

Abstract:

gde 21 – geschichtsdidaktik empirisch 21

KI im Fach Geschichte – Wie künstliche Intelligenz bei der inhaltlichen und sprachlichen Beurteilung von Schülerantworten genutzt werden kann.

Historisches Denken ist mehr als eine Ansammlung chronologisch erzählter Daten und Fakten. Stattdessen sollen Schüler*innen im Geschichtsunterricht lernen, mit vorgegebenen Narrativen (z.B. in Schulbüchern) kritisch umzugehen (De-Konstruktion), sowie auf Fragen an die Vergangenheit quellen- und evidenzbasierte Antworten zu geben (Re-Konstruktion). Historisches Denken ist also untrennbar mit der Sprache verbunden. Zur Vermeidung einer *construct underrepresentation* (Messick, 1995) werden die Kompetenzen historischen Denkens daher häufig in offenen Aufgabenformaten erfasst (Barricelli, 2005; Handro & Schönemann, 2010; Hartung, 2013; Hodel et al., 2013; Mierwald & Brauch, 2015; Nitsche & Waldis, 2016; Smith, Breakstone & Wineburg, 2018; VanSledright, 2014). Um eine objektive, reliable und valide Beurteilung der Texte sicherzustellen, sollten diese von mindestens zwei Personen bewertet werden, ein Verfahren, das in groß angelegten Studien mit einem hohen Ressourcenaufwand verbunden ist.

In einem interdisziplinären Projekt von Geschichtsdidaktik, Computerlinguistik und Empirischer Bildungsforschung gingen wir deshalb der Frage nach, ob und wie künstliche Intelligenz in Form computerlinguistischer Methoden für die Auswertung von Schülerantworten im Fach Geschichte genutzt werden kann. Da mit Hilfe der computerlinguistischen Ansätze *Automatic Content Assessment* und *Linguistic Complexity Assessment* die inhaltliche Richtigkeit respektive die sprachliche Komplexität von Texten automatisiert bewertet werden kann, adressierten wir folgende Forschungsfragen:

- (1) Kann mit Hilfe von Automatic Content Assessment die Richtigkeit von Schülerantworten automatisiert beurteilt werden?
- (2) Kann mit Hilfe von Linguistic Complexity Assessment die sprachliche Komplexität der Schülerantwort, die der inhaltlichen Komplexität der Aufgabe entsprechen sollte, automatisiert beurteilt werden?

Die Studie nutzt die Daten einer Interventionsstudie ($N = 962$) im Geschichtsunterricht, in der Schüler*innen im Post-Test basierend auf drei historischen Dokumenten sieben offene Aufgaben bearbeiteten, die sich hinsichtlich ihrer Komplexität und des damit verbundenen inhaltlichen und sprachlichen Anspruchs substantiell voneinander unterschieden. Der verwendete Datensatz besteht aus $N = 141$ randomisiert gezogenen Schülerantworten auf sieben Fragen, die von zwei Ratern beurteilt wurden. Die Interrater-Reliabilität war mit einem Range von $.75 < \kappa < .96$ (*Weighted Cohen's Kappa*) zufriedenstellend bis sehr gut.

Generell gilt für computerlinguistische Methoden, dass ein Text anhand vordefinierter sprachlicher Merkmale quantifiziert wird, indem die Ausprägung dieser Merkmale in Zahlen übertragen werden, die in eine mathematische Relation zur menschlichen Bewertung, die als *Gold Standard* dient, gestellt werden. Die so erlernte Relation erlaubt es, neue Texte

automatisch auf Basis ihres Merkmalsvektors zu bewerten. Ob die automatisierte Vorhersage zutrifft, wird überprüft, indem die von Menschen annotierten Texte in Trainings- und Testdaten geteilt werden. Mit den Trainingsdaten lernt der Computer die jeweils richtigen Muster, auf den Testdaten wird die Bewertung der Antworten automatisiert vorhergesagt und mit dem manuellen Rating verglichen.

Die Content Analysen wurden mit dem *CoMiC*-System (Comparing Meaning in Context, Meurers et al., 2011) berechnet. In den Aufgaben 1 und 5 stimmte der Computer mit beiden manuellen Ratings in einem fast perfekten Range ($\kappa \geq .8$) überein. Auch in den Aufgaben 6 und 7, in denen eigene Überlegungen angestellt und Schlussfolgerungen gezogen werden mussten, war die Übereinstimmung zwischen dem Computer und den Ratern substantiell hoch (A6: Interrater-Reliabilität: .83; CoMiC und Rater 1: .63; CoMiC und Rater 2: .68 / A7: Interrater-Reliabilität .88, CoMiC und Rater 1: .79; CoMiC und Rater 2: .66).

In den Komplexitätsanalysen wurde ein System mit 295 linguistische Merkmalen verwendet (Weiss & Meurers, 2018; Kühberger et al., 2019). Die automatisierte Vorhersage der Komplexität der Aufgabe durch die beobachtete sprachliche Komplexität der Schülerantwort entsprach der manuellen Klassifikation der Aufgabenkomplexität deutlich häufiger, als zufällig zu erwarten wäre: Mit 85,37% stimmte der Computer mit der manuellen Klassifikation überein (34,12% wären zufällig richtig gewesen). Dabei spiegelten die zum Einsatz kommenden sprachlichen Merkmale die Komplexität der Aufgabe weitgehend wider.

Anhand der ermutigenden Ergebnisse werden die Potenziale und Grenzen der computerlinguistischen Methoden für die empirische Forschung in der Geschichtsdidaktik diskutiert.

594 Wörter

Literatur

Barricelli, M. (2013). *Schüler erzählen Geschichte. Narrative Kompetenz im Geschichtsunterricht*. Schwalbach/Ts.: Wochenschau Verlag.

Handro, S. & Schönemann, B. (Hrsg.) (2010). *Geschichte und Sprache*. (Zeitgeschichte – Zeitverständnis, Bd. 20). Berlin: LIT.

Hartung, O. (2013). *Geschichte schreiben lernen. Empirische Erkundigungen zum konzeptionellen Schreibhandeln im Geschichtsunterricht*. Berlin: LIT.

Hodel, J. (2013). Schülernarrationen als Ausdruck historischer Kompetenz. *Zeitschrift für Didaktik der Gesellschaftswissenschaften* 2 (2013), 121-145.

Körber, A., Schreiber, W. & Schöner, A. (Hrsg.). (2007). *Kompetenzen historischen Denkens. Ein Strukturmodell als Beitrag zur Kompetenzorientierung in der Geschichtsdidaktik*. Neuried: ars una Verlagsgesellschaft.

Kühberger, C., Bramann, C. Meurers, D. and Weiss, Z. L. (2019, in press). Task Complexity in History Textbooks. A Multidisciplinary Case Study on Triangulation in History Education Research. Mixed Methods and Triangulation in History Education Research by the *International Journal of Historical Learning, Teaching and Research*.

- Meurers, D., Ziai, R., Ott, N. and Kopp, J. (2011). Evaluating answers to reading comprehension questions in context: Results for German and the role of information structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment Edinburgh*, S. 1-9.
- Mierwald, M. & Brauch, N. (2015). Historisches Argumentieren als Ausdruck historischen Denkens. Theoretische Fundierung und empirische Annäherungen. *Zeitschrift für Geschichtsdidaktik* 14, S. 104-120.
- Nitsche, M. & Waldis, M. (2016). Narrative Kompetenz von Studierenden erfassen – Zur Annäherung an formative und summative Vorgehensweisen im Fach Geschichte. In: *Zeitschrift für Didaktik der Gesellschaftswissenschaften* 7(1), 17-35.
- Schönemann, B., Thünemann, H. & Zülsdorf-Kersting, M. (2010). *Was können Abiturienten? Zugleich ein Beitrag zur Debatte über Kompetenzen und Standards im Fach Geschichte*. Berlin: LIT.
- Smith, B., Breakstone, J. & Wineburg, S. (2019). History assessments of thinking: A validity study. *Cognition and Instruction* 37(1). DOI: 10.1080/07370008.2018.1499646.
- Trautwein, U., Bertram, C., Borries, B. von, Brauch, N., Hirsch, M., Klausmeier, K., ..., Zuckowski, A. (2017). *Kompetenzen historischen Denkens erfassen – Konzeption, Operationalisierung und Befunde des Projekts „Historical Thinking – Competencies in History“ (HiTCH)*. Münster: Waxmann-Verlag.
- VanSledright, B. A. (2014). *Assessing historical thinking & understanding. Innovative designs for new standards*. New York: Routledge.
- Weiss, Z. L. & Meurers, D. (2019, in Druck). Broad Linguistic Modeling is Beneficial for German L2 Proficiency Assessment'. In A. Abel, A. Glaznieks, V. Lyding, and L. Nicolas (Hrsg.), *Widening the Scope of Learner Corpus Research. Selected Papers from the 4th Learner Corpus Research Conference 2017*. Louvain-La-Neuve: Presses Universitaires de Louvain.