Tra linguaggio umano e artificiale
La sintassi dei *large language models* e del linguaggio giornalistico tedesco

Paolo Valentinelli (Università di Trento)
paolo.valentinelli@studenti.unitn.it

Abstract

I recenti sviluppi nei modelli linguistici, come, ad esempio, *ChatGPT*, rappresentano una svolta significativa nella generazione ed elaborazione automatica di testi in lingua naturale. Questa loro abilità linguistica deriva da un processo di *pre-training* su corpora testuali prevalentemente in lingua inglese: oltre il 90% di questi dati. Questo studio preliminare analizza le implicazioni di tale squilibrio linguistico sul piano sintattico e strutturale, concentrandosi sulla produzione testuale in tedesco (1,5% dei dati di allenamento) e su quali tratti distintivi emergano nei testi generati rispetto a quelli prodotti da esseri umani. In particolare, si confrontano i testi artificiali di quattro *large language models* (o1 di OpenAI, Claude 3.5 Sonnet di Anthropic, Llama 3.1 di Meta e Mistral Large 2 di Mistral AI) con articoli di giornalismo politico pubblicati da testate come *Süddeutsche Zeitung* e *BILD*. Sebbene i testi generati possano rappresentare la forma dello stile giornalistico tedesco a livello superficiale, i risultati mostrano che essi presentano deviazioni strutturali profonde che rivelano che, in primis, l'equivalenza strutturale con la scrittura umana non è stata raggiunta, ma, soprattutto, che l'inglese influenza in maniera significativa l'*output* dei testi generati in lingua tedesca.

Bibliografia

Amstad, T., 1978. *Wie verständlich sind unsere Zeitungen?*. Universität Zürich: Dissertation.

Anthropic, 2024a. *Claude 3.5 Sonnet*.

———, 2024b. *Model Card Addendum: Claude 3.5 Haiku and Upgraded Claude 3.5 Sonnet*.

Brown, T. B., et al., 2020. *Language Models are Few-Shot Learners*. arXiv.

Bucher, H.-J. & Straßner, E., 1991. *Mediensprache. Medienkommunikation. Medienkritik*. Tübingen: Gunter Narr Verlag.

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

Bußmann, H., 1990. *Lexikon der Sprachwissenschaft*. Stuttgart: Kröner.

Chen, B., et al., 2024. *Unleashing the Potential of Prompt Engineering in Large Language Models: A Comprehensive Review*. arXiv.

Chiang, W.-L., et al., 2024. *Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference*. arXiv.

Chowdhery, A., et al., 2023. „PaLM: Scaling Language Modeling with Pathways“. *Journal of Machine Learning Research*, 24(240), S. 1–113.

de Beaugrande, R.-A. & Dressler, W. U., 1994. *Introduzione alla linguistica testuale*. Bologna: Il Mulino.

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

———, et al., 2016. *Sintassi marcata dell'italiano dell'uso medio in prospettiva contrastiva con il francese, lo spagnolo, il tedesco e l'inglese: Uno studio basato sulla scrittura dei quotidiani online*. Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien: Peter Lang Edition.

Desmond, M. & Brachman, M., 2024. *Exploring Prompt Engineering Practices in the Enterprise*. arXiv.

Devlin, J., et al., 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv.

Dudenredaktion, 2020. *Die deutsche Rechtschreibung. Das umfassende Standardwerk auf der Grundlage der amtlichen Regeln*. 28., völlig neu bearbeitete und erweiterte Aufl. Berlin: Dudenverlag.

Fan, L., et al., 2024. „A Bibliometric Review of Large Language Models Research from 2017 to 2023“. *ACM Transactions on Intelligent Systems and Technology*, 15(5), S. 1–25.

Farquhar, S., et al., 2024. „Detecting hallucinations in large language models using semantic entropy“. *Nature*, Band 630, S. 625–30.

Fritz, G. & Straßner, E., 1996. *Die Sprache der ersten deutschen Wochenzeitungen im 17. Jahrhundert*. Tübingen: Max Niemeyer Verlag.

Herrmann, G., 2022. „Wie die Homepage entsteht“. *Süddeutsche Zeitung*. https://www.sueddeutsche.de/kolumne/sz-homepage-77-jahre-1.5660484 [Letzter Zugriff am 19. März 2025].

Hochreiter, S. & Schmidhuber, J., 1997. „Long short-term memory“. *Neural Computation*, 9(8), S. 1735–80.

Höhle, T. N., 1978. *Lexikalische Syntax*. Tübingen: Niemeyer.

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

Holl, D., 2010. *Modale Infinitive und dispositionelle Modalität im Deutschen*. Berlin: Akademie Verlag.

Hooffacker, G. & Meier, K., 2017. *La Roches Einführung in den praktischen Journalismus. Mit genauer Beschreibung aller Ausbildungswege Deutschland · Österreich · Schweiz*. 20., neu bearbeitete Aufl. Wiesbaden: Springer VS.

Hoser, P, 2014. „Süddeutsche Zeitung (SZ)“. *Historisches Lexikon Bayerns*. https://www.historisches-lexikon-bayerns.de/Lexikon/S%C3%BCddeutsche_Zeitung_(SZ) [Letzter Zugriff am 19. März 2025].

Jackley, M., 2024. *What Is a Large Language Model (LLM)?*. https://www.oracle.com/artificial-intelligence/large-language-model/ [Letzter Zugriff am 19. März 2025].

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

Kahneman, D., 2013. *Thinking, Fast and Slow*. New York City: Farrar, Straus and Giroux.

Kaplan, J., et al., 2020. *Scaling Laws for Neural Language Models*. arXiv.

Khurana, D., et al., 2017. *Natural Language Processing: State of The Art, Current Trends and Challenges*. arXiv.

Koch, P. & Oesterreicher, W., 1986. „Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte“. In: *Romanistisches Jahrbuch*. Berlin, New York: Walter de Gruyter.

Kojima, T., et al., 2023. *Large Language Models are Zero-Shot Reasoners*. arXiv.

La Roche, W. v., 1988 [1975]. *Einführung in den praktischen Journalismus. Mit genauer Beschreibung aller Ausbildungswege*. München: List.

Lenerz, J., 1977. *Zur Abfolge nominaler Satzglieder im Deutschen*. Tübingen: Stauffenberg.

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

Llama Team @ Meta AI, 2024. *The Llama 3 Herd of Models*.

Lüger, H.-H., 1977. *Journalistische Darstellungsformen aus linguistischer Sicht: Untersuchungen zur Sprache der französischen Presse mit besonderer Berücksichtigung des „Parisien libéré“*. Freiburg im Breisgau: Albert-Ludwigs-Universität Freiburg. Dissertation.

———, 1995. *Pressesprache*. Berlin, New York: Max Niemeyer Verlag.

Lünenborg, M., 2013. „Boulevardisierung“. In: G. Bentele, H. Brosius & O. Jarren. *Lexikon Kommunikations- und Medienwissenschaft*. Wiesbaden: Springer VS.

Meta AI, 2023. *Introducing LLaMA: A foundational, 65-billion-parameter large language model*. https://ai.meta.com/blog/large-language-model-llama-meta-ai/ [Letzter Zugriff am 19. März 2025].

Metzeltin, M. & Jaksche, H., 1983. *Textsemantik: ein Modell zur Analyse von Texten*. Tübingen: Gunter Narr.

Meyn, H., 1985 [1974]. *Massenmedien in der Bundesrepublik Deutschland. Zur Politik und Zeitgeschichte*. Neu bearb. Aufl. Berlin: Colloquium-Verlag.

Mikolov, T., et al., 2013. *Efficient Estimation of Word Representations in Vector Space*. arXiv.

Mistral AI Team, 2024. *Large Enough*.

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

Ollivier, M., et al., 2023. „A deeper dive into ChatGPT: history, use and future perspectives for orthopaedic research. Knee Surgery, Sports Traumatology, Arthroscopy". *Official Journal of the ESSKA*, 31(4), S. 1190–1192.

OpenAI, 2023. *GPT-4 Technical Report*. arXiv.

———, 2024a. *Best practices for prompt engineering with the OpenAI API. How to give clear and effective instructions to OpenAI models*. https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api? [Letzter Zugriff am 19. März 2025].

———, 2024b. *GPT-4o System Card*. arXiv.

———, 2024c. *GPT-4o mini: advancing cost-efficient intelligence. Introducing our most cost-efficient small model*. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ [Letzter Zugriff am 19. März 2025].

———, 2024d. *Learning to Reason with LLMs*.

———, 2024e. *OpenAI o1 System Card*. arXiv.

Phoenix, J. & Taylor, M., 2024. *Prompt Engineering for Generative AI: Future-Proof Inputs for Reliable AI Outputs*. Sebastopol: O'Reilly Media.

Plöchinger, S., 2012a. „Schöner, schlichter, besser". *Süddeutsche Zeitung*. https://www.sueddeutsche.de/kolumne/unsere-neue-seite-schoener-schlichter-besser-1.1246740 [Letzter Zugriff am 19. März 2025].

———, 2012b. „Opulenter, innovativer, lesbarer". *Süddeutsche Zeitung*. https://www.sueddeutsche.de/kolumne/frisches-layout-fuer-sz-de-opulenter-innovativer-lesbarer-1.1531529 [Letzter Zugriff am 19. März 2025].

Raabe, J., 2013. „Boulevardpresse“. In: G. Bentele, H. Brosius & O. Jarren. *Lexikon Kommunikations- und Medienwissenschaft*. 2. überarbeiteten und erweiterten Aufllage Hrsg. Wiesbaden: Springer VS, S. 33–4.

Radford, A., et al., 2018. *Improving Language Understanding by Generative Pre-Training*.

———, et al., 2019. *Language Models are Unsupervised Multitask Learners*. OpenAI.

Reiners, L., 1976 [1943]. *Stilkunst. Ein Lehrbuch deutscher Prosa*. München: C. H. Beck.

Renze, M. & Guven, E., 2024. *The Effect of Sampling Temperature on Problem Solving in Large Language Models*. arXiv.

Reynolds, L. & McDonell, K., 2021. „Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm“. *CHI EA ‘21: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, S. 1–7.

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.

Schneider, W., 2001 [1999]. *Deutsch für Profis. Wege zu gutem Stil*. München: Goldmann.

Schuster, M. & Paliwal, K. K., 1997. „Bidirectional recurrent neural networks“. *IEEE Transactions on Signal Processing*, 45(11), S. 2673–81.

Schwiesau, D. & Ohler, J., 2003. *Die Nachricht. In Presse, Radio, Fernsehen, Nachrichtenagentur und Internet. Ein Handbuch für Ausbildung und Praxis*. Wiesbaden: Springer Fachmedien.

Seibicke, W., 1969. *Wie schreibt man gutes Deutsch? Eine Stilfibel*. Mannheim: Bibliographisches Institut / Dudenverlag.

Shen, Y., et al., 2023. „ChatGPT and Other Large Language Models Are Double-edged Swords“. *Radiology*, 307(2).

Shi, F., et al., 2022. *Language Models are Multilingual Chain-of-Thought Reasoners*. arXiv.

Straßner, E., 1981. „Sprachstrukturen“. In: *Sprache im Fernsehen. Spontan? – Konkret? – Korrekt?*. Mainz: V. Hase & Koehler, S. 169–184.

Sun, Y., et al., 2024. „AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content“. *Humanit Soc Sci Commun*, 11(1278).

Sutskever, I., et al., 2014. „Sequence to Sequence Learning with Neural Networks“. *Advances in Neural Information Processing Systems*, Band 27.

Tealab, A., 2018. „Time series forecasting using artificial neural networks methodologies: A systematic review“. *Future Computing and Informatics Journal*, 3(2), S. 334–40.

Vaswani, A., et al., 2017. „Attention Is All You Need“. *Advances in Neural Information Processing Systems*, Band 30.

Wei, J., et al., 2023. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv.

Weinrich, H., 2007. *Textgrammatik der deutschen Sprache*. 4. durchges. Aufl. Hildesheim: Olms.

White, C., et al., 2024. *LiveBench: A Challenging, Contamination-Free LLM Benchmark*. arXiv.

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.

Yang, S., 2023. *Best Practices in Prompt Engineering: Learnings and Thoughts from Andrew Ng's New Course*. [Letzter Zugriff am 19. März 2025].

Zhang, A., et al., 2023. *Dive into Deep Learning*. Cambridge: Cambridge University Press.

Zhao, W. X., Zhou, K., Li, J. & Tang, T., 2023. *A Survey of Large Language Models*. arXiv.

Zheng, L., et al., 2024. *LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset*. arXiv.

Zifonun, G., Hoffmann, L. & Strecker, B., 1997. *Grammatik der deutschen Sprache*. 3. Bände. Berlin, New York: De Gruyter.

RIFERIMENTI SITOGRAFICI

Anthropic. *https://claude.ai/* [Letzter Zugriff am 19. März 2025].

Arbeitsgemeinschaft Onlineforschung e.V. *https://www.agof.de* [Letzter Zugriff am 19. März 2025].

BILD. *https://www.bild.de/* [Letzter Zugriff am 19. März 2025].

Chatbot Arena (LMSYS). *https://lmarena.ai/* [Letzter Zugriff am 19. März 2025].

Google Books Ngram Viewer. *https://books.google.com/ngrams/* [Letzter Zugriff am 19. März 2025].

Google Trends. *https://trends.google.com/* [Letzter Zugriff am 19. März 2025].

Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern e.V. *https://www.ivw.de/* [Letzter Zugriff am 19. März 2025].

JMP. *https://www.jmp.com/* [Letzter Zugriff am 19. März 2025].

LiveBench. *https://livebench.ai/#/* [Letzter Zugriff am 19. März 2025].

Mistral AI Team. *https://mistral.ai/* [Letzter Zugriff am 19. März 2025].

Nvidia. *https://build.nvidia.com/meta/llama-3_1-405b-instruct* [Letzter Zugriff am 19. März 2025].

OpenAI. *https://chatgpt.com/* [Letzter Zugriff am 19. März 2025].

Süddeutsche Zeitung. *https://www.sueddeutsche.de/* [Letzter Zugriff am 19. März 2025].