SLLD SPRACHLICH-LITERARISCHES LERNEN UND DEUTSCHDIDAKTIK

Nadine Anskeit, Dirk Betzel, Marco Ennemoser & Martin Fix

Die Qualität von Texten Jugendlicher im diachronen und digitalen Wandel untersuchen – Pilotstudie zum Ludwigsburger Aufsatzkorpus 2.0 („LuKo 2.0“)

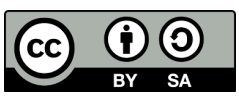
Investigating the quality of learner texts in diachronic and digital change – pilot study on the Ludwigsburg Essay Corpus 2.0

Abstract: In einer Pilotstudie wurde untersucht, wie sich die Textproduktionsleistungen von Schüler/-innen achter Klassen beim Schreiben zu einem Bildimpuls zwischen 1998 und 2023 verändert haben. Ferner wurde der Einfluss des Schreibmediums (handschriftlich vs. digital) und die Eignung von KI zur Textqualitätsbewertung überprüft. Die Ergebnisse, die sich ausschließlich auf narrative Texte beziehen, zeigen, dass handschriftliche Texte aus 2023 im Vergleich zu 1998 schlechter bewertet wurden, während digitale Texte qualitativ vergleichbar und sogar länger waren. Die Studie legt nahe, dass eine Verschlechterung der Schreibfähigkeiten bei narrativen Texten aus diachroner Perspektive nicht pauschal zutrifft, vielmehr scheinen Schreibfähigkeiten in Abhängigkeit zum Schreibmedium zu stehen. Limitierend sind die geringe Stichprobengröße und mögliche bildungsbezogene Unterschiede. In der geplanten Hauptstudie sind diese Faktoren zu kontrollieren. Der Einbezug von ChatGPT-Textqualitätsurteilen, ergänzend zu den Urteilen menschlicher Raterinnen, führte zu vergleichbar guten Reliabilitätsschätzungen, wenngleich die paarweisen Übereinstimmungen zwischen menschlichen Raterinnen höher ausfielen als zwischen Mensch und KI.

Keywords: Narration, Textproduktionsleistungen, Schreibmedium (handschriftlich vs. digital), KI-basierte Textqualitätsbewertung, Schreibfähigkeiten, diachroner Wandel

Abstract: A pilot study examined changes in the text production performance of eighth-grade students when writing in response to a visual prompt, depending on the time of data collection (1998 and 2023), as well as the influence of the writing medium (handwritten vs. digital) and the suitability of AI for text quality assessment. The results, which pertain exclusively to narrative texts, show that handwritten texts from 2023 were rated lower compared to those from 1998, while digital texts were qualitatively comparable and even longer. The study suggests that a deterioration in writing skills for narrative texts from a diachronic perspective is not universally applicable; rather, writing skills seem to depend on the writing medium. Limitations include the small sample size and potential educational disparities. These factors are to be controlled in the planned main study. The inclusion of ChatGPT text quality assessments, alongside those of human raters, yielded comparably good reliability estimates, although the pairwise agreements among human raters were higher than those between humans and AI.

Keywords: Narration, Text Production Performance, Writing Medium (Handwritten vs. Digital), AI-Based Text Quality Evaluation, Writing Skills, Diachronic Change



Zeitschrift für Sprachlich Literarisches Lernen und Deutschdidaktik 1 (2025)
veröffentlicht am 05.03.2025
https://doi.org/10.46586/SLLD.Z.2025.12036

Gefördert durch DFG Deutsche Forschungsgemeinschaft

1 | Ausgangslage und Ziele der Pilotstudie und der geplanten Hauptstudie

„Heutzutage hat jeder einen anderen Style oder auch Charakter. Kein Mensch ist gleich. Es kann noch so viele Ähnlichkeiten geben und doch ist alles anderst." (Achtklässlerin „Hope", 2023)

Was „Hope" (Deckname) hier in ihrem Text schreibt, lässt sich auch einer diachronen Studie voranstellen: Haben die Schüler/-innen heute einen anderen Schreibstil als vor 25 Jahren? Die Klage, dass sich ihre schriftsprachlichen Leistungen permanent verschlechterten, ist allgegenwärtig. Untersuchungen zur Veränderung schriftsprachlicher Leistungen von Lernenden fokussieren jedoch vorwiegend formalsprachliche Merkmale wie z. B. die Rechtschreibung (IQB Bildungstrends) oder die Kommasetzung (Berg & Romstadt 2021). Tiefenstrukturelle Textproduktionsleistungen sind hingegen kaum Gegenstand diachron angelegter Vergleichsuntersuchungen, und wenn, dann vor allem bezogen auf Abituraufsätze (vgl. Sieber 1998, Grimm 2003; eine Ausnahme für die Grundschule bilden Steinig et al. 2009). Während die Qualität formalsprachlicher Leistungen abzunehmen scheint, bleibt mit Blick auf Textproduktionsleistungen weitgehend offen, ob und ggf. in welche Richtung sich diese in den letzten Jahrzehnten verändert haben könnten und wohin die weitere Entwicklung geht.

Daher soll der Blick unserer Studie auf Veränderungen in der Textqualität aus diachroner Perspektive gerichtet werden. Hierzu kann auf eine Schreibaufgabe eines in den Jahren 1997-1998 (kurz: 1998) an der Pädagogischen Hochschule Ludwigsburg entstandenen Forschungsprojekts zurückgegriffen werden. Damals wurden in 24 achten Klassen aus der Region Stuttgart insgesamt rund 2300 Aufsätze geschrieben, die als „Ludwigsburger Aufsatzkorpus" (kurz: LuKo) publiziert vorliegen (Fix & Melenk 2000). Aus LuKo wurde nun die Aufgabe „Schreiben zu Bildimpulsen" für eine Replikation unter erweiterter Fragestellung herangezogen, um Textproduktionsleistungen von 1998 und 2023 zu vergleichen. Zur Vorbereitung einer größer angelegten Hauptstudie wurden im Jahr 2023 in einer Pilotstudie rund 150 Texte von Schüler/-innen der achten Jahrgangsstufe unter ähnlichen unterrichtlichen Bedingungen wie vor 25 Jahren erhoben. Im Zentrum der Pilotstudie sowie der geplanten Hauptstudie steht die Frage, ob sich aus diachroner Perspektive die Leistung, einen kohärenten, funktional angemessenen Text zu einer Schreibaufgabe zu verfassen, verändert hat. Die Variable *Textqualität* wird somit unter Berücksichtigung des Faktors *Erhebungszeitpunkt* (1998 vs. 2023) beleuchtet.

Ein wesentlicher Unterschied zwischen dem Erhebungszeitpunkt 1998 und heute besteht in der Verfügbarkeit digitaler Schreibmedien. Während das Schreiben 1998 schulisch und privat in erster Linie handschriftlich geprägt war und der Umgang mit Textverarbeitungsprogrammen im Unterricht eher zu den Ausnahmen zählte, haben sich die Ausgangsbedingungen inzwischen grundlegend verändert: Nahezu alle Schüler/-innen nutzen außerhalb der Schule digitale Medien für das (interaktionsorientierte) Schreiben, wohingegen digitales Schreiben im Unterricht zwar zunehmend an Bedeutung gewinnt, jedoch noch keine Selbstverständlichkeit zu sein scheint (ICILS 2019, S. 19 ff.).

Die digitale Transformation wird in der öffentlichen Debatte häufig als zentrale Triebkraft für eine angenommene Verschlechterung schriftsprachlicher Leistungen ausgemacht. Es dürfte auch kein Zweifel daran bestehen, dass sich das in der Medientheorie schon früh beschworene „Ende der Gutenberg-Galaxis" (Bolz 1993) im Sprachgebrauch der jüngeren Generationen manifestiert. Aber die Annahme, dass es sich dabei zwingend um eine Verschlechterung handeln soll, könnte auch durch die bekannte Generationenfalle beeinflusst sein. Diese besagt, dass die ältere Generation sich ändernde Normen als Verlust von Vertrautem empfindet. Von

dieser zumeist öffentlich geführten Debatte ist der fachliche Diskurs zu unterscheiden, indem eher die Potenziale digitalen Schreibens hervortreten. Unter der Voraussetzung einer entsprechenden Flüssigkeit beim Schreiben mit Tastatur oder Touchpad wird die höhere Schreibgeschwindigkeit im Vergleich zum motorisch anspruchsvolleren Handschreiben betont, zugleich können Korrekturprogramme entlastend auf hierarchieniedrigere Prozesse wirken und einfachere Möglichkeiten der Textrevision die Prozessorientierung des Schreibens unterstützen. Unabhängig davon knüpft digitales schulisches Schreiben stärker an die Lebenswelt heutiger Jugendlicher an und könnte deshalb auch zu einer höheren Motivation beitragen, Überarbeitungen am Text vorzunehmen.
Daraus lassen sich für das Untersuchungsdesign Konsequenzen ableiten: Um Aussagen über mögliche Veränderungen zur Textqualität aus diachroner Perspektive treffen zu können, genügt es nicht, das 1998 durchgeführte Forschungsprojekt nur zu wiederholen, vielmehr ist das analoge Schreibsetting um ein digitales zu erweitern. Die Variable *Textqualität* wird somit zusätzlich in Abhängigkeit zum *Schreibmedium* in den Blick genommen. Durch die Erweiterung um das digitale Schreiben werden zum einen – aus synchroner Perspektive – Erkenntnisse gewonnen, ob Jugendliche mit der Hand qualitativ andere Texte als digital schreiben – die bisherigen Forschungsbefunde dazu fallen recht unterschiedlich aus (Rödel 2020, S. 73 ff.). Zum anderen kann aus diachroner Perspektive überprüft werden, ob mögliche Veränderungen in der Textqualität bei handschriftlich verfassten Texten (1998 vs. 2023) auch im Vergleich zu den digital verfassten festzustellen sind.
Zu beiden Fragestellungen können im Rahmen der Pilotstudie lediglich Indizien für Veränderungen identifiziert werden. Sie können aber zur Bildung von Hypothesen für die geplante Hauptstudie genutzt werden, bei der mögliche intervenierende Variablen (wie z. B. die Konstellation der Probandengruppe) stärker kontrolliert werden können. Übergeordnetes Ziel ist es, den sich (möglicherweise) vollziehenden Wandel der Textproduktionsleistungen Jugendlicher empirisch genauer zu untersuchen und zu beschreiben, bevor die Deutschdidaktik Antworten darauf finden kann, wie sie sich daraufhin konzeptionell und methodisch ausrichtet.
Neben den zuvor erläuterten Forschungsfragen zur Veränderung von Textproduktionsleistungen möchten wir zusätzlich die Reliabilität KI-basierter Textbewertungen überprüfen, indem wir diese mit den Bewertungen menschlicher Raterinnen vergleichen. Im vorliegenden Beitrag steht die Qualität KI-basierter Beurteilung im Vordergrund. Perspektivisch interessieren in der Hauptstudie aber auch die damit verbunden Möglichkeiten einer prozessbezogenen Unterstützung in Form von Rückmeldungen zu Texten Lernender.

2 | Zum Forschungsstand

2.1 | Textqualität in diachronen Vergleichsuntersuchungen und digitale Transformation

Im Mittelpunkt der vorliegenden Pilotstudie als auch der geplanten Hauptstudie steht das Konstrukt „Textqualität“. Grundlegend kann *Textqualität* danach bemessen werden, ob ein Text mit Blick auf das gesetzte Handlungsziel und die anvisierten Adressaten kommunikativ angemessen ist. Diese funktionale Bestimmung impliziert bereits solche Kriterien, die in den Bildungsstandards für den Mittleren Schulabschluss zentral angeführt werden: Adressatenorientierung, Textkohärenz, Schreibfunktion, stilistische/formale Stimmigkeit (KMK

2022, S. 20 ff.); gleichermaßen sind sie Bestandteil verschiedener Definitionen von Schreibkompetenz (u.a. Fix 2025, S. 33f.).[1]
Ein differenziertes Modell zur Textqualität liegt mit dem *Zürcher Textanalyseraster* (Nussbaumer & Sieber 1995) vor, das sowohl in aktuellen wissenschaftlichen Untersuchungen (u. a. Grabowski 2022) als auch – modifiziert (Becker-Mrotzek & Böttcher 2014) – in schulischen Kontexten Anwendung findet. Es kann als grundlegender Bezugsrahmen zur Bestimmung von Textqualität aufgefasst werden, in dem angenommene Textqualitätsmerkmale fünf Dimensionen zugeordnet werden:

- *Grundgrößen* (z. B. Textlänge oder Wortschatz)
- *Sprachformale Korrektheit*
- *Angemessenheit*: *Verständlichkeit/Kohärenz*
- *ästhetische Angemessenheit*
- *inhaltliche Relevanz/Wagnis*

Diese Grundgrößen bieten einen Rahmen zur Einordnung der bislang vorliegenden empirischen Forschungsergebnisse:

Die Textlänge, gemessen an der Anzahl der Wortformen (tokens), zählt zu den Grundgrößen und wird in vielen Untersuchungen als erster Indikator für Textqualität erhoben (Grabowski 2022, S. 263). Betrachtet man die Ergebnisse diachroner Vergleichsuntersuchungen von Abituraufsätzen (Sieber 1998; Grimm 2003; Berg & Romstadt 2021), ergibt sich hierzu einheitlich ein Anstieg der durchschnittlichen Textlänge zwischen den 1970er und 1990er Jahren. Dass die Zunahme nicht ausschließlich auf Abiturarbeiten und damit auf eine besondere Gruppe von Schreiber/-innen beschränkt ist, konnten Steinig & Betzel (2014) an Texten von Grundschulkindern nachweisen (Zunahme der durchschnittliche Textlänge von 75,7 Wörtern im Jahr 1972 auf 105,6 Wörtern im Jahr 2002). Geht man davon aus, „dass die geschriebene Textlänge einen guten Indikator zur Erklärung von Textqualität darstellt“ (Neumann 2012, S. 78)[2], lassen die genannten Ergebnisse eher auf eine zunehmende Textqualität in jüngeren Vergleichsgruppen schließen. Auch die lexikalische Vielfalt eines Textes als weitere Grundgröße korreliert Studien zufolge positiv mit Textqualität (Mathiebe 2018, S. 183 f.). Im diachronen Vergleich ermittelte Grimm (2003) in den neueren Abituraufsätzen einen etwas geringeren Wert, Betzel & Steinig (2016) konnten hingegen in den Grundschultexten von 2002 im Vergleich zu 1972 einen signifikanten Anstieg der lexikalischen Vielfalt feststellen. Insgesamt deuten die unter der Dimension *Grundgrößen* erfassten statistischen Maße auf keine abnehmende Textqualität hin, vielmehr lässt die durchgängig festgestellte Textlängenzunahme eher Gegenteiliges vermuten.

Ein anderes Bild ergibt sich für die zweite Dimension, die *sprachformale Korrektheit*. Sie scheint bereits bei Abiturarbeiten in neueren Texten zurückzugehen, allerdings auf nach wie vor recht hohem Niveau und bei hoher Varianz, wie Berg & Romstadt (2021) am Beispiel der Interpunktion hervorheben. Bei Viertklässler/-innen hat sich die Anzahl der Rechtschreibfehler auf 100 Wörter im Text von 1972 bis 2012 mehr als verdoppelt (Steinig & Betzel 2014, S. 362). In eine ähnliche Richtung deuten die aktuellen IQB-Trendanalysen für die Primar-und die Sekundarstufe, wenngleich die Rechtschreibung hier nicht textbezogen erhoben wurde und

[1] Da nur jeweils ein Text zu einer Schreibaufgabe erhoben wurde, sollen hier keine generellen Aussagen zur Schreibkompetenz getroffen werden. Analysiert wird vielmehr die Qualität einer punktuellen Leistung (Schoonen, S. 2012; Grabowski 2022, S. 136).

[2] Für eine differenzierte Betrachtung des Zusammenhangs von Textlänge und Textqualität mit Blick auf Elaborationen und Kontextualisierungen verweisen wir auf Pander Maat et al. (2023).

lediglich Aussagen über einen Zeitraum von rund zehn Jahren getroffen werden können. Insgesamt zeigt sich die Tendenz, dass die sprachformale Korrektheit in neueren Texten abzunehmen scheint. Welcher Zusammenhang zur Textqualität besteht, wenn man Orthographie als Service für Leser/-innen begreift, muss zunächst offenbleiben.

Zu den drei weiteren Dimensionen der Textqualität liegen für den deutschsprachigen Raum bislang keine diachronen Vergleichsstudien vor. Einige Hinweise finden sich jedoch in der Arbeit von Sieber (1998). Unter dem Begriff *Parlando* führt er ein Merkmalsbündel textueller Auffälligkeiten auf verschiedenen linguistischen Ebenen an, die in neueren Maturaarbeiten stärker hervortreten, ohne diese aus defizitorientierter Perspektive zu betrachten. Neben den bereits genannten sprachformalen Merkmalen konstatiert er eine größere Nähe zur gesprochenen Sprache, die sich u. a. an einem hohen Maß an Implizitheit, „Problemen im Bereich der textuellen Verknüpfungen“ und einer grundlegenden Orientierung „an einer fiktiven Gesprächssituation“ zeige (ebd., S. 142 f.). Ob solche kohärenz- und adressatenbezogenen Textmerkmale auch in Texten weniger sprachversierter Schüler/-innen zu finden sind und ob diesen die erforderliche Balance zwischen sprachlicher Nähe und den Anforderungen an einen schriftlichen Text gelingt, ist eine Frage, die in der geplanten Hauptstudie beleuchtet werden soll.

Eine Bewegung weg von konzeptioneller Schriftlichkeit hin zu einer medial schriftlich auftretenden Mündlichkeit beschreiben auch weitere Studien der letzten Jahrzehnte (z.B. Androutsopoulos 2007). Dabei macht Storrer (2018) auf den Unterschied zwischen interaktionsorientiertem und textorientiertem Schreiben aufmerksam: In internetbasierter schriftlicher Kommunikation dominieren interaktionsorientierte Formen. Neben dem Auftreten von mehr konzeptioneller Mündlichkeit könnte im schulischen Schreiben also auch ein Transfer von interaktionsorientiertem Schreiben auf die Lesererwartung erfolgen. Dürscheid u. a. weisen diese Hypothese zwar zurück, indem sie nach einem Vergleich von Deutschaufsätzen mit der internetbasierten schriftlichen Freizeitkommunikation feststellen, dass die Jugendlichen zwischen dem privaten und dem schulischen Schreiben klar unterscheiden und einfach verschiedene Register nutzen, weshalb kein Einfluss des Schreibens in den neuen Medien auf das Schreiben in der Schule festzustellen sei (Dürscheid et al. 2010, S. 263). Sie räumen jedoch ein: „Was nun aber noch fehlt, ist eine Antwort auf die Frage, ob sich möglicherweise in diachroner Hinsicht Veränderungen in der schulischen Textproduktion zeigen, ob also Jugendliche im Jahr 2009 informeller schreiben, als sie es noch vor 20 Jahren getan haben“ (ebd., S. 266).

Auf einen Wandel der Schriftkultur insgesamt deuten die wenigen Befunde zu Schülertexten hin, die auch die tiefenstrukturelle Ebene einbeziehen. In einem solchen „intermedialen Style“ (Wurzenberger 2016) kann z. B. Kohärenz anders erfahren werden, als es beim traditionellen linearen Erzählen der Fall ist, indem auf dialogisches oder serielles Erzählen aus Serien, Filme und Games aller Art zurückgegriffen wird. Insgesamt könnte sich so bei Jugendlichen allmählich eine neue Textproduktionsweise herausbilden, die sich zunehmend der Reproduktion und Rekombination von vorgefundenen Ideen, Versatzstücken und Entwürfen, letztlich auch aus Textgeneratoren der KI, bedienen wird, die an das eigene Schreibziel und die adressatenbezogene Schreibfunktion angepasst werden müssen. Es bleibt offen, ob sich eine solche neue „Literacy“ Jugendlicher nur als weiteres Register neben der traditionellen Schriftlichkeit etabliert oder ob diese allmählich abgelöst wird, indem etablierte Textnormen in der digital geprägten Kommunikation zunehmend aufgegeben werden und sich letztlich auch

der Registerwechsel zwischen privatem Schreiben und dem schulischen normativen Erwartungshorizont auflösen wird.

Insgesamt zeigen die Befunde des Forschungsüberblicks, dass eine lineare Abwärtsbewegung der Textproduktionsfähigkeiten von Kindern und Jugendlichen pauschal betrachtet eine zu schlichte Annahme wäre. Auch Storrer (2018) vermerkt, es gebe bisher „keine empirischen Anhaltspunkte dafür, dass die netzaffine Jugend durch das interaktionsorientierte Schreiben in ihren textorientierten Schreibkompetenzen beeinträchtigt würde." (ebd., S. 20). Dass aber die Deutschdidaktik bei der Modellierung schulischen Textschreibens auf die Etablierung einer neuen intermedialen Schriftlichkeit reagieren muss (z.B. durch einen noch höheren Stellenwert für den Aufbau von Überarbeitungskompetenzen, die beim Anpassen vorgefundener Textbausteine an die individuellen Schreibziele erforderlich sind), liegt auf der Hand.

2.2 | Erfassung und Kodierung der Textqualität über Raster und Kriterien

Um die Textqualitäten der Schreibprodukte Jugendlicher möglichst objektiv, reliabel und valide bestimmen zu können, sind geeignete Analyseinstrumente erforderlich, die von der konkreten Forschungsfrage abhängig sind und eine genaue Konstruktdefinition voraussetzen (Neumann 2017: 205). Zudem ist das Verhältnis zwischen quantitativen und qualitativen Erhebungsverfahren zu klären. Neumann (2017, S. 208) weist darauf hin, dass Grundlage einer jeden guten quantitativen Studie zur Erfassung von Textqualität immer auch eine qualitative Bestimmung möglicher Abstufungen von Urteilskriterien sowie die Zusammenstellung von Benchmarktexten ist und eine stark dichotomisierte Darstellung nur noch der Abgrenzung beider Verfahren dient. Linguistisch orientierte Modellierungen zur Erfassung der Textqualität – wie das o. g. Zürcher Textanalyseraster (Nussbaumer & Sieber 1995) oder der Basiskatalog zur Textbeurteilung (Becker-Mrotzek & Böttcher 2014) – sind stärker qualitativ ausgerichtet und nehmen die Einschätzung einzelner inhaltlicher und sprachlicher Merkmale von kleineren Schülertextkorpora in den Blick. Quantitative Verfahren setzen skalierbare Daten und eine Beurteilung durch mehrere Rater voraus. Häufig werden in der Schreibforschung – adaptiert auf die Anforderungen des jeweiligen Textmusters sowie der Schreibaufgabe – je nach Fragestellung bestimmte Merkmale für analytische Ratings ausgewählt und mit holistischen Maßen ergänzt. Beide Verfahren bringen gewisse Vor- und Nachteile mit sich (Böhme et al. 2009). Bei der vorliegenden Pilotstudie soll zur Bestimmung narrativer Textqualitäten von Schreibprodukten aus der achten Jahrgangstufe sowohl auf einen holistischen als auch analytischen Zugang zurückgegriffen werden, wobei sowohl menschliche Raterinnen als auch KI-basierte Ratings eingesetzt werden.

2.2.1 | Analytische Kodierung

Analytische Verfahren ermöglichen eine differenzierte Erfassung von inhaltlichen und sprachlichen Merkmalen der Texte. Sie orientieren sich an einer kriterienbasierten Einschätzung, wobei einzelne Items entweder über dichotome Antwortkategorien oder mehrstufige Skalen bewertet werden. Das ermöglicht eine schnelle und gut anwendbare Analyse, die auch von weniger geübten Ratern bewältigt werden kann. Zudem liefern analytische Ratings gute Ergebnisse, die auch für individuelles Feedback an die Schreiber/-innen genutzt werden können (Neumann 2017, S. 210 f.).

2.2.2 | Holistische Kodierung

Die amerikanischen Studien des National Assessment of Educational Progress (siehe z.B. Persky et al. 2003) gehören zu den umfangreichsten und detailliertesten methodischen Vorarbeiten zur Erfassung von Textqualität. Die Texte von Schüler/-innen werden hier anhand textsortenspezifischer Globalskalen eingestuft, wobei den Ratern Kodiermanuale mit entsprechenden Leitfragen zur Textqualität vorliegen (Böhme et al. 2017, S. 62; Neumann 2017, S. 209). Ein Problem bei holistischen Verfahren besteht darin, dass es bei der Bewertung der Textqualität auch zu Fehleinschätzungen kommen kann. Um Verzerrungen entgegenzuwirken ist es erforderlich, dass die Texte von mindestens zwei unabhängigen Rater/-innen beurteilt werden. Da sich Fehlentscheidungen auf diese Weise gut korrigieren lassen, haben sich holistische Verfahren – wie bspw. das NAEP-Rating – in der Praxis bewährt (Neumann 2017, S. 209 f.).

2.2.3 | Einsatz von KI

Die jüngsten Entwicklungen im Bereich der KI auf Basis so genannter Large Language Models (LLM) eröffnen auch für den Unterricht ganz neue Möglichkeiten. Besondere Potenziale werden unter anderem im Bereich der personalisierten Lernunterstützung bzw. eines formativen Feedbacks gesehen und inzwischen auch auf Ebene der Kultusministerien diskutiert (Kultusministerkonferenz, 2021). Die Überlegungen schließen auch die Möglichkeit zur KI-gestützten Bewertung von Texten ein, die gegebenenfalls unmittelbar als lernunterstützendes Feedback im Schreibunterricht genutzt werden kann (Bräunig & Holberg, 2024). Überblicksarbeiten, die sich mit den bisherigen Möglichkeiten und Grenzen der KI-basierten Textbewertung auseinandersetzen (Ramesh & Sanampudi, 2022), können mit dem aktuellen Entwicklungstempo kaum noch schritthalten. Wesentliche Vorteile liegen beispielsweise in der Effizienz: KI-basierte Systeme ermöglichen die schnelle und konsistente Auswertung großer Textmengen, wodurch menschliche Ressourcen und Zeit eingespart werden. Zudem sind sie in der Lage, Muster und Trends zu erkennen, die für menschliche Beobachter schwer zu identifizieren sind. Die Anwendung objektiver Kriterien durch diese Systeme unterstützt zusätzlich die Standardisierung von Bewertungen. Trotz dieser Vorzüge stehen KI-Tools vor diversen Herausforderungen. Insbesondere scheinen sie nach ersten Erfahrungen Schwierigkeiten bei der Erfassung von Nuancen, kreativen Elementen sowie kulturellen Kontexten in Texten zu zeigen, was potenziell zu Fehlinterpretationen oder ungenauen Bewertungen führen könnte. Darüber hinaus sind sie stark von den Trainingsdaten abhängig, was zu inhärenten Voreingenommenheiten oder Einschränkungen in Bezug auf die Vielfalt und Qualität der Ergebnisse führen kann, wenn die Trainingsdaten nicht hinreichend repräsentativ sind. In der Gesamtschau bieten KI-basierte Tools für die Beurteilung von Textqualitäten eine Mischung aus Präzision, Skalierbarkeit und Schnelligkeit (Ramesh & Sanampudi, 2022). Dennoch erfordern sie menschliche Überwachung und Anpassung, um ihre Effektivität zu maximieren und ihre Grenzen zu erkennen.

3 | Pilotstudie: Aufbau und Methoden

3.1 | Forschungsfragen und Aufbau der Pilotstudie

Die Qualität der handschriftlich verfassten Texte wird für die beiden Erhebungszeitpunkte 1998 vs. 2023 verglichen. Da in der Erhebung von 2023 zusätzlich ein digitales Schreibsetting

durchgeführt wurde, können zudem Unterschiede in der Textqualität auch im Hinblick auf das verwendete Schreibmedium beleuchtet werden. Somit lassen sich zwei grundsätzliche Perspektiven unterscheiden – eine diachrone und eine synchrone Perspektive. Dabei stehen die folgenden Forschungsfragen[3] im Zentrum:

Diachrone Vergleichsperspektive:

(1) Lassen sich Unterschiede in der Textlänge und der Textqualität feststellen?
   a. *Vergleich handgeschriebener* Texte von *1998* mit *handgeschriebenen* Texten von *2023*
   b. *Vergleich handgeschriebener* Texte von *1998* mit *digital geschriebenen* Texten von *2023*

Synchrone Perspektive:

(2) Lassen sich Unterschiede in der Textlänge und der Textqualität feststellen?
   c. *Vergleich handgeschriebener* Texte von 2023 mit *digital geschriebenen* Texten von 2023

Die Bewertung der Textqualität erfolgt sowohl durch menschliche Raterinnen als auch durch den Einsatz von KI. Ergänzend zu den Forschungsfragen (1) und (2) erfolgt ein auswertungsbezogener Vergleich der Bewertung durch Mensch und KI bei allen Texten:

(3) Wie stark ist die Übereinstimmung in der Bewertung der Textqualität zwischen menschlichen Raterinnen und der Bewertungen durch ein KI-Tool (Chat GPT 4.0)?

Für die Pilotstudie ergibt sich damit der folgende Aufbau:

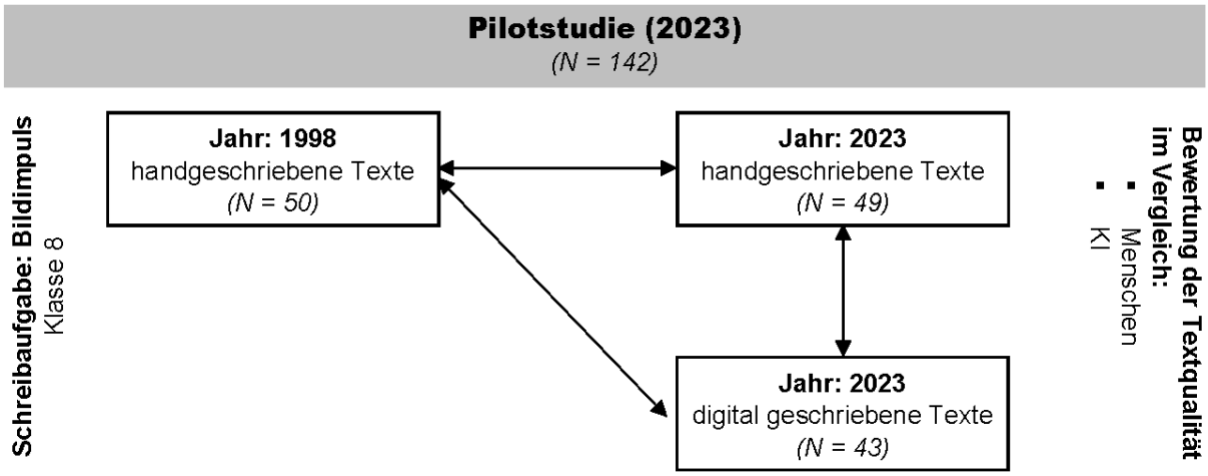


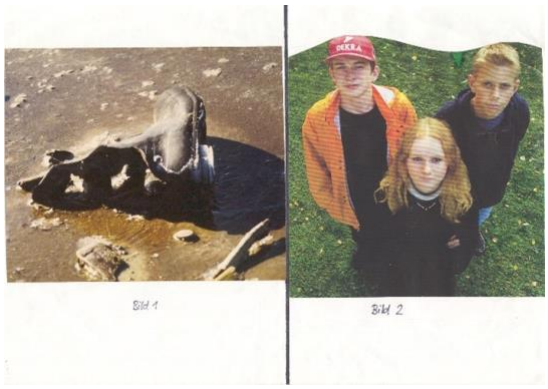
Abb. 1: Aufbau der Pilotstudie

3.2 | Setting der Erhebung und Erstellung eines Pilotkorpus

Um die Vergleichbarkeit zu gewährleisten, fanden alle Erhebungen unter ähnlichen Schreibbedingungen zur gleichen Schreibaufgabe statt. Diese wurde aus „LuKo“ (Fix & Melenk 2000) ausgewählt und übernommen. Sie bestand darin, einen Text nach Anregung durch einen Bildimpuls für Mitschüler/-innen (Adressat/Funktion) zu verfassen, der zudem im Rahmen einer Schreibkonferenz überarbeitet wurde (Wirkung/soziale Interaktion). Das Aufgabensetting erfüllte bereits 1998 weitgehend die Anforderungen profilierter Schreibaufgaben (Bachmann

[3] Da sich aus der Literaturlage keine gerichteten Hypothesen ableiten lassen, wurde für die Pilotstudie auf eine entsprechende Festlegung verzichtet. Die Ergebnisse sollen es ermöglichen, gerichtete Hypothesen für die Hauptstudie zu formulieren.

& Becker-Mrotzek 2010). Die peer-orientierte Adressatenorientierung sollte dazu beitragen, dass der Einfluss schulischer Textsortennormen etwas in den Hintergrund tritt. Aber auch ohne die Vorgabe einer Textsorte evozieren die beiden Bildimpulse, zu denen Ideen entwickelt werden sollten, zumeist ein narratives Textmuster. Insofern überrascht es nicht, dass trotz der offen gestellten Aufgabe in der Regel auf Erzählungen zurückgegriffen wurde. Die Schüler/-innen konnten sich zwischen einem Gegenstand (ein alter Stiefel im Schlamm) oder drei Jugendlichen, die nach oben blicken, entscheiden (vgl. Abb. 2):



Beide Bilder enthalten Unbestimmtheitsstellen, die auf ein erzählwürdiges Ereignis und einen entsprechenden Planbruch hinlenken (was ist dem Besitzer des Stiefels passiert, wohin schauen die drei?). Bild 1 unterstützt stärker die Ausprägung in Richtung außeralltäglicher Kriminal- oder Science-Fiction-Erzählungen (wenn der Stiefelbesitzer als Opfer eines Verbrechens oder die Landschaft als extraterrestrisch wahrgenommen wird), Bild 2 die Ausarbeitung alltagsnäherer Figuren, bei denen der Planbruch durch ein Erlebnis oder eine weitere Figur, die von oben kommt, entsteht. In beiden Fällen kann erwartet werden, dass die in der Erzähldidaktik bedeutsamen Kriterien der Ungewöhnlichkeit, des plötzlichen Planbruchs und der Auflösung in einer pointierten Koda in dieser Altersstufe von vielen Schüler/innen auch ohne explizite Vorgabe berücksichtigt werden (vgl. Augst 2010, S. 65; Wild 2020, S. 214ff.). Diese Elemente werden in den Kriterienrastern der Auswertung berücksichtigt (vgl. 3.3.1).

In Anlehnung an das Setting von LuKo 1998 umfasste die Unterrichtseinheit vier Unterrichtsstunden und gliederte sich grob in zwei Phasen: Die erste Doppelstunde (Phase A) beinhaltete neben einer kurzen Einführung eine Ideensammlung in Form eines Clusters bzw. einer Mindmap und berücksichtigte so Planungsaspekte, bevor die Schüler/-innen ihre ersten Textentwürfe verfassten. In der zweiten Doppelstunde (Phase B) konnte zunächst an den Entwürfen weitergearbeitet werden, bevor eine Schreibkonferenz durchgeführt wurde, auf deren Grundlage unter Zuhilfenahme einer Checkliste anschließend Textrevisionen vorgenommen werden durften (aber nicht mussten, einige machten davon keinen Gebrauch), um letztlich eine finale Fassung fertigzustellen. Die ersten Textentwürfe wurden zwar dokumentiert, ausschließlich die finale, ggf. überarbeitete Fassung wurde jedoch in den Auswertungsprozess der Pilotstudie einbezogen. Der auf diese Weise prozessorientiert vorstrukturierte Unterricht wurde von Masterstudierenden in Anwesenheit der jeweiligen Deutschlehrkräfte durchgeführt. Die entstandenen Texte wurden transkribiert und mit Code-Namen versehen, damit die Raterinnen keinen Anhaltspunkt zum Entstehungszeitpunkt hatten.

An der Pilotstudie nahmen insgesamt 7 Realschulklassen aus dem Großraum Stuttgart und Karlsruhe teil. Für das Pilotkorpus wurde ein Umfang von ca. 150 Texten angestrebt, um eine

ausreichende Anzahl für die Entwicklung des Ratingverfahrens zu haben, aber auch, um bereits erste Vergleichstendenzen feststellen zu können.

Bei der Erhebung der Texte war somit aus einem Überangebot (ca. 120 Texte 2023, ca. 400 Texte 1998) auszuwählen. Die Auswahl wurde nach bestimmten Kriterien vorgenommen:

- In Filter 1 wurden Texte unter 50 Wörtern ausgeklammert, ebenso die nicht-narrativen Texte, hierbei handelte es sich aber um sehr wenige Einzelfälle. Außerdem wurden Schüler/innen, die am zweiten Termin fehlten, nicht einbezogen.
- Filter 2: Da frühere Hauptschüler/-innen im heutigen Schulsystem Baden-Württembergs u. a. auch in der Realschule sind, wurden aus der Erhebung von 1998 hälftig Haupt- und Realschüler/-innen einbezogen. Bei der weiteren Reduktion wurden jeweils zwei Klassen von 2023 und 1998 weitgehend als Klumpenstichprobe erhalten, bei denen die Ergebnisse vollständig vorliegen.
- Filter 3: Der Genderanteil Jungen-Mädchen lag bei den analog schreibenden Realschulklassen 2023 etwa 2:1. Diese Relation wurde auch bei der Auswahl aus den Klassen von 1998 - jeweils zufällig ausgewählt - beibehalten.

Das so bereinigte Korpus (N = 142) setzte sich wie folgt zusammen (s. Abb. 1):

a. 50 Texte aus dem alten Korpus (LuKo), die Anfang 1998 (analog) entstanden waren,
b. 49 Texte aus Realschulklassen, die 2023 mit analogen Schreibmedien im Unterricht erhoben wurden,
c. 43 Texte aus Realschulklassen, die 2023 mit digitalem Medium auf Tastatur geschrieben wurden

3.3 | Auswertungsmethoden

Zur Veranschaulichung der durchgeführten Ratings wird ein Beispieltext[4] herangezogen, an dem der Einsatz der Analyseverfahren erläutert wird sowie Potenziale und Herausforderungen KI-basierter Ratings thematisiert werden. Die Textauswahl erfolgte auf der Grundlage theoretischer Erwartungen an narrative Textmuster, die im Vorfeld der Untersuchung festgelegt wurden (siehe hierzu auch 3.3.1.1). Diese Erwartungen sind unabhängig von den späteren Raterurteilen und dienen im Folgenden lediglich zur Veranschaulichung der durchgeführten Analysen.

> *Ich bin Julia und erzähl euch heute über eine Sache die mir passiert die ich nicht vergessen kann. Tim, Jan und ich mussten ein Schulprojekt machen und entschlossen zu Tims Haus zu gehen. Wir fingen an mit unseren Projekt bis auf ein mal ein Lauter Knall kamm. Wir erschrockten uns alle und haben direkt draußen geschaut ob da was ist aber nichts war da. Wir machten weiter an unseren Schulprojekt bis wieder ein Knall von draußen kam. Diesmal wollten Tim und Jan unbedingt wissen woher es kommt. Wir ziehen unsere Jacke an und lauften nach draußen. Tim war vor uns und zeigte den weg von wo die Knalle kammen. Bis wir an der Garage waren sahen wir den Vater von Tim und er baute ein magischen Stuhl auf. Tim fragte warum die Knalle so laut sind. Tim´s Vater m sagte das es sich so gehört. Also liefen wir wieder rein bis der Vater von Tim da stand. Wir schauten uns verwirrt an und wahren verwundert wie es*

[4] Die Beispielanalyse dient nicht dazu, die Raterurteile nachträglich zu rechtfertigen oder zu erklären. Vielmehr soll gezeigt werden, wie im ausgewählten Text verschiedene narrative Elemente realisiert werden und wie dies durch die Analyseinstrumente erfasst werden kann.

> *sein kann das Tims Vater vor uns drinnen war. Tim´s Vater begrüßte uns und Tim meinte dann das er eben noch draußen ware. Tim Vater mei schaute verwirrt und hatte keine Ahnung über was wir roden. Tim und Jan meinte lauften einfach weiter. Wer war in der Garage wen Tim´s Vater die Ganze Zeit drinnen war. De Bis heute ist dieser Fall mir ein Rätsel.*

Schülertext [01109]: Erzählung von Julia (Textkorpus 1998)

Zur Bestimmung der Qualität der Schülertexte werden zwei Ratingverfahren eingesetzt: zum einen ein analytisches Kodierschema mit einer mehrstufigen Skala zu inhaltlichen und sprachlichen Aspekten, zum anderen ein Schema zur holistischen Beurteilung in Form eines textmusterspezifischen Globalurteils.

3.3.1 | Analytische Kodierung

3.3.1.1 | Kriterien

Der analytischen Kodierung liegen aufgabenspezifische Kriterien zugrunde, die sich auf die Charakteristika narrativer Texte sowie das aufgrund des Bildimpulses selbst gesetzte Thema der entsprechenden Aufgabe („Erzählung zu einem Bildimpuls“) beziehen. Der Kodierbogen umfasst vier textmusterspezifische Items, die auf Grundlage des Zürcher Textanalyserasters und seiner Adaptionen (z. B. Fix 2025, S. 208f.; Kruse u.a. 2012, S. 95ff.; Wild 2020, S. 214ff., Sturm 2023, S. 97) entwickelt wurden. Von diesen vier klassischen Kriterien (Aufbau – Inhalt - Stil - Form) berücksichtigen vor allem die Kriterien 1) und 2) die in der erzähldidaktischen Forschung (vgl. z.B. das Kriterienraster RESTLESS bei Wild 2020, S. 214ff.) als bedeutsam herausgearbeiteten Aspekte der Kohärenz und des erzählwürdigen Ereignisses (mit Planbruch bzw. Komplikation, Erzählsituation, Figuren, Ort, Rahmenhandlung sowie Lösung des Problems):

1) *Der Text ist strukturell und inhaltlich kohärent*
   (z. B. nachvollziehbarer, verständlicher Textaufbau [Einleitung, Komplikation/Planbruch, Koda], Erzählstruktur mit einem „roten Faden“ ohne thematische Sprünge, gelungene thematische Entfaltung und Übergänge, stimmiges Verhältnis Details-Ganzes)

2) *Der Text ist ideenreich und originell*
   (z. B. Qualität der Einfälle, inhaltliches und sprachliches Wagnis, erzählwürdiges Ereignis, entwickelte Figuren und Orte, Ungewöhnlichkeit, ggf. Plötzlichkeit, unterhaltend im Hinblick auf die Adressaten)

3) *Der Text weist einen angemessenen Erzählstil auf*
   (adressatenorientierter und textsortenrelevanter Sprachgebrauch, z. B. variantenreicher Wortschatz und Satzbau, Markierung von Plötzlichkeit, wörtliche Rede, anschauliche Attribute)

4) *Der Text ist sprachformal korrekt (Grammatik und Orthografie)*
   (auffällige Abweichungen zur in dieser Stufe erwartbaren Sprachnormorientierung)

Die vier Kriterien werden auf einer fünfstufigen Skala hinsichtlich ihres Auftretens („trifft weitestgehend nicht zu“ = 1 bis „trifft weitestgehend zu“ = 5) bewertet und anschließend aggregiert.

3.3.1.2 | Raterdesign

Die Auswertung der Daten wurde von zwei geschulten menschlichen Raterinnen durchgeführt, die Deutsch für das Lehramt studierten und als Hilfskräfte im Projekt tätig waren.[5] Die Raterinnen wurden in mehreren Sitzungen anhand von Beispielanalyen geschult und nahmen an Diskussionen teil, in denen auf Basis von Probekodierungen und Rückmeldungen ein vertrauter Umgang mit den Kodieranweisungen erzielt werden sollte. Außerdem wurde ein KI-basiertes Rating durch ChatGPT (Version 4) ohne Training vorgenommen. Obwohl inzwischen eine wachsende Anzahl an KI-Chatbots zur Verfügung steht, bleibt ChatGPT bislang die meistgenutzte KI, die zudem breit zugänglich ist und in vielen wissenschaftlichen Kontexten eingesetzt wird (vgl. Fletcher, 2023). Die Prompteingabe für ChatGPT lautete:

> *„Bitte bewerte die folgenden Schülertexte der 8. Jahrgangsstufe anhand der folgenden Kriterien. Vergib dabei folgende Bewertung für jedes Kriterium:“* [6]

Im Gegensatz zu den beiden Raterinnen, die ihre Bewertungskriterien während einer Schulung aufeinander abstimmen konnten, wurde die KI nicht mit Beispieltexten trainiert. Sie erhielt also kein spezifisches Feedback, um ihre Bewertungen mit denen der menschlichen Raterinnen besser in Einklang zu bringen. Dadurch war die KI im Vergleich zu den menschlichen Bewerterinnen im Nachteil. Die Entscheidung, auf ein vergleichbares Training der KI zu verzichten, zielte darauf ab, in dieser ersten Pilotphase eine eher konservative Einschätzung der KI-unterstützten Textbewertung zu ermöglichen, in der die potenziellen Ungenauigkeiten bewusst in Kauf genommen werden, um daraus Hinweise für spätere Trainingsimpulse in der Hauptstudie zu gewinnen. Um für die anschließenden Vergleiche der Textkorpora eine möglichst zuverlässige Beurteilung zu erhalten, wurden durch die beiden unabhängigen Raterinnen und die KI Dreifachkodierungen vorgenommen, d.h. jeder Schülertext wurde von zwei studentischen Raterinnen und der KI bewertet. Für das analytische Rating liegen somit 423 vollständige Kodierbögen vor. Die Raterinnen erhielten die vorhandenen Texte allesamt in digitalisierter Fassung und in zufälliger Reihenfolge angeordnet, so dass für die Beurteilerinnen nicht ersichtlich war, welcher Text zu welchem Zeitpunkt bzw. mit welchem Medium erstellt worden ist.

3.3.1.3 | Beispielanalyse

Analytische Variable	Beispielanalyse zu Julias Text	
	Menschliche Raterinnen (1/2)	KI-Rating
Codes: 1= trifft weitestgehend nicht zu; 2= trifft eher nicht zu; 3= teils/teils; 4= trifft eher zu; 5= trifft weitestgehend zu		
1)	3 / 2	5
2)	**4** / 3	**4**
3)	3 / **4**	**4**
4)	**3** / **3**	**3**

Tabelle 1: Beispielanalyse zu Julias Text

Die Frage, ob Julias Text das Kriterium struktureller und inhaltlicher Kohärenz erfüllt, wird durch die Raterinnen und das KI-Rating unterschiedlich bewertet. Während bei den menschlichen Raterinnen die weniger gelungene thematische Entfaltung des Textes und die Übergänge

[5] Unser Dank gilt den beiden Studentinnen Stephanie Geltner (PH Ludwigsburg) und Hannah Lutz (PH Karlsruhe) für ihr sorgfältiges Rating sowie allen weiteren Studierenden, die bei der Erhebung der Texte mitgewirkt haben.

[6] Es folgen die Codes 1-5, siehe Tabelle 1

stärker ins Gewicht fallen (insbesondere zum Ende, eine Koda wird nicht herausgearbeitet), bewertet die KI die strukturelle und inhaltliche Kohärenz positiver, möglicherweise, weil sie nicht einbezieht, dass kürzere bzw. mittellange Texte (wie Julias Text mit 235 Wörtern) tendenziell weniger anfällig für Kohärenzbrüche sind als längere Texte. Ideenreichtum und Originalität werden durch alle drei Ratings recht positiv bewertet. Es tritt ein ungewöhnliches Ereignis ein und die Figuren und Orte wurden ebenfalls als erzählwürdig eingestuft. Lediglich die Qualität der Einfälle und die Frage, ob die Geschichte mit Blick auf die Leser/-innen unterhaltsam wirkt, wurden von den Raterinnen kritisch bemängelt. Auch der Erzählstil von Julias Text scheint weitestgehend angemessen zu sein, wobei insbesondere der adressatenorientierte („… und heute erzähle ich euch…“) und textsortenrelevante Sprachgebrauch und die Markierung von Plötzlichkeit („… bis auf ein mal ein lauter Knall kam“) positiv hervorzuheben sind. Die sprachformale Korrektheit des Textes wird hingegen weniger positiv bewertet. Auffallend sind u.a. Fehler bei der Groß- und Kleinschreibung und die fehlerhafte Bildung der Vergangenheitsform bei unregelmäßigen Verben sowie syntaktische Abweichungen. Insgesamt lässt sich Julias Text auf Basis des analytischen Ratings mit einer gemittelten Punktzahl von 3,5 als eher durchschnittlich einschätzen.

3.3.1.4 | Interraterreliabilität

Als Maß für die Interraterreliabilität wird die Intraklassenkorrelation (Intra Class Correlation, ICC) berichtet, da es sich um intervallskalierte Daten handelt, die von verschiedenen Raterinnen beurteilt wurden (vgl. Wirtz & Caspar 2002, S. 157 ff.). Weil für den späteren Vergleich der Aufsatzkorpora alle drei vorliegenden Urteile zu einem Gesamtwert kombiniert werden, können etwaige Unterschiede in den Urteilstendenzen der Raterinnen (z. B. Strenge) unberücksichtigt bleiben, weshalb unjustierte ICCs berechnet wurden.

Tabelle 2 gibt einen Überblick über die ermittelten Werte der Intraklassenkorrelationen für die vier Items.

	ICC	Paarweise Übereinstimmungen		
	alle 3 Urteile	Raterin 1 / 2	Raterin 1 / ChatGPT	Raterin 2 / ChatGPT
Item-Kohärenz	.70**	.73**	.56**	.48**
Item-Originalität	.72**	.75**	.55**	.56**
Item-Erzählstil	.59**	.54**	.43**	.48**
Item-Formale Korrektheit	.79**	.77**	.69**	.68**
Gesamtwert	.79**	.81**	.65**	.65**

Tabelle 2: Intraklassenkorrelation für die analytische Kodierung (menschliche Raterinnen und ChatGPT)

Für den Gesamtwert der analytischen Kodierung ergibt sich eine gute Intraklassenkorrelation von .79. Die relativ hohe Übereinstimmung bestätigt sich weitgehend auch bei differenzierter Betrachtung der Einzelitems. Die Werte für Kohärenz, Originalität und formale Korrektheit liegen zwischen .70 und .79. Lediglich die Bewertung des Erzählstils fällt mit .59 etwas hinter die anderen Items zurück. Um die Eignung der KI-Urteile genauer beurteilen zu können, wurden zusätzlich ICCs für paarweise Übereinstimmungen berechnet. Dabei zeigt sich, dass die Urteilsübereinstimmungen nahezu identisch ausfallen, wenn man lediglich die beiden menschlichen Raterinnen in die Berechnungen einbezieht. Demgegenüber fallen die paarweise berechneten Übereinstimmungen zwischen einem menschlichen Urteil und der KI numerisch

etwas geringer aus als zwischen den beiden Beurteilerinnen. Vor dem Hintergrund, dass die Prompts für die KI in der vorliegenden Pilotstudie bewusst minimalistisch gehalten wurden, während die menschlichen Raterinnen Gelegenheit hatten, ihr Urteilsverhalten in mehrstündigen Sitzungen und Diskussionen über Beispieltexte aufeinander abzustimmen, war dies aber erwartbar (vgl. 5.).

3.3.1.5 | Interne Konsistenz

Die interne Konsistenz des analytischen Kodierschemas wurde durch Reliabilitätsanalysen ermittelt. Angesichts der geringen Itemzahl und der weniger restriktiven Annahmen wurde dafür McDonald's Omega anstelle von Cronbach's Alpha berechnet (McDonald, 1999). Die entsprechenden Kennwerte sind in Tabelle 3 dargestellt. Die ergänzende Berechnung von Cronbach's Alpha ergab keine Hinweise auf die bei Omega gelegentlich beobachteten Reliabilitätsüberschätzungen (Malkewitz et al., 2023). Die Abweichungen zwischen den beiden Kennwerten waren durchweg marginal (max. 0.02). Für den über alle drei Urteile gemittelten Gesamtwert ergibt sich eine gute Reliabilität von .86. Um auch hier differenziertere Kenntnisse über die Eignung der KI zur Textbeurteilung zu gewinnen, wurden die Reliabilitätskennwerte ergänzend für alle drei Ratings separat berechnet. Im direkten Vergleich fällt die Reliabilitätsschätzung für die KI mit .91 tendenziell höher als bei den menschlichen Raterinnen (.78 bzw. .74) aus.

	alle 3 Urteile	Raterin 1	Raterin 2	ChatGPT
McDonald's Omega	.86**	.78**	.74**	.91**

Tabelle 3: Interne Konsistenz des analytischen Ratings

3.3.2 | Holistische Kodierung

3.3.2.1 | Niveaustufen

Die verwendete holistische Kodierung basiert auf einer vom IQB vorgenommenen Übersetzung und Anpassung des NAEP Holistic Scoring Guide für narrative Aufgaben in der Version für die achte Jahrgangsstufe (Böhme et al. 2009; Institut zur Qualitätsentwicklung im Bildungswesen (IQB) 2012; Persky et al. 2003). Sie wurde mit Blick auf die konkreten Schreibaufgaben der vorliegenden Studie geringfügig modifiziert. Die Raterinnen wurden vorab instruiert, bestimmte Teilaspekte bei der Beurteilung zu berücksichtigen und diese möglichst gleich gewichtet in ihr Globalurteil einfließen zu lassen. Diese Teilaspekte beziehen sich einerseits auf inhaltliche Erwartungen und andererseits auf Anmerkungen zur sprachlichen Bewertung. Da die Raterinnen bei der abschließenden Zuordnung der Texte zu einer Niveaustufe eine Entscheidung treffen müssen, ist es erforderlich, eine Abwägung vorzunehmen. Dabei sollten inhaltliche Aspekte den höchsten Stellenwert haben, gefolgt von sprachlichen Merkmalen und Kriterien der sprachlichen Richtigkeit.

Stufenbeschreibung der holistischen Globalskala (Narration)
Stufe 5

Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist folgende Merkmale auf:

- Der Text entfaltet eine gelungene Handlungsfolge, die durch gut gewählte Details entwickelt und gestaltet wird.

- Der Text verfügt über einen stimmigen Aufbau mit klarer Erzählstruktur und gut ausgebauten Übergängen zwischen den Teilen der Handlung und ist durchgängig kohärent. Die Textsorte wird durchweg eingehalten.
- Der Text zeichnet sich durch abwechslungsreichen Satzbau und meist treffende Wortwahl aus. Fehler in der Grammatik, in der Orthografie oder in der Kommasetzung treten kaum auf und beeinträchtigen nicht das Verständnis des Schülertextes.

(…)[7]

Stufe 1

Der Schülertext ist in Bezug auf den Schreibanlass zu beurteilen. Ein Text dieser Kategorie weist eines oder mehrere der folgenden Merkmale auf:

- Der Text bemüht sich um eine Bearbeitung der Aufgabe, stellt jedoch keine zusammenhängenden Inhalte bereit, oder der Text paraphrasiert lediglich die Aufgabenstellung.
- Der Text zeigt keine erkennbare Struktur oder besteht aus einer einzigen isolierten Aussage. Von der zutreffenden Textsorte wird abgewichen.
- Der Text weist kaum oder überhaupt keine Beherrschung des Satzbaus und kaum oder gar keine Beachtung der Satzgrenzen auf. Die Wortwahl ist überwiegend oder über den gesamten Text hinweg unzutreffend.
- Eine Vielfalt an Fehlern in Grammatik oder Sprachgebrauch, Orthografie und in der Kommasetzung verhindert in weiten Teilen das Verständnis des gesamten Schülertextes.

Stufe 0

- Der Text ist zu kurz und bietet keine hinreichende Substanz für eine zuverlässige Bewertung.

Tabelle 4: Auszug aus der modifizierten Stufenbeschreibung der holistischen Globalskala zu narrativen Texten (IQB 2023)

3.3.2.2 | Raterdesign

Die holistische Kodierung der Schülertexte erfolgte analog zur Kodierung des analytischen Verfahrens. Bei Unsicherheiten in der Bewertung konnten sich die menschlichen Raterinnen an geeigneten Benchmarktexten orientieren. Die KI hingegen erhielt, wie auch bei der analytischen Bewertung, keine Gelegenheit, ihr Urteilsverhalten aufgrund von Zusatzinformationen bzw. spezifischen Feedbackprompts anzupassen.

Die Prompteingabe für das holistische Rating durch ChatGPT lautete wie folgt:

> *„Bitte bewerte die folgenden Schülertexte der 8. Jahrgangsstufe anhand des folgenden Rasters:“* [8]

Für das holistische Rating liegen insgesamt 423 vollständige Kodierbögen vor.

3.3.2.3 | Beispielanalyse

Die Beispielanalyse kann sich hier auf die Einschätzung der Raterinnen bezüglich der entsprechenden Niveaustufen beschränken: Julias Text wird sowohl von den menschlichen Raterinnen als auch dem KI-basierten Rating auf Stufe 3 eingeschätzt. Julia bemüht sich um die Entwicklung einer Handlungsfolge. Die Ich-Erzählerin muss mit ihren Mitschülern ein Projekt

[7] Aus Platzgründen werden die mittleren Stufen 2-4 hier nicht abgebildet. Die komplette Skala kann unter folgendem Link eingesehen werden: https://cloud.ph-karlsruhe.de/index.php/s/ctNSL6aqwDjDRa3

[8] Das Raster wurde in den Prompt komplett eingefügt.

durchführen, sie gehen zu Tims Haus, auf einmal ertönt ein Knall, sie gehen diesem auf die Spur. Es wird jedoch deutlich, dass der Text klare Schwächen in der Erzählstruktur zeigt und Teile der Handlung – insbesondere zum Ende der Erzählung – unverbunden im Text stehen. Julias Text zeigt meist Sicherheit im Satzbau und der Beachtung von Satzgrenzen, wobei allerdings Fehler in der Orthographie und Kommasetzung den Lesefluss beeinträchtigen. Die hier im Rahmen des Globalurteils beschriebene Bewertung zu Julias Text deckt sich somit mit der durchgeführten Beispielanalyse des analytischen Ratings.

3.3.2.4 | Interraterreliabilität

Als Maß der Interraterreliabilität wird auch für die holistischen Ratings die unjustierte Intraklassenkorrelation (ICC) berichtet. Tabelle 5 gibt einen Überblick über die ermittelten Werte der Intraklassenkorrelationen für die Textqualität nach den NAEP-Stufen für menschliche Raterinnen und KI-basierte Ratings.

	ICC	Paarweise Übereinstimmungen		
	alle 3 Urteile	Raterin 1 / 2	Raterin 1 / ChatGPT	Raterin 2 / ChatGPT
Gesamtwert	.81**	.84**	.69**	.66**

Tabelle 5: Intraklassenkorrelation für die holistische Kodierung (menschliche Raterinnen und ChatGPT)

Für die holistische Kodierung der Schülertexte ergeben sich somit sehr ähnliche Intraklassenkorrelationen wie bei der analytischen Bewertung. Die Übereinstimmung über alle drei Urteile hinweg liegt mit .81 in einem guten Bereich. Auch hier wurden ergänzend paarweise Intraklassenkorrelationen berechnet, um die Eignung der KI-Urteile differenzierter bewerten zu können. Während die ICC bei isolierter Betrachtung der beiden menschlichen Raterinnen mit .84 in einer vergleichbaren Größenordnung liegt, fallen die Übereinstimmungen zwischen Mensch und Maschine mit .69 und .66 etwas niedriger aus[9].

3.3.3 | Erfassung der Textlänge

Für eine erste ökonomische Einschätzung der Textqualität kann auch die Textlänge herangezogen werden (Neumann 2012, S. 78). Tabelle 6 (siehe n. Seite) enthält die Korrelationen zwischen den erhobenen Maßen zur Erfassung der Textqualität. Diese zeigen zunächst, dass zwischen dem Gesamtwert des analytischen Kodierschemas und dem holistischen Urteil ein sehr enger Zusammenhang besteht. Die Korrelation in Höhe von .89 deutet darauf hin, dass die beiden Maße für eine globale Gesamteinschätzung der Textqualität im Grunde austauschbar sind. Dies entspricht auch Ergebnissen anderer Studien (Grabowski 2022). Demgegenüber fallen die Korrelationen dieser beiden Maße mit der Textlänge, die einen eher groben Indikator zur Beurteilung der Textqualität darstellt, mit .60 und .47 erwartungsgemäß niedriger aus.

[9] Aus ersten Testläufen liegen auch KI-Bewertungen durch die kostenfreie Version von ChatGPT 3.5 vor, so dass auch eine Schätzung der Übereinstimmung zwischen den beiden KI-Urteilen möglich ist. Die ICC von .68 deutet auf eine vergleichbar gute Übereinstimmung hin wie zwischen den beiden menschlichen Raterinnen.

		separat für alle 3 Urteilenden		
	kombinierte Urteile	Raterin 1	Raterin 2	ChatGPT
holistisch – analytisch	.89**	.82**	.72**	.91**
holistisch – Textlänge	.60**	.53**	.66**	.25**
analytisch – Textlänge	.47**	.48**	.52**	.18*

Tabelle 6: Korrelationen der Maße zur Erfassung der Textqualität

Ebenfalls in Tabelle 6 dargestellt sind separat für die einzelnen Raterinnen bzw. die KI berechnete Korrelationen, die abermals eine differenziertere Bewertung der KI-Performanz ermöglichen sollen. Dabei zeigt sich, dass die Korrelation zwischen holistischem und analytischem Urteil bei allen drei Bewertungen in einem sehr hohen Bereich liegt, auch wenn dieser Zusammenhang für Raterin 2 numerisch etwas niedriger ausfällt. Im Hinblick auf die Korrelationen mit der Textlänge ergibt sich hingegen für die ChatGPT-Urteile ein etwas anderes Bild als für die menschlichen Ratings. Während die Qualitätsurteile der beiden Raterinnen in ähnlicher Höhe mit der Textlänge korrelieren, wie dies beim kombinierten Gesamturteil der Fall ist (zwischen .48 und .66), fallen diese Zusammenhänge für die KI mit .25 und .18 augenscheinlich geringer aus, weil die KI die Textlänge offenbar ohne entsprechende Prompt-Präzisierung nicht als relevant einstuft. Für das holistische Urteil ist der ermittelte Unterschied auch statistisch signifikant ($z = -1.75$, $p < .05$).

3.3.4 | Fazit zu den gewählten Auswertungsmethoden

Insgesamt deuten die Ergebnisse darauf hin, dass die Qualität der Texte sowohl mit dem analytischen als auch dem holistischen Kodierschema zuverlässig eingeschätzt werden kann: Das analytische Kodierschema ermöglicht eine differenzierte Erfassung von inhaltlichen und sprachlichen Aspekten der Schülertexte, die sich auch gut für ein individuelles Feedback eignet. Die holistische Kodierung der Texte mittels textmusterspezifischer Globalurteile ist zwar vergleichsweise weniger detailliert, basiert aber auf umfangreichen Vorarbeiten der NAEP-Studien und lässt sich in der Praxis gut durchführen. Somit sollen beide Ratingverfahren für die Hauptstudie weiterverfolgt werden.

Beide Verfahren eignen sich auch für den Einbezug eines KI-Urteils. Letzteres weist zwar bei paarweisen Analysen gegenüber einer Übereinstimmung zwischen Mensch und Mensch etwas schlechtere Inter-Rater-Reliabilitäten auf. Bei Einbezug aller drei Urteile liegen jedoch die Intraklassenkorrelationen in einer vergleichbaren Größenordnung wie für das menschliche Duo. Aufgrund dieser Ergebnisse werden für den anschließenden Vergleich der Aufsatzkorpora (analog 1998, analog 2023, digital 2023) die Urteile von Mensch und Maschine kombiniert (Mittelwert aus allen drei Ratings).

Angesichts der sehr hohen Korrelationen zwischen dem holistischen und dem Gesamtwert des analytischen Kodierschemas wird zudem auf eine separate Betrachtung der beiden Maße verzichtet. Stattdessen werden die beiden Werte zu einem Gesamtwert „Textqualität“ kombiniert (Mittelwert) und in den nachfolgenden Analysen als abhängige Variable verwendet. Dessen ungeachtet werden die einzelnen Items der analytischen Kodierung (Kohärenz, Originalität, Erzählstil, Formale Korrektheit) separat in die Analysen einbezogen, um überprüfen zu können, ob sich für die einzelnen Qualitätsaspekte differenzierte Ergebnisse zeigen.

4 | Erste Befunde aus der Pilotstudie zu den gestellten Forschungsfragen (diachron und in Abhängigkeit vom Schreibmedium)

4.1 | Unterschiede in der Textqualität

Tabelle 7 zeigt die deskriptiven Statistiken für Textqualität und Textlänge sowie die Ergebnisse der durchgeführten Varianzanalysen. Wie aus der Tabelle hervorgeht, ergeben sich in allen Analysen signifikante Unterschiede zwischen den Textkorpora. In Bezug auf die Maße der Textqualität bestätigen post-hoc Analysen, dass die handschriftlich angefertigten Aufsätze aus dem Jahr 2023 konsistent schwächer beurteilt werden als die beiden anderen Textkorpora (alle *p*'s < .01). In Effektstärken (Cohen's *d*) ausgedrückt, bewegt sich der Qualitätsrückstand analoger Texte aus dem Jahre 2023 gegenüber den anderen Korpora im Bereich zwischen 0.94 und 1.69. Es handelt sich somit gängigen Konventionen zufolge durchweg um große Effekte (*d* > .08; Cohen, 1988). Nicht minder interessant ist der Befund, dass die digital verfassten Texte aus 2023 über alle Rating-Indizes hinweg auf einem sehr vergleichbaren Niveau liegen wie die analogen Texte von 1998. Die Befunde legen somit nahe, dass keinesfalls generell von einer diachronen Verschlechterung der Textqualität gesprochen werden kann. Nachteile der jüngeren Generation scheinen eher dann zutage zu treten, wenn die Texte noch traditionell von Hand geschrieben werden müssen.

	1998 analog		**2023 analog**		**2023 digital**		
	M	*SD*	*M*	*SD*	*M*	*SD*	*F*
Textqualität[1]	3.46	.61	2.47	.56	3.37	.67	F(2, 138) = 46.8**
Differenziert[2]							
- Kohärenz	3.39	.75	2.40	.74	3.38	.68	F(2, 139) = 29.1**
- Originalität	3.49	.70	2.83	.70	3.55	.63	F(2, 139) = 16.4**
- Erzählstil	3.76	.53	2.87	.57	3.58	.55	F(2, 139) = 35.7**
- Form	3.51	.66	2.52	.74	3.41	.60	F(2, 139) = 31.5**
Textlänge	274.76	134.08	189.58	93.04	366.37	155.97	F(2, 139) = 21.2**

Tabelle 7: Deskriptive Statistiken und Ergebnisse der Varianzanalysen zur Überprüfung von Korpusunterschieden.
[1]Mittelwert aus holistischem (NAEP) Rating und analytischem Rating; [2]Differenzierte Einzelitems der analytischen Kodierung

4.2 | Unterschiede in der Textlänge

Neben den berichteten Unterschieden in der Textqualität ergeben sich zwischen den Textkorpora auch signifikante Abweichungen in der Textlänge. Die durchgeführten post-hoc Tests zeigen, dass sich in diesem Falle alle drei Gruppen signifikant voneinander unterscheiden. Mit Abstand am kürzesten waren die analogen Texte aus dem Jahre 2023 (durchschnittlich knapp 190 Wörter). Mit fast der doppelten Anzahl an Wörtern (366) am längsten waren die zur selben Zeit digital erstellten Texte, während die analogen Texte aus 1998 dazwischen rangierten (275 Wörter). Angesichts der Befunde zur Textqualität kommt dieses Ergebnis eher unerwartet. Denn obwohl den analogen Texten von 1998 und den digitalen Texten von 2023 im Mittel eine vergleichbare Qualität attestiert wird, weisen sie im Hinblick auf die Länge der Texte beachtliche Unterschiede zugunsten der digitalen Texte auf (Cohen's *d* = 0.63).

4.3 | Unterschiede in der Bewertung zwischen Mensch und KI

Der Einbezug der Urteile von ChatGPT führte zu vergleichbar guten Reliabilitätsschätzungen. Zwar fielen die paarweisen Übereinstimmungen der KI mit den Raterinnen etwas geringer aus als bei einem rein menschlichen Paar. Wenn man allerdings berücksichtigt, dass die KI im Gegensatz zu den menschlichen Raterinnnen keinerlei Training bzw. Feedback oder ergänzende Prompts anhand von Beispielbewertungen erhalten hat, sind die ermittelten Übereinstimmungen durchaus ermutigend. Es ist anzunehmen, dass die Inter-Rater-Reliabilitäten durch wenige differenzierende Prompts verbessert werden können. Zu den registrierten Abweichungen zählt, dass die Qualitätsurteile von ChatGPT deutlich geringer mit der Textlänge korrelieren als bei menschlichen Raterinnen. Die Beobachtungen legen nahe, dass ChatGPT auch bei sehr kurzen Texten zu dem Urteil gelangen kann, dass der Text beispielsweise formal-sprachlich korrekt und kohärent ist, während menschliche Rater bei ihrer Bewertung in Rechnung stellen, dass dies bei einer Textlänge von zwei oder drei Sätzen nicht hinreichend ist, um auf eine gute Fähigkeit zur Produktion formal-sprachlich korrekter und kohärenter Texte zu schließen. Solche Abweichungen sollten sich in der geplanten Hauptstudie mit ergänzenden Prompts und entsprechendem Training der KI relativ leicht beheben lassen.

5 | Zusammenfassende Diskussion und Ausblick

Das Ziel der vorliegenden Pilotstudie bestand darin zu untersuchen, ob sich im Vergleich mit einem Textkorpus von 1998 Hinweise darauf finden lassen, dass sich die Textproduktionsleistungen heutiger Schülerinnen und Schülern, wie häufig angenommen, tatsächlich verändert haben. Da Texte heute zunehmend digital verfasst werden, sollte darüber hinaus untersucht werden, ob hierbei das gewählte Schreibmedium (handschriftlich vs. digital) eine Rolle spielt. In einer Nebenfragestellung sollten zudem erste Erkenntnisse zur Eignung von KI zur Beurteilung der Textqualität gewonnen werden.

Um die Texte der drei einbezogenen Korpora vergleichen zu können, wurde ein holistisches und ein analytisches Rating empirisch überprüft. Die Befunde deuten darauf hin, dass beide Raster eine reliable Einschätzung der Textproduktionsleistungen ermöglichen.

In Bezug auf die Hauptfragestellung der Pilotstudie – die Überprüfung von Unterschieden zwischen Texten aus 1998 und analog sowie digital erstellten Texten aus 2023 – ergab sich ein etwas unerwartetes, aber durchaus plausibel erklärbares Muster. Der Blick auf handschriftlich erstellte Texte scheint die häufig geäußerten Befürchtungen zu bestätigen. Die Texte aus 2023 wurden substanziell schlechter bewertet und waren deutlich kürzer als in den beiden anderen Korpora. Bei digital erstellten Texten ergab sich hingegen ein ganz anderes Bild. Die Qualitätsbeurteilungen waren qualitativ durchweg mit den Texten aus 1998 vergleichbar. Darüber hinaus waren die digital erstellten Texte sogar substanziell länger. Die Befunde liefern demnach keine stützende Evidenz für die häufig geäußerte Befürchtung, dass sich die Textproduktionskompetenzen von Schülerinnen und Schülern in den letzten Jahrzehnten pauschal verschlechtert hätten.

Limitierend für diesen Befund ist in einer kleineren Pilotstudie die Tatsache, dass die Texte aus wenigen Klassen stammen und somit nicht ausgeschlossen werden kann, dass Unterschiede auch auf Differenzen in der jeweiligen bildungs- und spracherwerbsbiografischen Konstellation der Probandengruppe zurückzuführen sind. Daraus leitet sich die Notwendigkeit ab, die Befunde mit größeren Stichproben zu replizieren, in denen die Vergleichbarkeit der Gruppen noch mehr sichergestellt werden kann, was in der geplanten Hauptstudie auch unter

Berücksichtigung der medialen Vorerfahrung geschehen soll. Denn die Feststellung, dass die Schüler/innen auch heute noch vergleichbar gute narrative Texte hervorbringen wie vor 25 Jahren, wenn sie das digitale Schreibmedium verwenden dürfen, ist von hoher fachdidaktischer Relevanz. In einem größeren Korpus könnte man zudem noch differenzierter verschiedene Facetten der Textqualität oder auch Überarbeitungsprozesse untersuchen. Auch die Rolle des Textmusters sollte dabei bedacht werden: Interessanterweise wählten die Schüler bei diesem Aufgabensetting kaum Textmuster, die sie in der Sekundarstufe I in der 7./8. Jahrgangsstufe lernen, sondern griffen beim freien Schreiben auf das Erzählen zurück, für das ihnen möglicherweise eine vertraute, gut verankerte Normvorstellung aus der Grundschulzeit vorliegt. Insofern wäre es auch untersuchenswert, ob sich der Trend zu gleichbleibender Textqualität beim digitalen Schreiben auch bei anderen Textmustern (z.B. argumentierendes Schreiben) bestätigt.

Insgesamt sind die Befunde auch für die Praxis außerordentlich vielversprechend. Um die Textproduktionskompetenzen von Schülerinnen und Schülern zu verbessern, ist ein individuelles Feedback zur Qualität von Textentwürfen essenziell, aber enorm ressourcenaufwändig. Für die Lehrerschaft scheint sich hier durch den Einbezug von KI-Tools eine Entlastungsmöglichkeit abzuzeichnen, zumal die hier verwendeten Prompts nicht nur zu einer Bewertung geführt haben, sondern darüber hinaus differenzierte Verbesserungsvorschläge ausgegeben wurden. Noch waren nicht alle gegebenen Kommentare hilfreich, aber bei entsprechender Optimierung der Prompts ist zu erwarten, dass für die Schüler/innen sinnvolle Tipps und Hilfen generiert werden können, auch wenn diese weiterhin einen fachlichen Blick durch die Lehrkraft erfordern werden.

Eine auf diese neuen Möglichkeiten ausgerichtete Schreibdidaktik kann der Generation, die Schülerin „Hope“ eingangs beschrieb, Hoffnung machen: Da „heutzutage jeder einen anderen Style“ hat, könnte die digitale Transformation – mit kritisch-reflektierender, fachdidaktischer Begleitung – dazu beitragen, einen Unterricht zu ermöglichen, der die Schreibfähigkeiten motivierend weiterentwickelt und dabei den individuellen Voraussetzungen noch besser gerecht wird.

Literaturverzeichnis

Androutsopoulos, J. (2007). Neue Medien – neue Schriftlichkeit? *Mitteilungen des Deutschen Germanistenverbandes, 54*(1), 72-94.

Augst, G. (2010). Zur Ontogenese der Erzählungskompetenz in der Primar- und Sekundarstufe. In T. Pohl & T. Steinhoff (Hrsg.), *Textformen als Lernformen* (S. 63-95). Gilles & Francke.

Bachmann, T., Becker-Mrotzek, M. (2010). *Schreibaufgaben situieren und profilieren. In T. Pohl & T. Steinhoff (Hrsg.), Textformen als Lernformen.* (S. 191–209). Gilles & Francke.

Becker-Mrotzek, M., & Böttcher, I. (2014). *Schreibkompetenz entwickeln und beurteilen* (5. Aufl.). Cornelsen.

Becker-Mrotzek, M., & Grabowski, J. (Hrsg.). (2022). *Schreibkompetenz in der Sekundarstufe: Theorie, Diagnose und Förderung*. Waxmann.

Becker-Mrotzek, M., Grabowski, J., & Steinhoff, T. (Hrsg.). (2017). *Forschungshandbuch empirische Schreibdidaktik*. Waxmann.

Berg, K., & Romstadt, J. (2021). Reifeprüfung – das Komma in Abituraufsätzen von 1948 bis heute. In *Die Sprache in den Schulen - Eine Sprache im Werden. Dritter Bericht zur Lage der deutschen Sprache* (S. 205-238). Erich Schmidt.

Betzel, D., & Steinig, W. (2016). Wortschatz und Orthographie. In B. Mesch & C. Noack (Hrsg.), *System, Norm und Gebrauch – drei Seiten einer Medaille? Orthographische Kompetenz und Performanz im Spannungsfeld zwischen System, Norm und Empirie* (S. 24-52). Schneider.

Böhme, K., Schipolowski, S., Canz, T., Krelle, M., & Bremerich-Vos, A. (2017). Kompetenzstufenmodelle im Bereich Schreiben. In M. Becker-Mrotzek, J. Grabowski, & T. Steinhoff (Hrsg.), *Forschungshandbuch empirische Schreibdidaktik* (S. 55–74). Waxmann.

Böhme, K., Bremerich-Vos, A., & Robitzsch, A. (2009). Aspekte der Kodierung von Schreibaufgaben. In D. Granzer, O. Köller, A. Bremerich-Vos, M. Van den Heuvel-Panhuizen, K. Reiss, & G. Walther (Hrsg.), *Bildungsstandards Deutsch und Mathematik. Leistungsmessung in der Grundschule.* (S. 290–329). Beltz.

Bolz, N. W. (1993). *Am Ende der Gutenberg Galaxis*. Fink.

Bräunig, S., & Holberg, S. (2024). Formatives Feedback: ein Überblick. *Schule NRW*, 07/08-24. Ministerium für Schule und Bildung NRW. https://www.schule.nrw.de/schulnrw

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. Aufl.). L. Erlbaum Associates.

Dürscheid, C., Wagner, V., & Brommer, S. (2010). *Wie Jugendliche schreiben*. *Schreibkompetenz und neue Medien.* De Gruyter.

Eickelmann, B., Bos, W., Gerick, J., Goldhammer, F., Schaumburg, H., Schwippert, K., Senkbeil, M., & Vahrenhold, J. (Hrsg.). (2019). *ICILS 2018. #Deutschland. Computer- und informationsbezogene Kompetenzen von Schülerinnen und Schülern im zweiten internationalen Vergleich und Kompetenzen im Bereich Computational Thinking*. Waxmann. Verfügbar unter https://kw.uni-paderborn.de/fileadmin-kw/fakultaet/Institute/erziehungswissenschaft/Schulpaedagogik/ICILS_2018_Deutschland_Berichtsband.pdf

Fix, M. (2025). *Texte schreiben – Schreibprozesse im Deutschunterricht* (3. Aufl.). Brill/Schöningh UTB.

Fix, M., & Melenk, H. (2000). *Schreiben zu Texten – Schreiben zu Bildimpulsen. Das Ludwigsburger Aufsatzkorpus*. Schneider.

Fletcher, R., & Nielsen, R. K. (2024). *What does the public in six countries think of generative AI in news?* Reuters Institute for the Study of Journalism. https://doi.org/10.60625/risj-4zb8-cg87

Grabowski, J. (2022). Operationalisierungen der Textqualität. In M. Becker-Mrotzek & J. Grabowski (Hrsg.), *Schreibkompetenz in der Sekundarstufe: Theorie, Diagnose und Förderung* (S. 133-148). Waxmann.

Grimm, H. (2003). *Veränderungen der Sprachfähigkeiten Jugendlicher: eine Untersuchung zu Abituraufsätzen von den Vierziger- bis zu den Neunzigerjahren*. Lang.

Institut zur Qualitätsentwicklung im Bildungswesen (IQB). (2023). Globalskalen zur Beurteilung von Schülertexten. *[unveröffentlichtes Material]*

Kruse, N., Reichardt, A., Herrmann, M., Heinzel, F., & Lipowsky, F. (2012). Zur Qualität von Kindertexten. Entwicklung eines Bewertungsinstruments in der Grundschule. *Didaktik Deutsch, 17*(32), 87-110.

Kultusministerkonferenz. (2021). *Lehren und Lernen in der digitalen Welt: Ergänzungspapier zur Strategie der Kultusministerkonferenz „Bildung in der digitalen Welt“*. https://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2021/2021_12_09-Lehren-und-Lernen-Digi.pdf

Kultusministerkonferenz (Hrsg.). (2022). *Bildungsstandards für das Fach Deutsch. Erster Schulabschluss (ESA) und Mittlerer Schulabschluss (MSA)*. Beschluss der Kultusministerkonferenz vom 15.10.2004 und vom 04.12.2003, i.d.F. vom 23.06.2022. Verfügbar unter https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2022/2022_06_23-Bista-ESA-MSA-Deutsch.pdf

Malkewitz, C. P., Schwall, P., Meesters, C., & Hardt, J. (2023). Estimating reliability: A comparison of Cronbach's α, McDonald's ωt and the greatest lower bound. *Social Sciences & Humanities Open, 7*(1), 100368. https://doi.org/10.1016/j.ssaho.2022.100368

Mathiebe, M. (2018). *Wortschatz und Schreibkompetenz. Bildungssprachliche Mittel in Schülertexten der Sekundarstufe I*. Waxmann.

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Neumann, A. (2012). Blick(e) auf das schulische Schreiben. Erste Ergebnisse aus IMOSS. *Didaktik Deutsch, 17*, 63-85.

Neumann, A. (2017). Zugänge zur Bestimmung von Textqualität. In M. Becker-Mrotzek, J. Grabowski, & T. Steinhoff (Hrsg.), *Forschungshandbuch empirische Schreibdidaktik* (S. 203-219). Waxmann.

Nussbaumer, M., & Sieber, P. (1995). Über Textqualitäten reden lernen – z.B. anhand des „Züricher Textanalysrasters". *Diskussion Deutsch, 141*, 36-52.

Pander Maat, H., de Glopper, K., Raaijmakers, K., Veerbeek, J., & Vermeulen, D. (2023). Fleshing out your text: How elaboration and contextualization moves differentially predict writing quality. *Journal of Writing Research, 15*(2), 363–393. https://doi.org/10.17239/jowr-2023.15.02.05

Persky, H. R., Daane, M. C. & Jin, Y. (2003). The Nation´s Report Card. Writing 2002 (NCES 2003-529). Verfügbar unter: https://nces.ed.gov/nationsreportcard/pdf/main2002/2003529a.pdf

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55, 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Rödel, M. (2020). *Schule, Digitalität & Schreiben. Impulse für einen souveränen Deutschunterricht*. Stauffenburg.

Schoonen, R. (2012). The validity and generalizability of writing scores. The effect of rater, task, and language. In E. van Steendam, M. Tillema, G. Rijlaarsdam, & H. van den Berg (Hrsg.), *Measuring writing. Recent insights into theory, methodology, and practices* (S. 23-32). Emerald.

Sieber, P. (1998). *Parlando in Texten. Zur Veränderung kommunikativer Grundmuster in der Schriftlichkeit*. Niemeyer.

Steinig, W., Betzel, D., Geider, F. J., & Herbold, A. (2009). *Schreiben von Kindern im diachronen Vergleich*. Waxmann.

Steinig, W., & Betzel, D. (2014). Schreiben Grundschüler heute schlechter als vor 40 Jahren? Texte von Viertklässlern aus den Jahren 1972, 2002 und 2012. In A. Plewnia & A. Witt (Hrsg.), *Sprachverfall? Dynamik – Wandel – Variation* (S. 353-371). De Gruyter.Storrer, A. (2018). „Interaktionsorientiertes Schreiben im Internet". In A. Deppermann & S. Reineke (Hrsg.), *Sprache im kommunikativen, interaktiven und kulturellen Kontext* (S. 219-244). De Gruyter. Open Access.

Sturm, Afra (2023): Schreiben wirksam fördern. Lernarrangements und Unterrichtsentwicklung für alle Stufen. Bern: hep Verlag.

Wild, J. (2020). *Schriftliche Erzählfähigkeiten diagnostizieren und fördern. Eine empirische Studie zum Erfassen von Textqualität in der Primar- und Sekundarstufe*. Waxmann.

Wirtz, M. A., & Caspar, F. (2002). *Beurteilungsübereinstimmung und Beurteilungsreliabilität: Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystem und Ratingskalen*. Hogrefe.

Wurzenberger, G. (2016). *Intermedialer Style. Kultureller Kontext und Potenziale im literarischen Schreiben Jugendlicher*. Verfügbar unter: https://www.transcript-verlag.de/media/pdf/18/14/94/oa97838394334614tcwdeXRJNl9J.pdf

Autor*inneninformation

Prof. Dr. Nadine Anskeit ist Professorin für deutsche Sprache und ihre Didaktik an der Pädagogischen Hochschule Karlsruhe

Prof. Dr. Nadine Anskeit
Pädagogische Hochschule Karlsruhe
Institut für deutsche Sprache und Literatur
Bismarckstraße 10
76133 Karlsruhe
nadine.anskeit@ph-karlsruhe.de

Prof. Dr. Dirk Betzel und Prof. Dr. Martin Fix sind Professoren für deutsche Sprache und ihre Didaktik an der Pädagogischen Hochschule Ludwigsburg.

Prof. Dr. Dirk Betzel
Pädagogische Hochschule Ludwigsburg
Institut für deutsche Sprache und Literatur
Reuteallee 46
71634 Ludwigsburg
dirk.betzel@ph-ludwigsburg.de

Prof. Dr. Martin Fix
Pädagogische Hochschule Ludwigsburg
Institut für deutsche Sprache und Literatur
Reuteallee 46
71634 Ludwigsburg
fix@ph-ludwigsburg.de

Prof. Dr. Marco Ennemoser ist Professor für Psychologie für den sonderpädagogischen Förderschwerpunkt Kommunikation und Sprache an der Pädagogischen Hochschule Ludwigsburg

Prof. Dr. Marco Ennemoser
Pädagogische Hochschule Ludwigsburg
Institut II: Sonderpädagogische Förderschwerpunkte
Reuteallee 46
71634 Ludwigsburg
ennemoser@ph-ludwigsburg.de