Mensch oder Maschine? Qualitative Linguistik als Werkzeug der KI-Erkennung

Mit der breiten Verfügbarkeit großer Sprachmodelle wie ChatGPT entstehen neue Herausforderungen für die forensische Linguistik. So besteht etwa die Gefahr von Fehlschlüssen, wenn Merkmale eines KI-generierten Texts fälschlicherweise einem menschlichen Ursprung zugeordnet werden. Andersherum lässt sich auch durch tatsächliche Täter*innen leicht behaupten, sie hätten einen fraglichen Text trotz stilistischer Ähnlichkeiten nicht geschrieben, sondern eine KI habe ihren Stil imitiert (sog. *Liar's Dividend*). Und nicht zuletzt gibt es neben vollständig menschlich verfassten vs. vollständig KI-generierten Texten diverse Zwischenstufen, von absatzweise wechselnder Autorschaft bis hin zu einem reinen Lektorat durch die KI, die voraussichtlich immer mehr Verbreitung finden werden. Die Autorenerkennung muss sich daher zwingend mit dem Thema auseinandersetzen, um weiterhin gerichtsfeste Aussagen über Autorschaft treffen zu können. Auf technischer Seite gibt es zwar viele Ansätze, aber bisher noch keine Lösungen, die generierte Texte verschiedener LLMs und verschiedener (insbesondere inkriminierter) Textsorten zuverlässig von menschlichen Texten unterscheiden können (z. B. Dawkins, Fraser & Kiritchenko 2025). Die menschliche Fähigkeit zur Erkennung generierter Texte wird dagegen in Studien als sehr begrenzt angegeben, allerdings deutet einiges darauf hin, dass diese Fähigkeit gezielt gefördert und verbessert werden kann (z. B. Milička et al. 2025). Dabei wird in der bisherigen Forschung aber nur selten qualitativ-linguistische Expertise einbezogen. Das Poster stellt eine Pilotstudie vor, in der forensisch-linguistische Expert*innen und andere BKA-Mitarbeitende 20 Texte auswerteten, von denen elf KI-generiert waren. Gefragt waren eine Einschätzung zum Ursprung (Mensch oder KI), Argumente für beide Seiten sowie eine Angabe zur empfundenen Sicherheit bei der Zuordnung. Die Ergebnisse deuten auf klare Vorteile forensisch-linguistischer Expertise und Methodik hin, sowohl bezüglich der Trennleistung als auch bezüglich der Zielsicherheit bei der Selbsteinschätzung und im Detailreichtum der Argumentation.

Literatur

Dawkins, Hillary, Kathleen C. Fraser & Svetlana Kiritchenko (2025): When Detection Fails: The Power of Fine-Tuned Models to Generate Human-Like Social Media Text. arXiv preprint V2, https://arxiv.org/abs/2506.09975.

Milička, J., Marklová, A., Drobil, O., & Pospíšilová, E. (2025). Humans can learn to detect AI-generated texts, or at least learn when they can't. arXiv preprint V1, https://arxiv.org/abs/2505.01877v3.