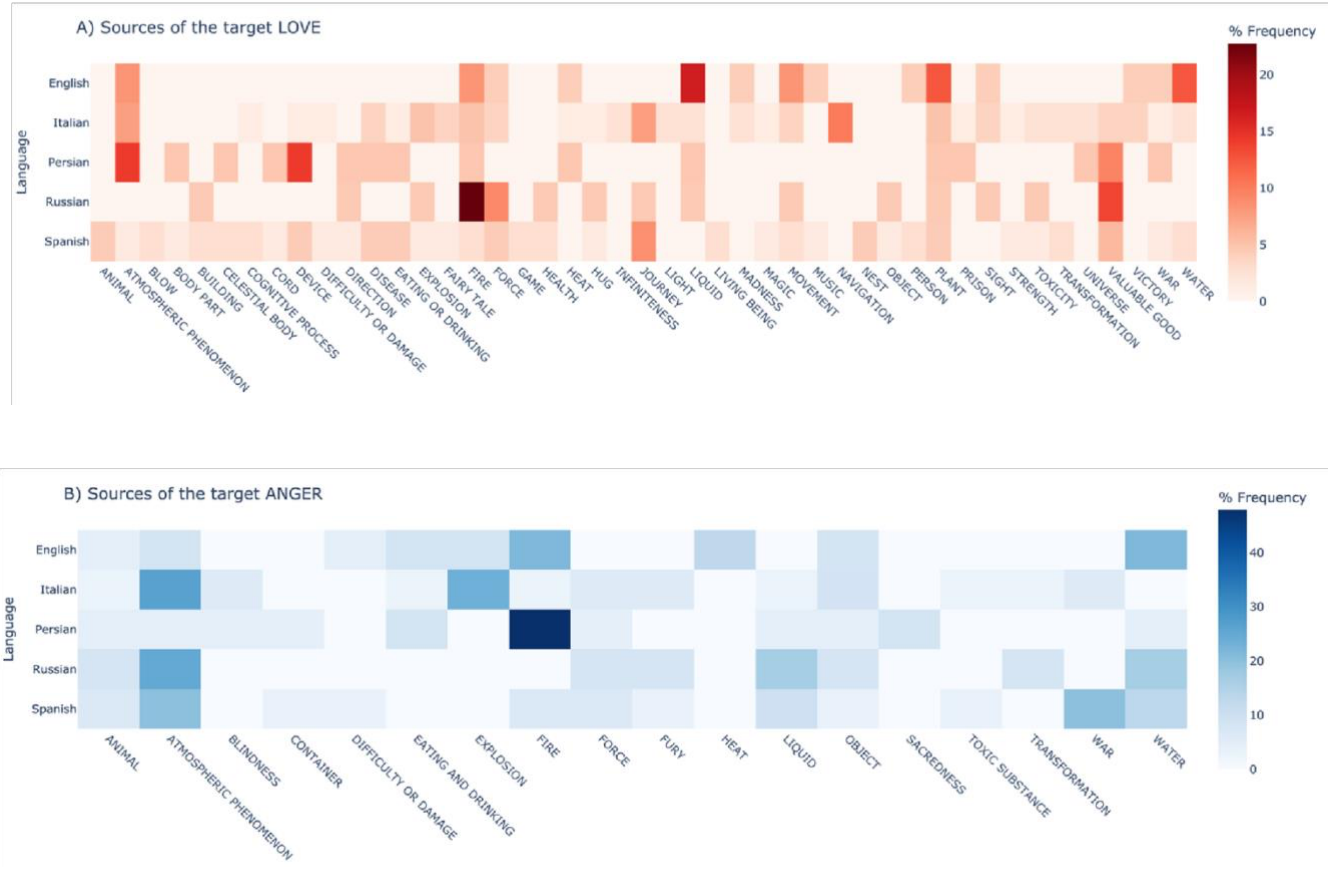
The “MetaMap” project leverages LLMs to build a "map of metaphor mappings” by processing large corpora for cross-cultural investigations of metaphorical language. Here, we present the preliminary results for two emotion concepts, i.e., LOVE and ANGER, as extracted from five IndoEuropean languages (English, Italian, Spanish, Persian, and Russian).

We selected corpora of journalistic and web texts comprising 300,000 sentences per language from the Leipzig Corpora Collection [8]. Sentences containing lemmas for the concepts LOVE and ANGER (and their derivatives) were processed using the GPT-4o Batch API to detect metaphors and identify their source and target domains. The pipeline was validated on a low-resource language (i.e., Latin), with accuracy = 0.77, precision = 0.85, and recall = 0.86.

Then, we calculated an index of cross-linguistic validity for each source-target mapping using the coefficient of variation (CV) [9], which we defined as the logarithm of the standard deviation of the source–target mapping frequency divided by the frequency mean. CV values can approximate 0, indicating great cross-linguistic stability, while higher values are positively related to

higher variability, with the maximum depending on the mappings' frequencies. Additionally, we measured the sparsity of each language's source domain distribution for the given targets using normalized entropy [10]. Entropy values close to 1 indicate a diverse and balanced use of sources, whereas values near 0 reflect reliance on a few dominant ones.

Figure 1. Heatmaps of the relative frequency of each source domain in the five analyzed languages for the target domains LOVE (A) and ANGER (B).



FIRE emerged as a cross-linguistically stable source domain for LOVE (CV = 0.70; "*I've since rekindled my love for country music*"), while ANGER was primarily conceptualized as an ATMOSPHERIC PHENOMENON (CV = 0.58; "*Schmidt temió que la ira se desatara también contra él*"). Other source domains showed greater instability and appeared primarily language-specific — for example, in the case of the NAVIGATION as a mapping for LOVE ("*La loro storia continua a gonfie vele*"; see Figure 1A) or HEAT as a source for ANGER ("*Senior was hot with anger*"; see Figure 1B).

Regarding entropy, all languages displayed high values (above 0.9), suggesting a tendency to use a wide range of source

domains in a distributed manner — with many sources appearing only once. Entropy falls below 0.9 only in Persian and Italian for the concept LOVE, indicating a slightly stronger reliance on a smaller set of dominant mappings.

These preliminary results indicate that: i) LLMs are effective tools for the automatic analysis of metaphors in texts, enabling large-scale, cross-linguistic studies of figurative language; ii) there seems to be universal tendencies in creating metaphors for emotions such as LOVE and ANGER, yet a large intra-linguistic variability also emerges, highlighting the relevance of further research into how metaphors lie at the intersection of nature and nurture.

References

1. Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.
2. Wierzbicka, A. (2007). The semantics of metaphor and parable: Looking for meaning in the Gospels. *Theoria et Historia Scientiarum, 6*(1), 85–106.
3. Schäffner, C. (2004). Metaphor and translation: Some implications of a cognitive approach. *Journal of Pragmatics,* 36, 1253–1269.
4. Anderson, W., Bramwell, E., & Hough, C. (2016). *Mapping English metaphor through time*.
   Oxford University Press.
5. Petruck, M. R. L. (2018). *Metanet*. John Benjamins Publishing Company.
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.
7. Tong, X., Choenni, R., Lewis, M., & Shutova, E. (2024). Metaphor understanding challenge dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (pp. 3517–3536). Association for Computational Linguistics.
8. Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association.

9. Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The wordbank project*. MIT Press. p.283.
10. Kulkarni, R., Rothstein, S., & Treves, A. (2013). A statistical investigation into the crosslinguistic distribution of mass and count nouns: morphosyntactic and semantic perspectives. *Biolinguistics*, 7, 132-168.

L'oral au prisme de la traduction automatique neuronale et de l'IA générative: retour d'expériences (français et italien) et perspectives didactiques

Maria Margherita Mattioda & Vincenzo Lambertini (Università di Torino)

Cette étude vise à analyser le traitement de l'oralité de la part de systèmes gratuits de traduction automatique neuronale (TAN), comme DeepL, Google Translate, et de systèmes de Reconnaissance automatique de la parole et de traduction automatique (RAPTAN), tels que Google Translate et Lara. Nous envisageons d'observer leurs performances concernant la traduction entre le français et l'italien et de les comparer avec les productions de l'Intelligence artificielle générative (ChatGPT4, Copilot).

Pour ce faire, deux corpus seront examinés. Le premier a été construit à partir d'un cours de traduction du français vers l'italien portant sur la traduction pédagogique (Ladmiral, 1979; Gile, 2005; Ahmadi, 2019) de divers types de textes audiovisuels d'information et de divulgation économique (France 24).

Le second est constitué d'enregistrements audiovisuels d'examens d'interprétation de dialogue entre le français et l'italien passés par des étudiants en première année de licence et relevant du domaine touristique.

Premièrement, nous proposons une comparaison entre les textes source (TS) et les textes cible (TC) issus des deux corpus mentionnés *supra* et traduits, d'une part, par les systèmes de TAN et RAPTAN et, d'autre part, par l'IAgen. Nos remarques reposent sur une grille d'analyse élaborée à partir de la littérature disponible dans le domaine de la linguistique contrastive français-italien (Carzacchi et Fonda, 2006; Sini et Merger, 2013; Bidaud, 2019, 2020; Lambertini *et al.*, 2021) revue au prisme de l'oralité et des spécificités des TS.

Cette comparaison nous permettra de vérifier l'adéquation des *outputs* des systèmes intelligents dans la traduction de ces types de texte par rapport aux spécificités des deux langues romanes prises en examen et de mettre en évidence les points forts et les points faibles des systèmes intelligents lorsqu'ils traitent des textes oraux spontanés ou semi-spontanés entre le

français et l'italien. La complexité du processus de traduction sera considérée au prisme de la variation entre l'oral et l'écrit, des enjeux posés par la co-élaboration / co-construction de la parole en interaction (Wadensjö, 1998 ; Merlini, 2015 ; Traverso, 2016 ; Niemants, 2018 ; Delizée, 2020) et des difficultés syntaxiques, sémantiques et lexicales des deux langues impliquées.

Enfin, nous constatons que l'observation des productions des systèmes de TAN et RAPTAN peut s'avérer un outil didactique de sensibilisation des apprenants aux avantages et aux inconvénients des technologies intelligentes (Aston, 2011 ; Monti, 2019), afin de renforcer leurs compétences d'analyse intra- et inter-linguistique, préalables au développement des capacités professionnelles en pré-édition et en post-édition, largement demandées par l'industrie des langues (ELIS, 2025).

References

Ahmadi Mohammad-Rahim (2019). «La traduction pédagogique, auxiliaire d'apprentissage et d'enseignement du français». *Plume. Revue semestrielle de l'Association iranienne de Langue et Littérature françaises*, n. 15 (29), pp. 25-37.

Aston Guy (2011). «Tecniche per migliorare la traduzione automatica: post-editing e pre-editing». In: Bersani Berselli Gabriele (ed.), *Usare la traduzione automatica*. Bologna: CLUEB, pp. 33-45.

Barysevich Alena, Costaris Claire (2021). « Traducteurs automatiques neuronaux comme outil didactique/pédagogique: DeepL dans l'apprentissage du français langue seconde ». *Nouvelle Revue Synergies Canada*, nº 14, pp. 1–16.

Bidaud Françoise (2019). *Traduire le français d'aujourd'hui*. Torino: UTET.

Bidaud Françoise (2020). *Grammaire française pour italophones*. Torino: UTET.

Bowker Lynne, Jairo Buitrago Ciro (eds.) (2019). *Machine translation and global research. Towards improved machine translation literacy in the scholarly community*. Bingley: Emerald. https://doi.org/10.1108/9781787567214.

Brignoli Laura (2021). «Esigenze d'oralità: dalla traduzione intralinguistica alla traduzione interlinguistica». *mediAzioni*, nº 31: A56–A79.

Carré Alice et al. (2022). « Machine translation for language learners ». In: Kenny Dorothy (ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence*. Berlin: Language Science Press.

Cennamo Ilaria (2018). *Enseigner la traduction humaine en s'inspirant de la traduction automatique*. Roma: Aracne.

Cirillo Letizia, Niemants Natacha (eds.) (2017). *Teaching dialogue interpreting: research-based proposals for higher education*. Amsterdam, Philadelphia: John Benjamins.

Delizée Anne (2020). « Quand dire, c'est agir sur l'autre : conscientiser les influences mutuelles lors d'une interaction bilingue interprétée ». *Bulletin de l'Institut de linguistique*, nº 31, p. 39-60

Ehrensberger-Dow Maureen, Massey Gary (2019). «Le traducteur et la machine. Mieux travailler ensemble ». *Des mots aux actes*, n. 8, Paris: Éditions Classiques Garnier, 47-62.

European Language Industry Survey (ELIS) (2025). *Trends, expectations and concerns of the European language industry.* https://elis-survey.org/wp-content/uploads/2025/03/ELIS-2025_Report.pdf

Fantinuoli Claudio, Prandi Bianca (2021). « Towards the evaluation of automatic simultaneous speech translation from a communicative perspective ». In: *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021), Bangkok, Thailand (online).* Association for Computational Linguistics, pp. 245–254.

Fantinuoli Claudio (2023). *EasyAI - Introducing Artificial Intelligence to the Humanities, Version 1*, Last updated: 7 Jan 2023 (https://easyai.uni-mainz.de/html/index.html).

Gile Daniel (2005). *La traduction: la comprendre, l'apprendre*. Paris: PUF.

Hernandez-Morin Katell (2019). « Évolution des technologies et des usages en traduction. Pratique et enseignement de la post-édition ». *Des mots aux actes*, n. 8, pp. 239-255.

Jiménez-Crespo Miguel Ángel (2017). « The role of translation technologies in Spanish language learning ». *Journal of Spanish Language Teaching* 4, pp. 181–193. https://doi.org/10.1080/23247797.2017.1408949.

Kenny Dorothy (ed.) (2022). *Machine translation for everyone: Empowering users in the age of artificial intelligence*. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.6653406.

Ladmiral Jean-René (1979). *Traduire: théorèmes pour la traduction*. Paris: Petite Bibliothèque Payot.

Lambertini Vincenzo, Baldi Lucia, Toni Patricia (2021). «Interpretare tra francese e italiano». In: Russo Mariachiara (ed.), *Interpretare da e verso l'italiano. Didattica e innovazione per la formazione dell'interprete*. Bologne: BUP, p. 191-210.

Loock, Rudy, Sophie Léchauguette, Benjamin Holt (2022). « The use of online translators by students not enrolled in a professional

translation program: beyond copying and pasting for a professional use ». In: Helena Moniz *et al.* (eds.), *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. Ghent: European Association for Machine Translation, pp. 23–29.

Loock, Rudy, Holt, Benjamin (2024). Augmented Linguistic Analysis Skills: Machine Translation and Generative AI as Pedagogical Aids for Analyzing Complex English Compounds. *Technology in Language Teaching & Learning*, 6 (3), 127. https://doi.org/10.29140/tltl.v6n3.1489

Mattioda Maria Margherita, Cennamo Ilaria (2023). «La traduzione automatica neurale: uno strumento di sensibilizzazione per la formazione universitaria in lingua e traduzione francese». *De Europa*, Special Issue.

Merlini Raffaela (2015). « Dialogue Interpreting ». In Franz Pöchhacker (ed.), *Routledge Encyclopedia of Interpreting Studies*. Londres, New York: Routledge, p. 101-106.

Monti Johanna (2019). *Dalla Zairja alla traduzione automatica. Riflessioni sulla traduzione nell'era digitale.* Naples: Paolo Loffredo Editore.

Niemants Natacha (2018). « L'interprétation des français parlés en interaction ». *inTRAlinea, Special Issue: Translation and Interpreting for Language Learners (TAIL)*. http://www.intralinea.org/specials/article/linterpretation_des_francais_ parles_en_interaction.

O'Brien Sharon (2020). « Translation, human–computer interaction and cognition 1 ». In: Fabio Alves, Arnt Lykke Jakobsen (eds.), *The Routledge Handbook of Translation and Cognition*. London: Routledge, pp. 376-388.

Sini Lorella, Merger Françoise (2013). *Le nouveau côte à côte*. Rome: Amon.

Traverso Véronique (2016). *Décrire le français parlé en interaction*. Paris: Ophrys.

Wadensjö Cecilia (1998). *Interpreting as Interaction*. London, New York: Longman.

Yamada Masaru (2019). «The impact of Google Neural Machine Translation on post-editing by student translators». *The Journal of Specialised Translation*, 31, pp. 87–106. https://jostrans.org/issue31/art_yamada.php.

Zanetti Mouillet Françoise, Carzacchi Fonda Michèle (2006), *L'acrobatraducteur. Réflexions et exercices grammaticaux pour la traduction italien-français*, Roma: Aracne.

Attitude and subjectivity in human-written and AI-generated editorials published in Il Foglio

Franz Meier (Universität Augsburg)

In March 2025, the Italian newspaper *Il Foglio* attracted international attention with the launch of *Il Foglio AI*, a month-long experiment presenting a daily four page supplement entirely generated by artificial intelligence (AI). Due to its success and the interest it garnered, the experiment has been extended beyond its original timeframe which one AI-generated supplement published per week. The journal's initiative aims to investigate both the potential and limitations of AI within journalistic practice, prompting broader discussions about the future of news production and the necessity of human editorial oversight to uphold journalistic integrity. Each edition of *Il Foglio AI* features around 25 AI-generated articles spanning various journalistic genres, including editorials. These opinion-based texts – traditionally authored by senior editorial staff and often unsigned – serve to articulate a newspaper's official position on current issues (cf. Barbano/Sassu 2012). Like its AI-produced counterpart, *Il Foglio* regularly publishes editorials which, in fact, are central to the journal's editorial identity. Distinct from many traditional newspapers, *Il Foglio*, founded in 1996 and known for its conservative-liberal editorial stance, is particularly recognized for its analytical tone and commentary-driven approach, often placing greater emphasis on reflection and interpretation than on straightforward news reporting (cf. Draghi 2005, Gualdo 2017).

The purpose of this contribution is to examine AI-generated editorials published in *Il Foglio AI* and to compare them to human-written editorials published in *Il Foglio*. Due to the nature of editorials as opinion-based texts, the both corpus-based and corpus-driven study focuses on the presence of markers of authorial stance or subjectivity, that is, the expression of the speaker's attitudes, beliefs, feelings, emotions, judgement, will, personality, etc. (Lyons 1982). Following Martin/White (2005), we investigate categories such as counter-expectation, intensification, reporting of mental processes, expression of engagement with external voices and, possibly, other modes as yet not identified (cf. also Pounds 2010). Based on these categories, the contrastive

study explores to what extent the authorical stance is conveyed explicitly to the readers depending on the type of editorial – human written or AI generated – concerned. The study of markers of subjectivity is especially interesting with regard to the fact that large language models (LLMs) such as ChatGPT do not experience the world. They only generate text based on patterns in data, but they cannot truly understand contexts like humans do. Previous research has shown that AI can provide fluent, fact-oriented texts (see e.g. De Cesare 2022), while more research is needed in the domain of opinion-based texts which seem to require human judgment and first-hand understanding to comment on complex or contested facts.

References

Barbano, A. and Sassu, V. (2012), *Manuale di giornalismo*. Rome/Bari: Laterza

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", in *CHIMERA. Romance Corpora and Linguistic Studies* 10, 179-210.

Draghi, C. (2005), "Il giornale attivista. I dieci anni del Foglio di Giuliano Ferrara", in *Problemi dell'informazione* 4, 414-437.

Gualdo, R. (2017), *L'italiano dei giornali*. Rome: Carocci editore.

Lyons, J. (1982), "Deixis and Subjectivity: Loquor Ergo Sum?", in R. Jarvella and W. Klein (eds), *Speech, Place and Action: Studies in Deixis and Related Topics*. Chichester: Wiley, 101-124

Martin, J.R. and White, P.R.R. (2005), *The Language of Evaluation: Appraisal in English*. Basingstoke: Palgrave.

Pounds, G. (2010), "Attitude and subjectivity in Italian and British hard-news reporting: The construction of a culture-specific 'reporter' voice", in *Discourse Studies* 12(1), 106-137.

Migliorare la chiarezza dei testi amministrativi con ChatGPT: analisi qualitativa degli interventi sintattici e degli effetti testuali e comunicativi

Mariachiara Pascucci (Università di Pisa/Universität Basel) &
Angela Ferrari (Universität Basel)

La complessità dei testi amministrativi italiani rappresenta una sfida significativa per una comunicazione efficace tra enti pubblici e cittadini. Recentemente, l'interesse per le potenzialità dei modelli linguistici di grandi dimensioni (LLM), come ChatGPT, per la semplificazione e l'ottimizzazione della chiarezza di tali testi è cresciuto notevolmente, facendo emergere l'esigenza di approfondire la conoscenza delle caratteristiche linguistiche dei prodotti di queste operazioni automatiche e di verificarne la qualità. Studi come Tavosanis (2024) esplorano questo problema, affrontando la delicata questione della valutazione dei testi generati.

Il presente lavoro, concepito come parte di un più ampio progetto di ricerca e con un'impostazione complementare rispetto alle attività di valutazione, si propone di descrivere le proprietà linguistiche delle rielaborazioni di testi amministrativi in italiano generate da ChatGPT, con particolare attenzione agli aspetti sintattici. La complessità sintattica, nelle sue varie manifestazioni, è infatti un tratto distintivo dell'italiano amministrativo, come evidenziato, tra gli altri, da Fortis (2005) nella sua ricognizione sistematica. Le raccomandazioni fornite dai manuali e delle linee guida per la redazione di testi amministrativi (Fioritto 1997, ITTIG/Accademia della Crusca 2011, Cortelazzo 2021) mirano quindi a mitigare tale caratteristica, suggerendo, ad esempio, di ridurre la lunghezza dei periodi, evitare le frasi complesse e privilegiare strutture semplici e lineari.

La presente analisi si basa su un corpus di rielaborazioni effettuate con ChatGPT-4o a partire da sezioni autonome di testi più ampi (linee guida e circolari ministeriali). Le riscritture sono state generate attraverso la tecnica del *role prompting*, con un prompt formulato per migliorare la chiarezza del testo preservando la completezza delle informazioni.

I testi generati sono stati analizzati sia come testi autonomi che in relazione agli originali. L'indagine quantitativa preliminare,

di cui verranno presentati i risultati, ha evidenziato, per le rielaborazioni automaticamente generate, una riduzione significativa della lunghezza dei periodi e una tendenza alla semplificazione della struttura frasale, in linea con quanto indicato dalla letteratura sul miglioramento della chiarezza. I testi rielaborati mostrano una sintassi meno articolata rispetto agli originali e, come già osservato da Cicero (2023) per i testi informativi generati *ex novo*, presentano una struttura semplice e lineare. Permangono tuttavia interrogativi fondamentali, già sollevati da Ferrari (2021), sull'effettiva utilità di tali interventi: è sempre vero che una sintassi più semplice favorisce la coerenza e la chiarezza? Questi dubbi assumono particolare rilievo se la semplicità sintattica viene analizzata in relazione alla dimensione testuale e comunicativa.

Questo studio propone dunque un'analisi qualitativa approfondita con il fine di esaminare nel dettaglio gli effetti della rielaborazione automatica sugli aspetti sintattici, con particolare attenzione alla riduzione della lunghezza dei periodi e agli interventi sulle subordinate. L'analisi si concentra su come tali trasformazioni influenzino l'organizzazione del testo, la segmentazione e la gerarchizzazione delle informazioni, utilizzando come riferimento il Modello Basilese della testualità (Ferrari et al. 2021), e sull'impatto effettivo che gli interventi introdotti hanno sulla chiarezza.

L'obiettivo è contribuire alla fase esplorativa sull'impiego dei LLM, arricchendo la riflessione sui criteri linguistici per l'analisi e la valutazione dei testi amministrativi generati automaticamente.

References

Cicero, Francesco. (2023). «L'italiano delle intelligenze artificiali generative». *Italiano LinguaDue*, 15(2), 733–761.

Cortelazzo, M. (2021). *Il linguaggio amministrativo: principi e pratiche di modernizzazione*. Roma: Carocci.

Ferrari, A. (2021). «La semplicità sintattica in prospettiva testuale. Riflessioni a partire dalla Guida alla redazione degli atti amministrativi». Italiano digitale, 16(1), 188–195.

Ferrari, A., Carlevaro, A., Evangelista, D., Lala, L., Marengo, T., Pecorari, F., Piantanida, G., & Tonani, G. (a cura di). (2024). *La comunicazione istituzionale durante la pandemia. Il Ticino, con uno sguardo ai Grigioni*. Bellinzona: Casagrande.

Ferrari, A./Lala, L. /Zampese, L. (2021), *Le strutture del testo scritto. Teoria e esercizi*, Roma: Carocci.

Fiorentino, G., & Ganfi, V. (2024). «Parametri per semplificare l'italiano istituzionale: revisione della letteratura». *Italiano LinguaDue, 16(1), 220–237*. https://doi.org/10.54103/2037-3597/23835

Fioritto, A. (a cura di). (1997). *Manuale di stile. Strumenti per semplificare il linguaggio delle amministrazioni pubbliche*. Bologna: Il Mulino. (Presidenza del Consiglio dei Ministri, Dipartimento della Funzione Pubblica)

Fortis, D. (2004). «Semplificazione del linguaggio amministrativo: validità e limiti delle linee guida». *Rivista italiana di comunicazione pubblica, (20), 48–83*. Milano: Franco Angeli.

Fortis, D. (2005). «Il linguaggio amministrativo italiano», in *Revista de Llengua i Dret, n. 43, 2005*, pp. 47-116.

ITTIG – Istituto di Teoria e Tecniche dell'Informazione Giuridica & Accademia della Crusca (a cura di). (2011). *Guida alla redazione degli atti amministrativi. Regole e suggerimenti*. Firenze.

Piemontese, E. (1996). *Capire e farsi capire. Teorie e tecniche della scrittura controllata*. Napoli: Tecnodid.

Tavosanis, M. (2024). «Valutare la qualità dei testi generati in lingua italiana». AI-Linguistica, 1(1), 1–24.

"Write a bedtime story for my four year old daughter!" Features and implications of ChatGPT-generated narrations

Robert Cornelis Schuppe (Technische Universität Dresden)

Narrativity plays an integral role in culture and communication. Its importance for human cognition has been a recurring focus of research from different disciplines (e.g. Turner 1996; Alber/Wenzel 2021). The analysis of stories generated by generative artificial intelligence however constitutes a newly emerging field of research (Albrecht 2024). The use of this type of technology has increased massively since the release of ChatGPT in November 2022. Among others, it can be utilized to generate stories for specific purposes. In how far these outputs follow typical patterns and in which ways stories generated by ChatGPT differ depending on specific prompts are questions yet to be researched in detail.

This paper aims at analyzing the features of German language bedtime stories generated by ChatGPT. The basis for this analysis is a corpus which has been collected over the period of more than a year by feeding a list of prompts into ChatGPT on a monthly basis. These prompts contain the request for a bedtime story as well as varying information about the age and gender of the target person. This data is used for an analysis located at the intersection of linguistic research on artificial intelligence on the one hand and sociolinguistics and gender linguistics on the other.
The paper employs a variety of corpus linguistic methods to analyze the choice of topics and the narrative structure of stories for different age groups and genders. The metadata on the age and gender of the stories' target persons provides the criteria for dividing the corpus into multiple sub-corpora. These sub-corpora are utilized for a keyword analysis as well as a comparison of n-gram-frequencies in order to gain insight on prevalent topics and constructions in the different types of stories.

Furthermore, the corpus is scrutinized with regard to the macro-structure of the stories.
Following Bubenhofer et al. (2013), each text is divided into a number of equally long parts. Subsequently, an additional set of sub-corpora is created where each sub-corpus contains the same segments of every individual story (e.g. all the beginnings or all the

ends of the stories). These sub-corpora are used for an additional contrastive analysis which sheds light on the overall structure and the prototypical narrative arc of the generated stories.

By using the outlined methodological approach, this paper aims at contributing to a better understanding of the output generated by large language models. This can help disentangling implications of AI-generated texts such as gender or age biases as well as understanding their overall characteristics related to macro-structure, thematic focus, and word choice.

References

Alber, Jan & Peter Wenzel (Eds., 2021): Introduction to Cognitive Narratology (WVT-Handbücher zum literatur- und kulturwissenschaftlichen Studium 24). Trier: WVT.

Albrecht, Steffen (2024): ChatGPT als doppelte Herausforderung für die Wissenschaft. Eine Reflexion aus der Perspektive der Technikfolgenabschätzung. In: Gerhard Schreiber & Lukas Ohly (Hg.): *KI:Text. Diskurse über KI-Textgeneratoren*. Berlin, Boston: De Gruyter. S. 13-27.

Bubenhofer, Noah, Nicole Müller & Joachim Scharloth (2013): Narrative Muster und Diskursanalyse: Ein datengeleiteter Ansatz. In: *Zeitschrift für Semiotik, Methoden der Diskursanalyse 35* (3-4), S. 419-444.

Turner, Mark (1996): *The Literary Mind*. New York, Oxford: Oxford University Press.

Gli errori grammaticali degli LLM: diversità tra sistemi e caratteristiche generali

Mirko Tavosanis (Università di Pisa)

Il contributo presenta i risultati di un confronto tra LLM. In particolare, al suo interno viene presa in esame la produzione in lingua italiana, su compiti confrontabili, di ChatGPT-4o, Minerva LLM e DeepSeek. I prompt usati includono richieste come “Puoi scrivere un comunicato stampa sulle nuove attività di miglioramento della scrittura nelle scuole della Toscana?” e “Puoi scrivere un articolo di opinione sull'importanza del contrasto al cambiamento climatico?” La lingua dei testi generati è stata quindi l’italiano neo-standard di taglio amministrativo o giornalistico.

Gli LLM presi in esame hanno caratteristiche molto diverse fra loro e di fatto hanno prodotto testi con caratteristiche differenziate per qualità e per lunghezza. Per esempio, le risposte di ChatGPT sono spesso piuttosto lunghe e basate su frasi brevi (14,3 token per periodo), mentre le risposte di Minerva LLM sono cinque volte più brevi ma composte di frasi mediamente più lunghe (17 token per periodo). Le risposte di DeepSeek sono nel complesso più lunghe di quelle di qualunque altro sistema, ma le singole frasi si collocano a un livello intermedio rispetto agli altri due (15,7 token per periodo). Valutatori esperti hanno poi esaminato la funzionalità linguistica dei testi, notandone la sostanziale correttezza.

Analizzando quantitativamente frasi scelte in modo casuale nella loro produzione, si nota però che tutti gli LLM presi in esame hanno una qualità grammaticale molto simile, con un numero di errori piuttosto contenuto: meno del 10% delle frasi include veri e propri errori grammaticali, con una differenza quantitativa ridotta tra gli esempi. Inoltre, le tipologie di errore sono molto simili tra i diversi modelli e riguardano soprattutto la coesione sintattica a distanza ravvicinata. In alcuni casi, addirittura, le strutture erronee generate in risposta allo stesso prompt sono sostanzialmente identiche tra modelli diversi.

I risultati dell’indagine corroborano quindi l’ipotesi che le caratteristiche grammaticali dei testi generati dipendano assai più dalle caratteristiche qualitative dei materiali di addestramento e dei meccanismi generali di addestramento e generazione che dalle

caratteristiche specifiche e dalla dimensione dei singoli sistemi. Per ottenere buoni risultati dal punto di vista grammaticale, la disponibilità di raccolte di testi di enormi dimensioni non sembra tanto cruciale quanto è stato ipotizzato in passato: anche modelli di dimensioni ridotte riescono ad "apprendere" la grammatica di una lingua in modo paragonabile a quello di strumenti addestrati su raccolte di testi assai più ampie. Anzi, le dimensioni dei testi raccolti sono tali da essere in alcuni casi paragonabili a quelle utilizzate dei testi usati dagli esseri umani per il proprio normale apprendimento, rendendo più verosimili le ipotesi che anche l'apprendimento linguistico umano abbia una forte componente statistica.

References

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.

Serianni, Luca. 1988. *Grammatica italiana*. Torino, UTET.

Tavosanis, Mirko. 2024. "Valutare la qualità dei testi generati in lingua italiana", *AI-Linguistica*, 1, pp. 1-24.

Linguistic features and ethical implications of ChatGPT-generated Italian translations of news: a case study on hate speech

Aurora Trapella (Università di Torino)

The emergence of Generative AI (GenAI) based on Large Language Models (LLM) and AI-smart agents such as chatbots is reshaping translation practices (Munday et al., 2022; Siu, 2024). These models, trained on vast datasets and capable of capturing broader contextual nuances (Niu & Jiang, 2024), are progressively integrated into journalistic workflows (Roush, 2023a, 2023b) to provide first-hand translations. As Neural Machine Translation (NMT) increasingly serves as an assimilation tool (Koehn, 2020) in news consumption, questions arise on the linguistic integrity, ideological neutrality, and the ethical implications of NMT use in translating politically sensitive content. With minimal post-editing and increasing dependence on AI systems by editors often lacking formal translation training (Troqe & Marchan, 2017), there is a growing risk that such tools may unintentionally alter meaning and perpetuate bias.

This study investigates the linguistic features and potential ideological shifts in AI-generated translations from English to Italian of news articles covering the controversial and challenging phenomenon of hate speech, because of its blurred boundaries with free speech (Zottola, 2020). This topic presents a critical test case for evaluating the discursive behaviour of GenAI tools in the translation of content with inherently politicised language.

Drawing on a small, specialised corpus of 15 English-language news articles focused on hate speech in the context of socio-political tensions during the current Trump administration and the ensuing controversy over “woke culture” (Yourish et al., 2025), the articles are translated into Italian using the free online version of ChatGPT 4.0 using a zero-shot translation prompt (Jiao 2024), without any human post-editing.

This study is guided by two primary research questions. Firstly, this study seeks to identify and describe the distinctive linguistic features of automatically generated Italian translations of news articles on hate speech produced by GenAI-enabled chatbots. Secondly, it explores the role and the ethical implications of the use

of ChatGPT in journalistic translation practices and how it affects politically sensitive content. Through a qualitative approach grounded in Critical Discourse Analysis (CDA) (Machin & Mayr, 2023), this paper examines how GenAI-enabled tools potentially affect meaning, drawing on studies such as Song's (2020) that illustrate subtle ideological shifts in translated political discourse.

Preliminary results indicate that the translated output features natural syntax and a style in line with news articles; however, language choices by the machine are not always neutral, as political stance can be subtly conveyed not only through lexical choices, but also omissions, shifts, lexical mitigations and reformulations of sentences including sensitive topics or potentially politically-loaded information. Moreover, the choice of contextually appropriate lexicon, cultural terms, references, and idiomatic expressions warrants investigation.

By exploring the linguistic features of GPT-generated translations, this study contributes to ongoing debates on *machine translationese* (Daems et al. 2018; Loock 2018; De Clercq et al. 2021), the features of automatically generated translations in the Italian language and the ethical challenges related to the use of AI in journalism, seeking to inform future standards for AI translation practices, given the increasingly relevant role of these tools.

References

Daems, J., De Clercq, O., & Macken, L. (2018). Translationese and Post-editese: How comparable is comparable quality?. *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, *16*. https://doi.org/10.52034/lanstts.v16i0.434

De Clercq, O., Sutter, G., Loock, R., Cappelle, B., & Plevoets, K. (2021). *Uncovering Machine Translationese Using Corpus Analysis Techniques to Distinguish between Original and Machine--Translated French. Translation Quarterly, 2021, 101, pp.21-45.* https://hal.science/hal-03406287v1

Jiao, H., Peng, B., Zong, L., Zhang, X., & Li, X. (2024). *Gradable ChatGPT translation evaluation*. arXiv. https://arxiv.org/abs/2401.09984

Koehn, P. (2020). Uses of Machine Translation. In *Neural Machine Translation* (pp. 19-28). Cambridge: Cambridge University Press. https://doi:10.1017/9781108608480.003

Loock, R. (2020). No more rage against the machine: how the corpus-based identification of machine-translationese can lead to student

empowerment. *The Journal of specialised translation (JoSTrans), 2020, 34, pp.150-170.* https://shs.hal.science/halshs-02913980v1

Roush, C. (2023a) What Reuters is telling its journalists about using artificial intelligence. Newstex Blogs Talking Biz News. https://talkingbiznews.com/media-news/what-reuters-is-telling-its-journalists-about-using-artificial-intelligence/

Roush, C. (2023b) FT editor in chief Khalaf on how it will use artificial intelligence. Newstex Blogs Talking Biz News. https://talkingbiznews.com/media-news/149829/

Song, Y. (2020). Ethics of Journalistic Translation and its Implications for Machine Translation: A Case Study in the South Korean Context. *Babel: Revue Internationale de La Traduction/International Journal of Translation, 66*(4–5), 829–846.

Troqe, R., & Marchan, F. (2017). News Translation: Text analysis, fieldwork, survey. In S. Hansen-Schirra, O. Czulo & S. Hofmann (Eds) *Empirical modelling of translation and interpreting* (pp. 277–310). Language Science Press.

Yourish, K., Daniel, A., Datar, S., White, I., & Gamio, L. (2025, March 7). These words are disappearing in the new Trump administration. The New York Times. https://www.nytimes.com/interactive/2025/03/07/us/trump-federal-agencies-websites-words-dei.html. Accessed April 5.

Zottola, A. (2020). When freedom of speech turns into freedom to hate: Hateful speech and 'othering' in conservative political propaganda in the USA. In G. Balirano & B. Hughes (Eds.), *Homing in on hate: Critical discourse studies of hate speech, discrimination and inequality in the digital age* (pp. 93–112). Napoli: Loffredo Editore.

Between Human and Artificial Language: Comparing the Syntax of Large Language Models and German Journalistic Texts

Paolo Valentinelli (Università di Trento)

Recent developments in language models represent a significant breakthrough in the automatic generation and processing of natural language texts. This linguistic capability stems from a pre-training process on text corpora predominantly in English – over 90% of the data. This preliminary study examines the implications of this linguistic imbalance in language model training (where German accounts for only 1.5% of the data) by systematically analysing the syntactic and structural properties of generated texts and comparing them with those found in contemporary German journalistic language. The choice of journalistic language as a reference is motivated by its codification in the specialised literature and its social and communicative relevance.

Six key syntactic features were identified and grouped into three macro-areas, based on the specialised literature: (i) sentence length and readability; (ii) word order and pre-field occupation; and (iii) diathesis and agent realisation. The investigation focuses on two widely circulated German newspapers that represent different communicative styles: *Süddeutsche Zeitung*, a 'quality' daily newspaper, and *BILD*, considered a tabloid. The selected texts come exclusively from the Politics section to ensure thematic and textual consistency. The corpus was then compared with a collection of texts generated by four large language models: OpenAI's o1, Anthropic's Claude 3.5 Sonnet, Meta AI's llama-3.1-405b-instruct, and Mistral's mistral-large-2407. Each model generated 25 texts using zero-shot prompting with chain-of-thought techniques. Lastly, the analysis was also supported by statistical tests (*t-tests* and *chi-squared tests*) and an equivalence test with a 10% tolerance threshold to assess both statistical significance and potential functional similarity between the two kinds of texts.

The results show that none of the analysed language models produce texts that are fully equivalent to human texts from a syntactic standpoint, according to the defined equivalence threshold. However, some models come closer than others with

respect to certain features. A noteworthy finding concerns the occupation of the pre-field, i.e., the type of constituent that appears in the first position of the sentence. In fact, all models exhibit a more limited range of constituents compared to human-authored texts, both functionally and formally. Word order also differs significantly: whilst in the analysed journalistic texts the +SVO and –SVO orders are distributed with relative balance, the models tend to apply a +SVO order more systematically and rigidly. These findings indicate a clear preference for constructions more closely aligned with English syntax, likely reflecting the linguistic bias inherent in the training data.

References

Burger, H. & Luginbühl, M., 2014. *Mediensprache. Eine Einführung in Sprache und Kommunikationsformen der Massenmedien*. 4., neu bearbeitete und erweiterte Aufl. Berlin, Boston: De Gruyter.

De Cesare, A.-M., 2007. „Le funzioni del passivo agentivo. Tra sintassi, semantica e testualità". *Vox Romanica*, 66, pp. 32–59.

Hofstätter, A., 2020. *Entwicklungen in der deutschen Nachrichtensprache*. München: LMU München. Dissertation.

Johnson, R. L., et al., 2022. *The Ghost in the Machine has an American Accent: Value Conflict in GPT-3*. arXiv.

Linden, P., 2008 [1998]. *Wie Texte wirken: Anleitung zur Analyse journalistischer Sprache*. 3. Aufl. Berlin: ZV Zeitungs-Verlag.

Mittelberg, E., 1967. *Wortschatz und Syntax der Bild-Zeitung*. Marburg: N. G. Elwert Verlag.

Schmitt, U., 2004. *Diskurspragmatik und Syntax. Die funktionale Satzperspektive in der französischen und deutschen Tagespresse unter Berücksichtigung einzelsprachlicher, pressetyp- und textklassenabhängiger Spezifika*. Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien: Peter Lang.

Wöllstein-Leisten, A., et al., 2006 [1997]. *Deutsche Satzstruktur. Grundlagen der syntaktischen Analyse*. Unveränderter Nachdruck der 1. Auflage 1997. Tübingen: Stauffenburg Verlag Brigitte Narr GmbH.

––––––, et al., 2022. *Die Grammatik. Struktur und Verwendung der deutschen Sprache. Satz – Wortgruppe – Wort*. 10. Aufl. Berlin: Dudenverlag.