

Jenna Koenen¹
Marvin Rost¹
Ines Sonnenschein²
Ruikun Hou¹
Babette Bühler¹
Enkelejda Kasneci¹

¹Technische Universität München
²FH Münster

Think-Aloud trifft NLP – Prozessanalyse im chemischen Lehlabor

Einführung

Das Labor gilt als Herzstück des Chemielernens, da hier domänenspezifische Arbeitsweisen praktisch erfahren und reflektiert werden (Agustian et al., 2022; Seery, 2020). Eine zentrale Rolle spielt dabei das Experimentieren, das als grundlegende Säule chemiebezogener Bildung gilt (Koenen & Tiemann, 2018). Wesentliches Merkmal experimentellen Arbeitens ist die Anwendung von Variablenkontrollstrategien (VKS), die es Lernenden ermöglichen, systematisch Ursache-Wirkungs-Zusammenhänge zu untersuchen (Schwchow et al., 2016; Schwchow & Nehring, 2018). Sie ist eng mit naturwissenschaftlichem Denken und Arbeiten, also mit Scientific Inquiry (SI), verknüpft – einem Prozess, der Praktiken wie Hypothesenbildung, Untersuchungsplanung, Messung, Dateninterpretation, Evaluation und metakognitive Regulation umfasst (Pedaste et al., 2015). Empirische Befunde zeigen, dass Lernprozesse im Labor individuell und nicht-linear verlaufen: Studierende kombinieren unterschiedliche Herangehensweisen und wechseln flexibel zwischen ihnen (Kranz et al., 2022; Sonnenschein, 2019). Um diese Vielfalt gezielt zu adressieren, sind Verfahren erforderlich, die individuelle Lernwege zeitnah analysieren können. Während klassische qualitative Inhaltsanalysen tiefgehende Einblicke liefern, sind sie zugleich zeit- und ressourcenintensiv. Hier eröffnen Verfahren des Natural Language Processing (NLP) und Machine Learning (ML) möglicherweise neue Möglichkeiten einer skalierbaren Analyse. Im Rahmen des Symposiums „Maschinelles Lernen in der chemiedidaktischen Forschung“ (Martin & Rost, dieser Tagungsband) widmete sich der vorliegende Beitrag mit einer Re-Analyse von Think-Aloud-Daten bei problemorientierten Experimentieraufgaben den Schwerpunkten der *Konstruktvalidierung* und der *kontinuierlichen Modellierung von Lernprozessen*.

Theoretischer Hintergrund

Scientific Inquiry wird häufig als zyklischer oder semi-zyklischer Prozess mit aufeinanderfolgenden Phasen, d.h. im Sinne eines Teilschrittansatzes beschrieben (Pedaste et al., 2015). Zwar sind verschiedene Frameworks vielschichtig und unterschiedlich in ihren konkreten Ausgestaltungen (Rönnebeck et al., 2016). Einerseits zeigen sich für empirische Untersuchungen durchaus generalisierbare Muster (Reith & Nehring, 2020). Andererseits ist die empirische Lage, gerade bei der Erfassung experimenteller Fähigkeiten und speziell für Validitätsbefunde in Assessments, äußerst divers (Schreiber & Gut, 2022). Re-Analysen können in diesem Zusammenhang hilfreich sein (Morrison, 2020), weil sie die Möglichkeit zur Revision und Erweiterung von Erkenntnissen bieten. Jedoch sind sie im Bildungsbereich bisher eher selten (Perry et al., 2022) und stellen damit ein Desiderat für die naturwissenschaftsdidaktische Forschung dar. Insbesondere die Replikation von inhaltsanalytischen Studien war bis vor kurzem allerdings praktisch nicht empirisch abbildbar

(Taylor et al., 2016), weil die Ressourcen für derartige Untersuchungen nicht zur Verfügung standen. Dies ändert sich mit zunehmender Geschwindigkeit durch das Aufkommen von neuen Werkzeugen aus dem Bereich NLP und ML (Kubsch, Krist, et al., 2022; Martin et al., 2023; Nelson, 2020; Rost et al., 2025).

Forschungsfragen

Mit dem Ziel des Vergleichs einer menschlichen qualitativen Inhaltsanalyse von Experimentierprozessen und eines NLP/ML-Ansatzes, ergeben sich die folgenden Fragestellungen.

FF1: Kann die Kategorisierung von Think-Aloud-Protokollen bei Titrationsexperimenten durch NLP-Methoden und ML repliziert werden?

FF2: Inwiefern trägt eine Modellierung von Verbaldaten zu einem vertieften Verständnis von Scientific Inquiry beim Experimentieren bei?

Design & Methode

In der Primärdatenerhebung wurden Studierende ($N = 14$) bei je drei kontextualisierte, problemorientierten Titrationsaufgaben zum Lauten Denken aufgefordert. Damit ergaben sich 42 Transkripte aus Gehaltsbestimmungen von Ascorbinsäure in Orangensaft, von Tanninen in schwarzem Tee sowie der Peroxidzahl in Speisefett. Die qualitative Inhaltsanalyse nach Mayring (2022) erfolgte durch ein Kodiermanual (Sonnenschein et al., 2023) mit 55 Hauptkategorien über den Verlauf der Think-Aloud-Protokolle. Die hier berichtete Re-Analyse mittels NLP/ML erfolgte durch Word Embeddings mit einem vortrainierten allgemeinen Sprachmodell (bert-base-german-uncased, Bayerische Staatsbibliothek, 2025) und der Anwendung eines Random Forest Classifiers (Breiman, 2001). Der Datensatz wurde so aufgeteilt, dass die Transkripte von 12 Studienteilnehmenden zum Training des Algorithmus auf die 55 Kategorien sowie die Transkripte von 2 Teilnehmenden für das Testen der Vorhersagegüte mit den Metriken F1 (macro), klassisches κ_{Cohen} und Accuracy inklusive Signifikanztest gegen die No-Information-Rate (Jurafsky & Martin, 2023) verwendet wurden.

Ergebnisse

Die maschinelle Klassifikation erreicht global $F1(\text{macro}) = .49$, $\text{Accuracy} = .49$ und $\kappa_{\text{Cohen}} = .44$; die Accuracy liegt signifikant über der No-Information-Rate ($p < .001$). Diese vergleichsweise niedrigen Übereinstimmungsmaße schwankten jedoch stark über die Kategorien. Die Abbildungen 1 und 2 visualisieren exemplarisch den Vergleich von menschlicher und maschineller Klassifikation des Test-Datensatzes für zwei Codes ("Interpretation und Auswertung", "Durchführung").

Ausblick

Die in Entwicklung befindliche In-Prozessanalyse durch moderne computationale Verfahren kann dazu beitragen, individuelle Herausforderungen und Stärken der Lernenden zu identifizieren. Im vorliegenden Setting wurde von sprachlicher Manifestation auf kognitive Prozesse geschlossen. Die Erweiterung um multi-modale Verfahren könnte zu einer weiteren Validierung beitragen (Kubsch, Caballero, et al., 2022) und iterative Workflows im Wechselspiel zwischen Mensch und Maschine können in der Folge eine tiefergehende Auseinandersetzung und Theorieentwicklung in den Naturwissenschaftsdidaktiken anstoßen.

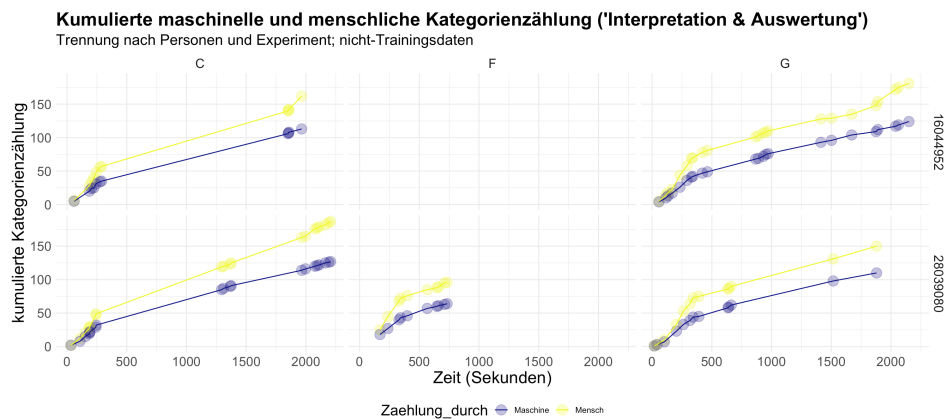


Abb. 2: Über den Verlauf der Think-Aloud-Protokolle aufgetragene, kumulierte Kategorisierungszählungen (blau: maschinell; gelb: menschlich; Zahlen am rechten Rand sind Personencodes; Buchstaben am oberen Rand repräsentieren die drei Experimente). Die Zählungen weichen systematisch voneinander ab: Die maschinelle Kategorisierung ordnet den Code „Interpretation und Auswertung“ grundsätzlich seltener zu als die menschlichen Coder.

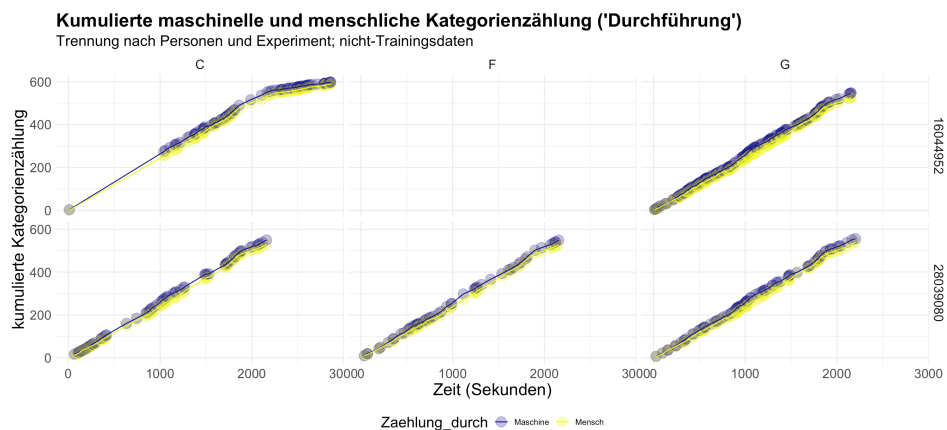


Abb. 1: Analoge Darstellung des Mensch-Maschine-Vergleichs für die Zählungen der Kategorie „Durchführung“. Die maschinelle Zuordnung fällt sehr eindeutig mit der menschlichen Zuordnung zusammen.

Literatur

- Bayerische Staatsbibliothek. (2025). Bert-base-german-uncased (Revision b705f0e). Hugging Face. <https://doi.org/10.57967/hf/4378>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Jurafsky, D. & Martin, J. H. (2023). *Speech and language processing. An introduction to natural language processing, computational linguistics, and speech recognition* (3. Aufl.). <https://web.stanford.edu/~jurafsky/slp3/>

- Kranz, J., Baur, A. & Möller, A. (2022). Learners' challenges in understanding and performing experiments: A systematic review of the literature. *Studies in Science Education*, 1–47. <https://doi.org/10.1080/03057267.2022.2138151>
- Kubsch, M., Caballero, D. & Uribe, P. (2022). Once more with feeling: Emotions in multimodal learning analytics. In M. Giannakos, D. Spikol, D. Di Mitri, K. Sharma, X. Ochoa, & R. Hammad (Hrsg.), *The Multimodal Learning Analytics Handbook* (S. 261–285). Springer International Publishing. https://doi.org/10.1007/978-3-031-08076-0_11
- Kubsch, M., Krist, C. & Rosenberg, J. M. (2022). Distributing epistemic functions and tasks—A framework for augmenting human analytic power with machine learning in science education research. *Journal of Research in Science Teaching*, tea.21803. <https://doi.org/10.1002/tea.21803>
- Martin, P. P., Kranz, D., Wulff, P. & Graulich, N. (2023). Exploring new depths: Applying machine learning for the analysis of student argumentation in chemistry. *Journal of Research in Science Teaching*, tea.21903. <https://doi.org/10.1002/tea.21903>
- Mayring, P. (2022). *Qualitative Inhaltsanalyse: Grundlagen und Techniken* (13., überarbeitete Auflage). Beltz.
- Morrison, K. (2020). Realizing the promises of replication studies in education. *Educational Research and Evaluation*, 25(7–8), 412–441. <https://doi.org/10.1080/13803611.2020.1838300>
- Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, 49(1), 3–42. <https://doi.org/10.1177/0049124117729703>
- Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., Manoli, C. C., Zacharia, Z. C. & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, 14, 47–61. <https://doi.org/10.1016/j.edurev.2015.02.003>
- Perry, T., Morris, R. & Lea, R. (2022). A decade of replication study in education? A mapping review (2011–2020). *Educational Research and Evaluation*, 27(1–2), 12–34. <https://doi.org/10.1080/13803611.2021.2022315>
- Reith, M. & Nehring, A. (2020). Scientific reasoning and views on the nature of scientific inquiry: Testing a new framework to understand and model epistemic cognition in science. *International Journal of Science Education*, 42(16), 2716–2741. <https://doi.org/10.1080/09500693.2020.1834168>
- Rönnebeck, S., Bernholt, S. & Ropohl, M. (2016). Searching for a common ground—A literature review of empirical research on scientific inquiry activities. *Studies in Science Education*, 52(2), 161–197. <https://doi.org/10.1080/03057267.2016.1206351>
- Rost, M., Resch, K. & Lembens, A. (2025). Using computational grounded theory to analyze pre-service chemistry teachers' reflective practice regarding technology integration in classrooms within a service-learning-oriented seminar. *Journal of Science Education and Technology*. <https://doi.org/10.1007/s10956-025-10236-x>
- Schreiber, N. & Gut, C. (2022). Die Validierungspraxis bei hands-on Experimentiertests in der naturwissenschaftsdidaktischen Forschung: Ein Review auf Basis eines allgemeinen Ordnungssystems für Validierungsstudien. *Zeitschrift für Didaktik der Naturwissenschaften*, 28(1), 8. <https://doi.org/10.1007/s40573-022-00145-6>
- Sonnenschein, I. (2019). *Naturwissenschaftliche Denk- und Arbeitsprozesse Studierender im Labor* (Bd. 289). Logos.
- Sonnenschein, I., Koenen, J. & Tiemann, R. (2023). Kodiermanual zur Beschreibung naturwissenschaftlicher Erkenntnisgewinnungsprozesse Studierender im Labor. <https://doi.org/10.5281/ZENODO.7896850>
- Taylor, J., Furtak, E., Kowalski, S., Martinez, A., Slavin, R., Stuhlsatz, M. & Wilson, C. (2016). Emergent themes from recent research syntheses in science education and their implications for research design, replication, and reporting practices. *Journal of Research in Science Teaching*, 53(8), 1216–1231. <https://doi.org/10.1002/tea.21327>