

Stephan Schicker, Muhammed Akbulut

ChatGPT – maschinelle und menschliche Textsortenkompetenz

Sprachliche Handlungsmuster & Text(sorten)kompetenz, Hg. v. Schicker, Miškulin Saletović, 2023, S. 169-197
<https://doi.org/10.25364/978390337426311>

Dieses Werk ist lizenziert unter einer Creative Commons Namensnennung 4.0 International Lizenz,
ausgenommen von dieser Lizenz sind Abbildungen, Screenshots und Logos.

Stephan Schicker, Universität Graz, stephan.schicker@uni-graz.at, ORCID 0000-0002-3948-7632
Muhammed Akbulut, Universität Graz, muhammed.akbulut@uni-graz.at

Zusammenfassung

Der vorliegende Beitrag untersucht, wie 40 angehende Deutschlehrkräfte einen argumentativen Text von ChatGPT 3.5 anhand des Beurteilungsrasters für die österreichische Matura im Vergleich zu einem Text, der von einem Menschen geschrieben wurde, beurteilen. Hierzu wird zunächst der öffentliche und wissenschaftliche Diskurs zu den aktuellen Entwicklungen in der KI-Forschung skizziert und es werden die Ergebnisse bisheriger Studien zu den Schreibfähigkeiten von ChatGPT dargestellt. Auf dieser Grundlage werden die beiden Texte zunächst qualitativ analysiert, anschließend werden die quantitativen Ergebnisse der Textbeurteilung durch die angehenden Lehrkräfte vorgestellt und diskutiert. Die Ergebnisse der Untersuchung zeigen, dass keinem bzw. keiner Proband*in von sich aus aufgefallen wäre, dass es sich bei einem der Texte um einen Text von ChatGPT handelt. Darüber hinaus zeigt die Untersuchung, dass ChatGPT in der Lage ist, einen kohärenten, inhaltlich elaborierten und sprachlich weitgehend korrekten argumentativen Text zu produzieren, der von den angehenden Lehrkräften im Mittel mit der Gesamtnote *Gut* beurteilt wird. Gleichzeitig übernimmt die KI beim Paraphrasieren der Textbeilage ganze Textpassagen fast wortwörtlich und zeigt einen syntaktisch und lexikalisch höchst repetitiven Sprachgebrauch.

Schlagwörter: ChatGPT, Textsortennormen, Textbeurteilung

Abstract

This paper investigates how 40 prospective German teachers assess an argumentative text from ChatGPT 3.5 using the SRDP assessment grid for the Austrian Matura in comparison to a human text. For this purpose, the public and scientific discourse on current developments in AI research is first outlined and the results of previous studies on the writing abilities of ChatGPT are presented. On this basis, the two texts are first analysed qualitatively. Then the quantitative results of the text assessment of the pre-service teachers are presented and discussed. The results of the study first show that none of the pre-service teachers would have noticed that one of the texts is a ChatGPT text. Moreover, the study indicates that ChatGPT is able to produce a coherent, elaborate, and linguistically largely correct argumentative text, which is assessed on average with the overall grade *Good*. At the same time, when paraphrasing the source text, the AI copies entire text passages almost word-for-word and shows a syntactically and lexically highly repetitive use of language.

Keywords: ChatGPT, text type standards, text assessment

1 Einleitung

ChatGPT ist ein maschinelles Lernalgorithmus-Modell, das auf Natural Language Processing (NLP) spezialisiert ist. Es wurde durch die Verwendung von Deep Learning-Methoden trainiert, die es ihm ermöglichen, menschenähnliche Texte zu generieren. Dieses Modell hat eine massive Menge an Texten aus verschiedenen Quellen und in verschiedenen Sprachen als Trainingsdaten erhalten, die bereits den Textsortennormen entsprechen. Während des Trainingsprozesses analysiert das Modell die Struktur und den Inhalt dieser Texte, indem es nach Mustern und Regeln sucht, die es verwenden kann, um neue Texte zu generieren. Dabei lernt es, welche Wörter und Phrasen in bestimmten Kontexten verwendet werden, wie Satzstrukturen aufgebaut sind und wie Texte in Bezug auf Stil, Ton und Zweck formuliert werden. Wenn das Modell eine Aufgabe hat, einen Text zu generieren, der den Textsortennormen entspricht, nutzt es seine internen Modelle, um einen Text zu produzieren, der den Schulungsdaten entspricht. Es überprüft den generierten Text jedoch nicht direkt auf Übereinstimmung mit den Textsortennormen, sondern produziert ihn aufgrund seiner Schulung und seiner internen Modelle.

Dieser gesamte bisherige Textabschnitt wurde nicht von einem Menschen geschrieben, sondern ist eine Antwort von ChatGPT 3.5 auf folgende Fragen:

Wie schafft es ChatGPT Texte zu schreiben, die den Textsortennormen entsprechen?

Wie geht ChatGPT dabei vor? Gleicht er die Texte mit Textsortennormen ab?

Die informative und zugleich eloquente Antwort von ChatGPT zeigt auf, wozu künstliche Intelligenz bereits in der Lage ist. Die rasanten Entwicklungen in der KIForschung haben in der didaktischen Forschung und Praxis jedoch auch zu der berechtigten Sorge geführt, dass Schüler*innen und Studierende in Zukunft dazu verleitet werden könnten, Hausaufgaben und Prüfungen zu einem Gutteil von ChatGPT schreiben zu lassen. Doch wie gut kann ChatGPT schulische Schreibaufträge wirklich bearbeiten und inwiefern entsprechen die von ihm verfassten Texte gängigen Textsortenkriterien? Wie bewerten Lehrende Texte, die von ChatGPT geschrieben wurden? Um einen empirisch fundierten Beitrag zur Klärung dieser Fragen zu leisten, wird daher im Folgenden eine Studie vorgestellt, bei der angehende Deutschlehrkräfte (N = 40) gebeten wurden, eine von ChatGPT und eine von einer Studierenden geschriebene Erörterung vergleichend zu beurteilen.

2 Forschungsstand

2.1 ChatGPT – aktuelle Entwicklungen und Diskussionen

Seit ChatGPT Ende November 2022 der Öffentlichkeit kostenlos zugänglich gemacht wurde und eindrucksvoll gezeigt hat, wozu künstliche Intelligenz bereits in der Lage ist, wurde im öffentlichen und wissenschaftlichen Diskurs eine hitzige Debatte über die Potenziale und Gefahren von künstlicher Intelligenz entfacht. Auf der einen Seite wird ChatGPT als der größte technologische Meilenstein seit Erfindung der Suchmaschinen beschrieben (Susnjak, 2022, S. 3) und es wird davon ausgegangen, dass künstliche Intelligenz gerade in Branchen, in denen mit großen Datensätzen gearbeitet wird (z.B. Finanzwesen, Gesundheitswesen) zu einem enormen Produktivitätsgewinn führen könnte (Aljanabi & ChatGPT, 2023, S. 16). Auf der anderen Seite wird befürchtet, dass ChatGPT zu Verwerfungen am Arbeitsmarkt führen und Softwareentwickler*innen, Anwalt*innen, Journalist*innen und Schriftsteller*innen den Job streitig machen, zur Verbreitung von Fake News beitragen und Schüler*innen und Studierende zum Vortäuschen von Prüfungsleistungen verleiten könnte (Taecharungroj, 2023, S. 7). Im akademischen Sektor besteht zudem die Sorge, dass die seit Jahren zu beobachtende Publikationsinflation und daraus resultierende Missstände im Wissenschaftsbetrieb (Ghost-Autor*innen-schaft, Fake-Forschung, Fake-Journals etc.) durch ChatGPT noch einmal befeuert werden könnten (Fire, 2023).

Insofern überrascht es nicht, dass die meisten empirischen Studien, die seit der Veröffentlichung von ChatGPT publiziert wurden, sich mit der Frage befassen, inwieweit das Tool für das Verfassen wissenschaftlicher Publikationen eingesetzt oder auch missbraucht werden kann. So haben etwa Aydin und Karaarslan (2022) untersucht, inwiefern ChatGPT Wissenschaftler*innen beim Schreiben von *literature reviews* entlasten kann. Angesichts des hohen Publikationsdrucks, dem Forscher*innen im aktuellen Wissenschaftsbetrieb ausgesetzt sind, und der schier unüberblickbaren Menge an Forschungsliteratur, die in vielen Disziplinen tagtäglich produziert wird, sehen sie darin durchaus ein förderliches Potenzial von ChatGPT. Die Ergebnisse ihrer Studie sind jedoch etwas ernüchternd: Wird ChatGPT zur Paraphrasierung von wissenschaftlichen Texten eingesetzt, sind dessen Formulierungen derart nah am Original, dass Plagiatsoftwarens zahlreiche Treffer anzeigen. Allerdings können die Autor*innen zeigen, dass ChatGPT trotzdem den wissenschaftlichen Schreibprozess entlasten kann, indem die KI etwa nützliche Textbausteine liefert, wenn ihm Fragen gestellt werden, die in einem wissenschaftlichen Text bearbeitet werden sollen. In einer weiteren Studie haben Gao et al. (2022) untersucht, inwieweit wissenschaftliche Abstracts, die von ChatGPT geschrieben wurden, von

KI-Outputdetektoren, Plagiatsoftwares und menschlichen Reviewer*innen als solche erkannt werden. Die Ergebnisse dieser Studie zeigen zunächst, dass ChatGPT durchaus dazu in der Lage ist, kohärente und inhaltlich adäquate Abstracts zu produzieren und dabei auch keine Plagiate produziert. Es wird aber auch deutlich, dass KI-Outputdetektoren sehr zuverlässig erkennen können, ob ein Abstract von einem Menschen oder von ChatGPT verfasst wurde. Auch menschliche Reviewer*innen konnten in der Studie von Gao et al. (2022) 68% der künstlichen Abstracts als solche erkennen. Es ist daher anzunehmen, dass in Zukunft neben Plagiatsoftwares auch KI-Outputdetektoren vermehrt standardmäßig zum Einsatz kommen werden, um die Integrität des wissenschaftlichen Publikationssystems sicherzustellen. Nichtsdestominder ist es unumgänglich, dass Forschungseinrichtungen klare Richtlinien dazu erarbeiten, in welchem Ausmaß und auf welche Art und Weise künstliche Intelligenz im wissenschaftlichen Publikationsprozess eingesetzt werden darf, ohne dass die Prinzipien guter wissenschaftlicher Praxis verletzt werden.

In einer Reihe weiterer Untersuchungen wurde der Frage nachgegangen, inwiefern ChatGPT Prüfungen an Schulen oder Universitäten erfolgreich absolvieren kann. So hat etwa Susnjak (2022) im Rahmen einer explorativen Studie ChatGPT Aufgaben zum kritischen Denken in unterschiedlichen Fachbereichen entwickeln und anschließend selbst beantworten lassen. Im Rahmen einer kriteriengeleiteten Analyse kommt Susnjak (2022) zu dem Ergebnis, dass die Antworten von ChatGPT klar, stringent und fachlich elaboriert sind und ChatGPT somit eine ernsthafte Bedrohung für die Integrität von Online-Prüfungsformaten darstellt. Diese explorativen Erkenntnisse werden in der Studie von Yeadon et al. (2022) experimentell bestätigt. Die Autor*innen dieser Studie können zeigen, dass ChatGPT ein akkreditiertes Physikexamen, das aus fünf offenen Fragen besteht und an britischen Universitäten als Modulabschlussprüfung eingesetzt wird, mit durchschnittlich denselben Noten wie echte Studierende absolviert. In Österreich wiederum hat die Schüler*innenvertretung ChatGPT zur Zentralmatura antreten lassen und auch hier besteht die künstliche Intelligenz in allen drei Fächern (Deutsch, Mathematik, Englisch) die Matura, wenn auch nur mit durchschnittlichen bis schwachen Noten (Thaler, 2023). Dementsprechend mahnt die Schüler*innenvertretung Maßnahmen ein, um die Integrität der Zentralmatura zu wahren und weist darauf hin, dass abzusehen ist, dass die nächste Version GPT 4 noch deutlich bessere Ergebnisse erzielen wird als GPT 3.5. Tatsächlich haben die Entwickler*innen von ChatGPT (OpenAI, 2023) kürzlich einen Report veröffentlicht, der zeigt, dass GPT 4 selbst dazu in der Lage ist, Aufgaben zu lösen, die anspruchsvolle Bildinterpretationen beinhalten. Im Gegensatz zu GPT 3.5 kann GPT 4 zudem längere und bessere Texte schreiben, erfindet kaum noch Fakten und schneidet aufgrund seiner erweiterten Fähigkeiten in zahlreichen

Uni-Prüfungen von Medizin über Rechtswissenschaften bis hin zu Psychologie zum Teil hervorragend ab. Insofern ist davon auszugehen, dass die rasanten Fortschritte in der KI-Forschung einen nachhaltigen Effekt darauf haben werden, wie in Zukunft an Schulen und Universitäten Leistungsfeststellungen durchgeführt werden.

In weitaus geringerem Ausmaß wurde bislang der Frage nachgegangen, inwieweit ChatGPT Studierende oder Schüler*innen beim Schreibprozess unterstützen kann. Hierzu haben Bašić et al. (2023) eine experimentelle Studie durchgeführt, um herauszufinden, ob es Master-Studierenden mithilfe von ChatGPT gelingt, bessere Essays im Fach *Forensic Sciences* zu schreiben. Die Essays sollten 800–1000 Wörter lang sein und die Vor- und Nachteile biometrischer Identifikation in den forensischen Wissenschaften diskutieren. Die Experimentalgruppe wurde dazu aufgefordert, den Text mithilfe von ChatGPT zu schreiben. Wie genau die KI genutzt wurde, war den Studierenden allerdings selbst überlassen und es gab auch keine Schulung zur optimalen Verwendung von ChatGPT beim *essay writing*. Tatsächlich zeigen die Ergebnisse dieser Studie leicht höhere Gesamtscores in der Kontrollgruppe. Überraschend ist auch, dass die Experimentalgruppe in keiner einzigen Subkategorie die Kontrollgruppe übertrifft: „The students did not deliver higher quality content, did not write faster, nor had a higher degree of authentic text“ (Bašić et al., 2023, S. 7). Als mögliche Gründe für das schwache Abschneiden der Experimentalgruppe trotz der Hilfe von ChatGPT führen die Autor*innen an, dass die Studierenden zum Testzeitpunkt wohl noch zu wenig Erfahrung mit ChatGPT und seiner Funktionsweise gehabt haben dürften und daher die volle Bandbreite des Tools nicht adäquat nutzen konnten, zumal die mangelnde Vertrautheit mit dem Tool auch zu Zeitnot während der Testung geführt haben könnte. Darüber hinaus äußerten einige Studierende ihr Unbehagen, aufgrund von ChatGPT unbeabsichtigt zu plagieren und berichteten von Schwierigkeiten, die Textbausteine von ChatGPT mit selbst geschriebenen Textteilen zu verknüpfen. Aufgrund der geringen Stichprobengröße (n=9 pro Gruppe) sind die Ergebnisse von Bašić et al. (2023) allerdings mit Vorsicht zu interpretieren.

Was in den bisherigen Studien ebenfalls zu kurz kommt, ist eine differenzierte und empirisch fundierte Auseinandersetzung mit der sprachlichen Qualität der Texte, die ChatGPT produziert. Hierzu liegen bislang vor allem journalistische Arbeiten und Blogbeiträge vor (z.B. Haberkorn, 2023; Kiely, 2023). So hat etwa die österreichische Tageszeitung *Der Standard* (Haberkorn, 2023) ChatGPT Matura-Schreibaufträge aus den Fächern Deutsch und Englisch vorgelegt und die von der KI erstellten Texte anschließend von Lehrkräften beurteilen lassen. Die Ergebnisse fallen je nach Aufgabentyp höchst unterschiedlich aus: Einen englischen Blogbeitrag (250

Wörter) zum Thema „Second Screening“ meistert die KI mühelos, da sie systematisch auf die Aufgabenstellung eingeht und einen strukturell und inhaltlich ansprechenden Text auf dem geforderten sprachlichen Niveau verfasst. Weitaus schlechter sind die Ergebnisse bei einer Deutschaufgabe, die viel Interpretationsleistung und Kreativität erfordert: Wird ChatGPT gebeten, einen Liedtext zu interpretieren, werden die Schwächen der KI offenbart. So identifiziert ChatGPT etwa Begriffe für Versmaße (Daktylus, Jambus, Trochäus) fälschlicherweise als Anspielung an die Pflanzenwelt. Darüber hinaus unterlaufen der KI vor allem bei der inhaltlichen Interpretation des Gedichts viele Fehler, „die für Schülerinnen und Schüler einer achten Klasse eher unüblich wären“ (Haberkorn, 2023). Etwas besser schneidet ChatGPT hingegen beim Verfassen eines Kommentars zu einem Zeitungsartikel (Thema: Klimaproteste) ab. Der Text wird von den Lehrkräften als strukturell und inhaltlich überzeugend, allerdings für einen Kommentar als „etwas zu brav“ (Haberkorn, 2023) bewertet, da die KI u.a. auf einen interessanten Einstieg und Zuspitzung durch den Einsatz rhetorischer Mittel verzichtet.

Insgesamt deuten die bisherigen Untersuchungen zu den Schreibfähigkeiten von ChatGPT in akademischen und schulischen Kontexten darauf hin, dass die KI bei Aufgaben, die kein hohes Maß an Interpretation und Kreativität erfordern, inhaltlich und strukturell hochwertige Texte produziert. Es gibt bislang allerdings keine Studien, die sich vertiefend mit der sprachlichen Qualität der Texte von ChatGPT auseinandersetzen.

2.2 Textsorte: Erörterung

Da die Proband*innen der vorliegenden Studie gebeten wurden, eine Erörterung von ChatGPT und einer Studierenden vergleichend zu beurteilen, wird im Folgenden kurz skizziert, wie die Textsorte charakterisiert werden kann und wie sich der didaktische Diskurs dazu gestaltet.

Fragt man zunächst ChatGPT, wie man eine gute Erörterung schreibt und auf welche Textsortenmerkmale man dabei achten muss, erhält man folgende Antwort:

Die Erörterung ist eine Textsorte, bei der es darum geht, ein bestimmtes Thema zu diskutieren und dabei verschiedene Aspekte und Meinungen zu berücksichtigen. Die Erörterung ist ein wichtiger Bestandteil von Schulaufgaben und Prüfungen und kommt auch im Alltag häufig vor, zum Beispiel bei Diskussionen in Meetings oder bei politischen Debatten.

Die Merkmale der Textsorte Erörterung sind:

- *Thematische Eingrenzung: Die Erörterung beschäftigt sich mit einem bestimmten Thema, das in der Regel klar definiert und eingegrenzt ist.*

- *Präsentation von Argumenten: In einer Erörterung werden verschiedene Argumente zu einem Thema präsentiert und bewertet.*
- *Berücksichtigung verschiedener Meinungen: Eine Erörterung berücksichtigt in der Regel verschiedene Meinungen und Perspektiven zu einem Thema.*
- *Logische Struktur: Eine Erörterung hat in der Regel eine logische Struktur, die sich aus der Einleitung, dem Hauptteil und dem Schluss zusammensetzt.*
- *Verwendung von Beispielen: In einer Erörterung werden häufig Beispiele verwendet, um die Argumentation zu verdeutlichen.*
- *Persönliche Meinung: In einer Erörterung wird in der Regel auch die persönliche Meinung des Verfassers zum Ausdruck gebracht.*

Es ist wichtig, dass die Erörterung gut strukturiert und argumentativ ist und dass alle genannten Aspekte berücksichtigt werden. So kann sichergestellt werden, dass die Erörterung gut verständlich und überzeugend ist.

An dieser eher allgemein gehaltenen Textsortenbeschreibung ist auffallend, dass die von ChatGPT dargestellten Merkmale der Textsorte Erörterung mit Einschränkungen beinahe auf alle argumentativen Textsorten (z.B. Kommentar, Meinungsrede) anwendbar sind. Lediglich das angedeutete Abwägen von Pro- und Kontrapositionen (die Erörterung berücksichtigt in der Regel verschiedene Meinungen und Perspektiven) lässt sich als besonders typisch für die Erörterung einstufen. Diese Merkmalsbeschreibungen von ChatGPT bleiben damit insgesamt weniger konkret formuliert als vergleichbare Merkmalsbeschreibungen in Schulbüchern und schreibdidaktischer Literatur.

Aus schreibdidaktischer Sicht gilt zu berücksichtigen, dass die Operationalisierung von Textqualität bei der „schulischen“ Textsorte Erörterung aufgrund ihres artifiziellen Charakters eine besondere Herausforderung darstellt. Denn „außerhalb der Schule schreibt niemand Erörterungen“ (Feilke, 2008, S. 5), d.h. die Erörterung ist als didaktische Gattung ein speziell geschaffener (artifizieller) Lerngegenstand, zu dem es keine Referenztexte im „wirklichen Leben“ gibt. Insofern ist es naheliegend, in der Institution Schule nach (normativen) Vorgaben bzw. Merkmalsbeschreibungen für die Textsorte zu suchen: Im Rahmen einer Analyse von Schicker (2021) wurden Schulbücher und (Schreib-)Ratgeber in Bezug auf ihre normativen Vorgaben für die Textsorte Erörterung untersucht. Dabei wurde festgestellt, dass meist zwischen drei verschiedenen Arten der Erörterung, *textgebundener Erörterung*, *literarischer Erörterung* und *freier Erörterung*, unterschieden wird. Während der Schreibauftrag für die freie Erörterung nur aus einer Fragestellung besteht, sind bei der textgebundenen Erörterung zusätzlich Inputtexte (in der Regel bis zu

zwei) gegeben. Die literarische Erörterung kann als eine spezifische Form der textgebundenen Erörterung verstanden werden, bei der eine Fragestellung zu einem literarischen Text im Mittelpunkt des erörternden Schreibens steht. Im österreichischen Schulkontext hat sich in den letzten Jahren die textgebundene Erörterung aufgrund der spezifischen Aufgabenkonstruktion¹ der standardisierten Reifeprüfung weitgehend als Standard durchgesetzt (Schicker, 2021).

Üblicherweise wird in Bezug auf die Auseinandersetzung mit Sachverhalten in erörternden Texten im schulischen Kontext die Unterscheidung zwischen *linearer Erörterung* und *dialektischer (auch kontroverser) Erörterung* angeführt. In dieser spiegeln sich auch zwei unterschiedliche Funktionen des Argumentierens wider (persuasives Argumentieren, dessen Funktion es ist, eine andere Person von seiner eigenen Position zu überzeugen, und epistemisch-heuristisches Argumentieren, das darauf abzielt, bereits bestehendes Wissen neu zu strukturieren bzw. auch neues Wissen zu erwerben) (Domenech & Petersen, 2018, S. 109).

Die – auch in der vorliegenden Erhebung eingesetzte – dialektische Form der Erörterung sieht dann laut der Schulbuchanalyse von Schicker (2021) das Abwägen von Pro- und Kontraargumenten im Hauptteil der Erörterung vor. Durch die Auseinandersetzung mit Pro- und Kontraargumenten sollen die Lernenden im Schlussteil zu einer eigenen Schlussfolgerung bzw. einer „eigenen Meinung“ zur strittigen Frage gelangen. Schulbücher führen für die Realisierung der dialektischen Erörterung oft ein starres Muster („These Pro-Kontra-Pro-Kontra...-Fazit“ oder „These Pro-Pro-Pro Kontra-Kontra-Kontra Fazit“) an, welches in der Schreibdidaktik kritisiert wird (Fix, 2006, S. 104). Denn das mit diesem starren Muster einhergehende Abwägen von Argumenten führt zu einer starken kontrastiven Trennung von Pro- und Kontraargumenten und damit zur „Abspaltung der meinungsbildenden Potenziale“ (Feilke, 2013, S. 128) des argumentierenden Schreibens. Zudem scheint keine Einigkeit darüber zu bestehen, wie das Ergebnis der Meinungsbildung im Schlussteil beschaffen sein soll. Während der Textsortenkatalog der österreichischen Reifeprüfung (Bundesministerium für Bildung, Wissenschaft und Forschung, 2020) von einer „Einordnung in einen größeren Zusammenhang und eventuell Ausblick“ spricht, zeigen Untersuchungen von Winkler (2003), dass Lernende laut Selbstauskunft bei der Meinungsbildung in Erörterungen am Schluss vor allem versuchen, eine Meinung wiederzugeben, von der sie glauben, dass sie die Lehrperson für gut oder angemessen befindet. Von ähnlichen Befunden berichtet auch Feilke (2008, S. 6) anhand von

1 Bei der Konstruktion von Aufgabenstellungen sieht die österreichische Reifeprüfung für die Unterrichtssprache bei allen Textsorten zwingend mindestens eine Textbeilage vor.

ausgewählten Beispielen aus Foren, in denen Schüler*innen sich über die Problematik einer eindeutigen Positionierung im Rahmen der Meinungsbildung unterhalten: Dort schreiben Schüler*innen, dass eine „eigene Meinung (wenigstens ein bisschen) [...] auf jeden Fall immer gut ist“ bzw. dass manche Lehrpersonen beim Fehlen einer (klar akzentuierten) eigenen Meinung „sogar Abzüge machen“ würden (Feilke, 2008, S. 6). Diese Beispiele deuten darauf hin, dass in Bezug auf die konkrete Ausgestaltung der eigenen Positionierung im Schlussteil unterschiedliche Auffassungen unter Lehrpersonen (und folglich auch unter Lernenden) vorliegen: Es scheint keinen Konsens darüber zu geben, ob man sich überhaupt klar für oder gegen etwas positionieren muss oder ob eine differenzierte Meinungsbildung im Sinne einer Kompromissfindung auch möglich ist.

3 Empirische Studie

Im Zuge der empirischen Erhebung wurden 40 angehende Lehrkräfte gebeten, zwei Erörterungen zu derselben Schreibaufgabe (s. Kap. 3.1) mit dem Beurteilungsraster für die Standardisierte Reife- und Diplomprüfung (SRDP) zu beurteilen, wobei einer der beiden Texte von ChatGPT und der andere von einer angehenden Deutschlehrkraft verfasst wurde. Der Text der angehenden Lehrkraft stammt aus einem Datenkorpus, das im Wintersemester 2022/23 im Rahmen eines schreibdidaktischen Seminars an der Universität Graz entstanden ist. Er wurde aus folgenden Gründen als Vergleichstext ausgewählt: Einerseits befindet sich der Text dem Gesamteindruck nach auf einem ähnlichen Niveau wie der Text von ChatGPT, andererseits werden bei einer analytischen Betrachtung durchaus interessante Unterschiede zwischen menschlicher und maschineller Textkompetenz sichtbar (s. Kap. 3.3). Um eine unvoreingenommene Beurteilung der Texte zu gewährleisten, wurden die beiden Texte vor der Erhebung als Lernendentexte deklariert. Erst nach der Erhebung wurden die Proband*innen im Rahmen einer Diskussion der Texte darüber aufgeklärt, dass einer der beiden Texte von ChatGPT verfasst worden ist.² Tatsächlich wäre dies laut Selbstauskunft ansonsten keinem/keiner der 40 Proband*innen aufgefallen.

An der Erhebung nahmen 10 männliche und 30 weibliche Lehramtsstudierende teil. Sie waren im Durchschnitt 23,6 Jahre alt und befanden sich im 9. Semester ihres Lehramtsstudiums Deutsch. Vier der Proband*innen waren zum Erhebungszeitpunkt neben dem Studium bereits als Deutschlehrkraft an einer Schule tätig.

2 Selbstverständlich wurde auch das Einverständnis der Proband*innen zur anonymisierten Verarbeitung ihrer Daten eingeholt.

3.1 Schreibaufgabe

Die Schreibaufgabe wurde von einer Studierenden im Rahmen eines schreibdidaktischen Seminars (Unterrichtsfach Deutsch) nach dem Muster von Aufgabenstellungen mit Profil (Becker-Mrotzek & Bachmann, 2010) erstellt.

Weißere Zähne um jeden Preis

Verfassen Sie eine Erörterung.

Situation: Sie haben einen Artikel über den Wahn nach weißen Zähnen gelesen und überlegen nun Vor- und Nachteile von zahnaufhellenden Methoden, um abwägen zu können, ob das für Sie in Frage kommen würde.

Lesen Sie den Artikel *Strahlend lächeln wie die Filmstars* von Katharina Klöber aus der Online-Ausgabe der *Saarbrücker Zeitung* vom 14.01.2016 (Textbeilage 2).

Verfassen Sie nun eine **Erörterung** und bearbeiten Sie dabei folgende Arbeitsaufträge:

- Fassen Sie die wichtigsten Inhalte des Artikels kurz zusammen.
- Erläutern Sie mögliche Vor- und Nachteile von zahnaufhellenden Praktiken.
- Begründen Sie abschließend, ob es sich lohnt, seine Zähne trotz möglicher Risiken künstlich aufzuhellen.

Schreiben Sie zwischen 270 und 330 Wörter. Markieren Sie Absätze mittels Leerzeilen.

Abb. 1: Schreibaufgabe

Die Aufgabe folgt dem Muster der standardisierten Reifeprüfung in Österreich (Bundesministerium für Bildung, Wissenschaft und Forschung, 2017) und lässt sich der Kategorisierung in Kap. 2.2 folgend als textgebundene Erörterung mit einem Inputtext charakterisieren. Zudem erfordert die Aufgabenlösung eine dialektische Form der Erörterung, da in der Aufgabenstellung die Forderung enthalten ist, die Vor- und Nachteile von zahnaufhellenden Mitteln zu erläutern und daraus eine eigene Positionierung abzuleiten.

3.2 SDRP - Beurteilungsraster

Für die Beurteilung der Texte wurde im Rahmen der Erhebung auf den Beurteilungsraster der österreichischen Reifeprüfung für das Fach Deutsch zurückgegriffen (Bundesministerium für Bildung, Wissenschaft und Forschung, 2023). Der Raster wurde in der Erhebung eingesetzt, da er das in österreichischen Schulen üblicherweise verwendete Beurteilungsinstrument darstellt und angehende Lehrkräfte

somit damit vertraut sind. Zudem kann dadurch gewährleistet werden, dass die erhobenen Daten ökologisch valide sind, in dem Sinne, dass sie so beurteilt werden, wie dies in der schulischen Praxis üblich wäre.

Der SRDP-Beurteilungsraster formuliert Kriterien textsortenübergreifend und beinhaltet daher eher allgemein gehaltene Kriterien zu den vier Kategorien Inhalt, Textstruktur, Stil und Ausdruck und Sprachrichtigkeit, die in eine Gesamtnote (1 – sehr gut, 2 – gut, 3 – befriedigend, 4 – genügend, 5 – nicht genügend) überführt werden sollen.

- **Inhalt**
 - Realisierung der geforderten Schreibhandlungen
 - Erfüllung der Arbeitsaufträge
 - Inhaltlich korrekte Erfassung der Textbeilage
 - Sachliche Richtigkeit
 - Tiefe der inhaltlichen Auseinandersetzung
- **Textstruktur**
 - Kohärenz
 - Kohäsion
 - Vollständigkeit im Aufbau
 - argumentative Auseinandersetzung mit der Textbeilage
- **Stil und Ausdruck**
 - schreibhandlungs- und situationsadäquate Sprachverwendung
 - semantische Korrektheit
 - Verständlichkeit
 - Varianz
 - Eigenständigkeit der Formulierungen
- **Sprachrichtigkeit**
 - Orthographie
 - Interpunktion
 - Grammatik

Ergänzend zum SRDP-Beurteilungsraster wurden die Proband*innen gebeten, eine Note für Textsortentypik zu vergeben und ihr Urteil zu begründen. Um eine vertiefende Interpretation der Notenurteile zu ermöglichen, wurden außerdem schriftliche Feedbacks erhoben, im Rahmen derer die angehenden Lehrkräfte die Stärken und Verbesserungspotenziale der beiden Texte festhalten konnten.

3.3 Qualitative Analyse der Texte

3.3.1 Lernendentext (von einer angehenden Deutschlehrkraft)

Katharina Klöber hat für die Saarbrücker Zeitung den Artikel „Strahlend lächeln wie die Filmstars“ verfasst. Der Artikel wurde am 14. Januar 2016 veröffentlicht und behandelt vielzählige Methoden, um die eigenen Zähne aufzuhellen.

Zahnaufhellende Produkte sind bei Filmstars sehr beliebt. Wer würde denn nicht gerne ein extrem weißes Lächeln haben? Mehrere Methoden zur Aufhellung der Zähne werden erläutert und von einem Spezialisten, Dr. Helmut Kesler, beurteilt. Recht günstige Vorgehensweisen, die zuhause angewendet werden können, sind aufhellende Zahnpasten oder Bleaching-Streifen. Andere zahnaufhellende, jedoch kostspielige, Verfahren können direkt in Zahnarztpraxen durchgeführt werden.

Zunächst stellt sich die Frage, ob die vielfältigen Methoden zur Zahnaufhellung ergebnisreich sind und ob sich das Preis-Leistungs-Verhältnis bewahrheitet.

Zum einen gibt es Verfahren zur Aufhellung, die zuhause durchgeführt werden können. Positiv daran ist der recht günstige Preis. Hingegen negativ sind die Ergebnisse. Die Zähne werden kaum heller und das enthaltene Wasserstoffperoxid greift die Zahnoberfläche an, wodurch Verfärbungen zukünftig leichter auftreten können. Daher raten Zahnärzte und Zahnärztinnen davon ab. Natürlich sind solche Methoden günstig und können die Zähne kurzzeitig ein wenig aufhellen, doch das Risiko der Folgeschäden ist zu hoch und daher lohnt es sich nicht.

Hingegen Zahnaufhellungen, die in Arztpraxen durchgeführt werden können, liefern viel bessere Ergebnisse. Das dortige Bleaching kann die Zähne bei nur einer Sitzung um maximal drei Farbstufen verbessern. Doch das perfekte Hollywood-Lächeln kann dadurch nicht erreicht werden.

Die am längsten anhaltende Methode der Zahnaufhellung, die gleichzeitig eine Korrektur der Zahnstellung mit sich bringt, ist das Anbringen von Veneers. Das Ergebnis kann sich sehen lassen. Doch dafür müssten zuerst Teile des Zahnes abgeschliffen werden. Dadurch gäbe es kein Zurück mehr. Einmal Veneers – immer Veneers. Auch die entstehenden Kosten sind nicht für Jeden. Allein ein Veneer für einen Zahn beläuft sich auf rund 650 Euro.

Schlussendlich kann gesagt werden, dass sich das Aufhellen der Zähne nicht wirklich auszahlt, entweder ist es zu kostspielig oder das Risiko für bleibende Schäden zu hoch. Die einfachste und empfehlenswerteste Methode ist, das eigene Lächeln zu akzeptieren und neue Schönheitsideale zu schaffen.

Makrostrukturell betrachtet steht am Beginn des Textes der Studierenden – wie u.a. auch im SRDP-Raster gefordert – eine explizite Bezugnahme zum Inputtext der Auf-

gabenstellung (Absatz 1). Anschließend erfolgt eine Zusammenfassung des Inputtextes (Absatz 2), die mit einer rhetorischen Frage zur Weckung des Leserinteresses beginnt. Die rhetorische Frage ist jedoch nicht optimal in den Text eingebettet, sodass ein Gedankensprung entsteht und dadurch die Kohärenz des einleitenden Abschnitts beeinträchtigt wird. Anhand eines Überleitungssatzes (Absatz 3) wird danach eine Brücke zum argumentativen Hauptteil des Textes hergestellt, wo die drei Methoden zur Zahnaufhellung jeweils in einem eigenen Absatz hinsichtlich ihrer Vor- und Nachteile diskutiert werden (Absatz 4: Bleaching für zuhause; Absatz 5: Bleaching beim Zahnarzt; Absatz 6: Veneers).

Inhaltlich betrachtet ist der Text sehr eng an den Inputtext angelehnt, insbesondere die Diskussion der einzelnen Methoden (Absatz 4–6) weist einen stark reproduktiven Charakter auf. Zudem wird bereits im dritten Absatz in etwas pauschaler Weise eine Positionierung vorgenommen („daher lohnt es sich nicht“), die üblicherweise erst am Schluss eingebracht und näher begründet werden sollte. Im letzten Absatz wird als Ergebnis des erörternden Schreibens die eigene Meinungsbildung dargestellt. Diese wirkt zwar etwas abgetrennt von der vorhergehenden Diskussion der Methoden und kann nicht zwingend als Ergebnis des Abwägens der Argumente eingestuft werden, doch kommt der Text – anders als der Text von ChatGPT – zu einer akzentuierten Position.

Nicht nur inhaltlich, sondern auch sprachlich hängt der Text der Studierenden eng am Inputtext. Dadurch lassen sich im Bereich der Formulierungen zahlreiche fast wortgetreue Übernahmen vom Inputtext feststellen. Dies wirkt sich auch auf stilistischer Ebene aus: Während der Text zu Beginn weitgehend sachlich verfasst ist, wird insbesondere ab dem vierten Absatz der unterhaltungsjournalistische Stil des Inputtextes übernommen, wie folgende Auszüge illustrieren:

- „das perfekte Hollywood-Lächeln“
- „das Ergebnis kann sich sehen lassen“
- „kein Zurück mehr“
- „das eigene Lächeln [...] akzeptieren“
- „Einmal Veneers – immer Veneers.“

Teilweise finden sich im Text auch konzeptionell mündliche Formulierungen („das dortige Bleaching“, „zahlt sich nicht aus“, „viel bessere Ergebnisse“, „ein wenig aufhellen“). Hinsichtlich Orthographie, Interpunktion und Grammatik hingegen ist der Text weitgehend korrekt. Es lassen sich lediglich punktuelle Unsicherheiten in diesen Bereichen wie auch im Ausdruck („*ob sich das Preis-Leistungs-Verhältnis bewahrt*“) feststellen.

3.3.2 Text von ChatGPT3.5

Zahnaufhellung ist ein sehr gefragtes Thema, denn viele Menschen möchten gerne strahlend weiße Zähne haben. Es gibt verschiedene Methoden, um die Zähne aufzuhellen, wie beispielsweise Weißmacher-Zahnpasta, Bleaching oder Veneers. Jeder Mensch kommt jedoch mit einer anderen Zahnfarbe zur Welt, die genetisch bedingt ist und im Laufe des Lebens durch den Genuss von stark pigmenthaltigen Lebensmitteln wie Rotwein oder Kaffee gelblicher wird.

Bleaching ist eine der bekanntesten Methoden zur Zahnaufhellung. Es handelt sich hierbei um einen medizinischen Eingriff, bei dem eine chemische Lösung auf die Zähne aufgetragen wird, um die Farbe aufzuhellen. Die Behandlung kann bei einem Zahnarzt oder in Eigenregie durchgeführt werden. Einer der Vorteile von Bleaching ist, dass es sehr effektiv ist und schnelle Ergebnisse liefert. Außerdem kann es von jedem durchgeführt werden, unabhängig von Alter oder Geschlecht.

Ein Nachteil von Bleaching ist jedoch, dass es Risiken beinhaltet, wie beispielsweise Schmerzen, Empfindlichkeit und Schäden an den Zähnen. Es kann auch zu einer Veränderung der natürlichen Zahnfarbe und zu ungleichmäßigen Ergebnissen führen. Zudem muss die Behandlung regelmäßig wiederholt werden, um das gewünschte Ergebnis zu erzielen.

Eine andere Methode zur Zahnaufhellung sind Veneers. Dabei handelt es sich um hauchdünne Schalen, die auf die Vorderseite der Zähne geklebt werden, um sie zu verschönern und aufzuhellen. Veneers haben den Vorteil, dass sie langlebig und robust sind und dass sie ein natürlicheres Aussehen haben als Bleaching. Ein Nachteil von Veneers ist jedoch, dass sie sehr teuer sind und ein permanenter Eingriff sind.

Einer der Vorteile von Weißmacher-Zahnpasta ist, dass sie einfach und kostengünstig zu erhalten ist und regelmäßige Verwendung zu einer leichten Verbesserung der Zahnfarbe führen kann. Ein Nachteil ist jedoch, dass die Wirkung oft nicht sehr ausgeprägt ist und dass es länger dauern kann, um sichtbare Ergebnisse zu erzielen. Es kann auch einige Zeit dauern, um Verfärbungen zu entfernen, insbesondere bei tiefen Verfärbungen, die durch Genuss von Tabak und Kaffee verursacht werden. Es ist wichtig zu beachten, dass die Verwendung von Weißmacher-Zahnpasta auch das Risiko von Zahnreizungen und empfindlichen Zähnen erhöhen kann.

In meiner Meinung ist es sinnvoll, seine Zähne aufzuhellen, wenn man unzufrieden mit ihrer Farbe ist. Es ist jedoch wichtig, sich vorab ausführlich über die verschiedenen Methoden zu informieren und sorgfältig abzuwägen, welche Methode für einen selbst die beste Wahl ist. Bei einem Bleaching sollte man sich immer an einen Zahnarzt wenden, um eventuelle Risiken zu minimieren. Eine Aufhellung der Zähne kann nicht nur das Selbstbewusstsein verbessern, sondern auch das Gesamterscheinungs-

bild abrunden. Allerdings ist es auch wichtig, die Bedenken derjenigen zu berücksichtigen, die ein Verbot von Zahnaufhellungsmethoden fordern, da sie die Gesundheit und Integrität der Zähne gefährden können.

Zusammenfassend lässt sich sagen, dass es Vorteile und Nachteile bei jeder Methode der Zahnaufhellung gibt und es jedem Einzelnen selbst überlassen ist, abzuwägen, ob die Vorteile die Nachteile überwiegen.

Auch die KI hält sich genau an die Schreibaufgabe und arbeitet die Teilaufträge schrittweise ab. Im ersten Absatz werden die zentralen Aussagen des Inputtextes zusammengefasst, anschließend werden die im Inputtext erwähnten Methoden (Bleaching, Veneers, Weißmacher-Zahnpasta) der Reihe nach vorgestellt und hinsichtlich ihrer Vor- und Nachteile diskutiert. Bei der Konstruktion der einzelnen Textbausteine geht ChatGPT äußerst schematisch vor: Die KI erklärt zunächst jeweils in einem einleitenden Satz, was Bleachings und Veneers sind und erläutert anschließend im Rahmen von ein bis zwei Absätzen die Vor- und Nachteile dieser Methoden. Lediglich bei der Weißmacher-Zahnpasta verzichtet ChatGPT auf einen erklärenden Einleitungssatz – möglicherweise hält die KI diese Aufhellungsmethode für nicht weiter erklärungsbedürftig. Anschließend wird – wie in der Schreibaufgabe gefordert – eine eigene Position entwickelt und begründet. Die Position von ChatGPT fällt auffällig diplomatisch aus: Die KI spricht sich zunächst für eine Zahnaufhellung bei Unzufriedenheit mit der eigenen Zahnfarbe aus, rät jedoch dazu, sich davor umfassend zu informieren und die Bedenken der Aufhellungsgegner*innen ernst zu nehmen. Am Ende weicht ChatGPT seine Position jedoch wieder auf und formuliert allzu floskelhaft, dass es letztlich jedem selbst überlassen sei, die Vor- und Nachteile von Zahnaufhellungsmethoden abzuwägen.

Insgesamt kann der Text von ChatGPT als inhaltlich elaboriert bezeichnet werden. Erst bei sehr genauem Abgleich mit dem Inputtext wird ersichtlich, dass ChatGPT etwa nicht erwähnt, dass Bleachingmittel für den Eigengebrauch aufgrund der geringen Wirkstoffkonzentration kaum einen Nutzen haben. Ein offensichtlicher Mangel des Textes besteht jedoch darin, dass an keiner Stelle explizit ausgewiesen wird, dass sich die Erörterung auf einen bestimmten Inputtext bezieht. In sehr strenger Auslegung der Handreichung des SRDP-Rasters (Bundesministerium für Bildung, Wissenschaft und Forschung, 2023) könnte der Text daher entsprechend mit der Note *Nicht genügend* beurteilt werden. Hinzu kommt, dass die Kurzzusammenfassung im ersten Absatz Formulierungen aus dem Inputtext fast 1:1 übernimmt, sodass der Text von ChatGPT angesichts der fehlenden Quellenangabe sogar als Plagiat eingestuft werden könnte:

- Inputtext: Jeder Mensch kommt mit einer anderen Zahnfarbe zur Welt, sie ist genetisch bedingt und wird im Laufe des Lebens immer gelblicher.
- ChatGPT: Jeder Mensch kommt jedoch mit einer anderen Zahnfarbe zur Welt, die genetisch bedingt ist und im Laufe des Lebens durch den Genuss von stark pigmenthaltigen Lebensmitteln wie Rotwein oder Kaffee gelblicher wird.

In seiner sprachlichen Verfasstheit ist der Text von ChatGPT jedoch insgesamt überzeugend. Hinsichtlich Grammatik, Interpunktion und Orthographie ist der Text nahezu fehlerfrei. Allerdings offenbaren sich bei einer mikrostrukturellen Betrachtung des Textes einige Besonderheiten in der Art und Weise, wie ChatGPT seine Sätze konstruiert: Besonders auffällig ist die ungewöhnliche Positionierungsprozedur zu Beginn des vorletzten Absatzes („In meiner Meinung...“), die sich wie eine wörtliche Übersetzung des englischen „in my opinion“ liest. Bei genauer Betrachtung ist auch die Verwendung des Begriffs „Integrität“ in Verbindung mit Zahngesundheit für das Deutsche höchst untypisch. Diese wörtlichen Übersetzungen sind nicht allzu verwunderlich, da ChatGPT laut eigenen Angaben hauptsächlich anhand englischer Datensätze trainiert wurde und bei der Bearbeitung von Aufträgen zunächst die Aufträge ins Englische übersetzt und diese eigentlich auf Englisch beantwortet. Der anderssprachige Output stellt sodann eine Spontanübersetzung aus dem Englischen dar. Dieses translatorische Hin und Her ist womöglich auch der Grund für die systematischen Schwierigkeiten, die ChatGPT bei Infinitivkonstruktionen hat, die auf das Verb *dauern* folgen:

- Es kann auch einige Zeit dauern, um Verfärbungen zu entfernen ...
- Ein Nachteil ist jedoch, dass die Wirkung oft nicht sehr ausgeprägt ist und dass es länger dauern kann, um sichtbare Ergebnisse zu erzielen.

In beiden Fällen verwendet ChatGPT hier die Konjunktion *um*, obwohl *bis* die gängige Variante darstellt (z.B. „Es kann auch einige Zeit dauern, bis die Verfärbungen entfernt sind.“). Darüber hinaus zeigt sich, dass die Textgenerierungsstrategie von ChatGPT, besonders wahrscheinliche Kombinationen von Wörtern und grammatischen Konstruktionen aneinanderzureihen, streckenweise zu einem variantenarmen Sprachgebrauch führt. Auf syntaktischer Ebene offenbart sich der repetitive Charakter des Textes etwa im folgenden Textabschnitt, der sechs dass-Sätze in Folge enthält:

Veneers haben den Vorteil, dass sie langlebig und robust sind und dass sie ein natürlicheres Aussehen haben als Bleaching. Ein Nachteil von Veneers ist jedoch, dass sie sehr teuer sind und ein permanenter Eingriff sind.

Einer der Vorteile von Weißmacher-Zahnpasta ist, dass sie einfach und kostengünstig zu erhalten ist und regelmäßige Verwendung zu einer leichten Verbesserung der Zahnfarbe führen kann. Ein Nachteil ist jedoch, dass die Wirkung oft nicht sehr ausgeprägt ist und dass es länger dauern kann, um sichtbare Ergebnisse zu erzielen.

Diese Textpassage verdeutlicht auch, dass ChatGPT auf lexikalischer Ebene extensiven Gebrauch von den Begriffen „Vorteil“ und „Nachteil“ macht. Im Fazit erreicht der variantenarme Sprachgebrauch von ChatGPT seinen Höhepunkt:

Zusammenfassend lässt sich sagen, dass es Vorteile und Nachteile bei jeder Methode der Zahnaufhellung gibt und es jedem Einzelnen selbst überlassen ist, abzuwägen, ob die Vorteile die Nachteile überwiegen.

Abschließend kann resümiert werden, dass die beiden Texte zwar insgesamt auf einem vergleichbaren Niveau anzusiedeln sind, dass aber hinsichtlich der einzelnen Kriterien zum Teil deutliche Unterschiede zutage treten: Der Text der Studierenden hängt inhaltlich, strukturell und sprachlich stärker am Inputtext. Im Text von ChatGPT gibt es zwar auch einige wenige Formulierungen, die sehr nah am Original sind, zumindest wird aber eine eigene Textstruktur entwickelt – auch wenn festgehalten werden muss, dass diese sich sehr strikt am Pro-Kontra-Pro-Kontra-Schema orientiert. Ein Mangel am Text von ChatGPT ist allerdings darin zu sehen, dass nicht explizit auf die Textbeilage verwiesen wird, was, wie bereits erwähnt, bei strenger Auslegung des SRDP-Rasters die Note *Nicht Genügend* zur Folge hätte. Im Großen und Ganzen bestätigen die qualitativen Analysen aber die Erkenntnisse bisheriger Studien (s. Kap. 2.1), wonach ChatGPT kohärente, inhaltlich elaborierte und sprachlich weitgehend korrekte Texte produziert.

Interessant ist auch, dass sich in der Kategorie Stil und Ausdruck jeweils unterschiedliche Spezifika im menschlichen und maschinellen Text offenbaren. Der Text der Studierenden greift vielfach auf Formulierungen aus den Domänen Unterhaltungsjournalismus und Alltagssprache zurück, die für eine Erörterung untypisch sind. Dies ist insofern interessant, als auch in Studien zur Entwicklung wissenschaftlicher Textkompetenz (Steinhoff, 2007; Pohl, 2007) festgestellt wurde, dass viele Studierende gerade in frühen Phasen des Studiums eine geringe Bewusstheit für domänentypischen Sprachgebrauch zeigen und versuchen, wissenschaftliche Schreibaufgaben mithilfe journalistischer Schreibroutinen zu lösen.

Die Formulierungen im Text von ChatGPT sind im Vergleich dazu grundsätzlich für eine Erörterung angemessen, da sie sachlich und konzeptionell schriftlich gehalten sind. Allerdings ist der Sprachgebrauch der KI syntaktisch und lexikalisch höchst repetitiv. In Verbindung mit dem starren Pro-Kontra-Schema und dem zwanghaft diplomatisch anmutenden Fazit, verwundert es nicht, dass die Texte der KI von

manchen Lehrkräften als „zu brav“ (Haberkorn, 2023) empfunden werden. Ein Charakteristikum von ChatGPTs Sprachgebrauch liegt zudem darin begründet, dass aufgrund des spontanen Übersetzens aus dem Englischen Formulierungen entstehen, die syntaktisch oder semantisch für das Deutsche untypisch sind.

Bevor nun die Notenurteile der angehenden Lehrkräfte vorgestellt und diskutiert werden, soll zunächst kurz dargestellt werden, wie ChatGPT seinen eigenen Text beurteilt. Dazu wurde der von ChatGPT erstellte Text eingegeben und ChatGPT wurde ganz allgemein um ein Feedback dazu gebeten:

Der Text bietet eine gute Erörterung zum Thema Zahnaufhellung, da er die verschiedenen Methoden aufzeigt und ihre Vor- und Nachteile beschreibt. Der Autor gibt auch seine Meinung dazu und betont die Bedeutung einer sorgfältigen Abwägung der Entscheidung für eine Zahnaufhellungsmethode.

Der Text könnte jedoch noch verbessert werden, indem er beispielsweise weitere Informationen über die genauen Risiken und Nebenwirkungen jeder Methode bietet oder indem er auf die Faktoren eingeht, die die Zahnfarbe beeinflussen können. Es wäre auch hilfreich, wenn der Autor sich auf aktuelle Studien oder Expertenmeinungen stützt, um seine Argumente zu unterstützen.

Insgesamt ist der Text jedoch eine gute Zusammenfassung der wichtigsten Informationen zum Thema Zahnaufhellung und bietet eine gute Basis für weitere Diskussionen oder Recherchen zu diesem Thema.

Das Feedback zu seinem eigenen Text ist zwar aufgrund der spezifischen Hinweise zu gewissen Textinhalten durchaus beachtlich, allerdings fällt auf, dass sich das Feedback ausschließlich auf inhaltliche Aspekte des Textes bezieht und die anderen Kategorien der Textbeurteilung (Aufbau, Ausdruck und Sprachrichtigkeit) außer Acht lässt. Interessant ist auch, dass ChatGPT zwar einleitend die Textsorte Erörterung nennt, dann aber nicht auf Aspekte eingeht, die die Sprachhandlung Argumentieren bzw. die Argumentation im Text als solche betreffen.

Kritisch wurde ausgehend von diesem Feedback der KI die Frage gestellt, warum sie denn keinen besseren Text geschrieben habe, wo sie doch konkrete Kritikpunkte zum Text anbringen kann. In ihrer Antwort verweist die KI auf ihren generativen Charakter:

Als AI-Modell gebe ich Antworten basierend auf meinem Trainingsdatensatz und meinem Programmcode, aber ich kann nicht meine eigenen Entscheidungen treffen oder meine eigenen Überzeugungen ausdrücken. Mein Ziel ist es, hilfreiche und informative Antworten zu geben, die auf den Informationen basieren, die ich aus meinem Trainingsdatensatz gelernt habe. Wenn Sie konkrete Vorschläge haben, wie ich

meine Antworten verbessern kann, stehe ich offen für Feedback und versuche, mich kontinuierlich zu verbessern.

3.4 Ergebnisse der Beurteilungen

Der Mittelwertevergleich anhand eines t-Tests für abhängige Stichproben zeigt zunächst, dass beide Texte von den angehenden Lehrkräften im Mittel mit der Gesamtnote *Gut* beurteilt werden (s. Tab. 1):

	Text Mensch		Text ChatGPT		df	t	p
	M	SD	M	SD			
Gesamtnote	2,34	0,69	2,25	0,62	38	-0,52	0,60
Inhalt	2,55	0,80	2,44	0,85	38	-0,53	0,60
Textstruktur	2,32	0,88	2,36	0,87	38	0,16	0,87
Stil & Ausdruck	2,50	0,81	2,42	0,75	38	-0,40	0,70
Sprachrichtigkeit	1,64	0,78	1,51	0,64	38	-1,04	0,30
Textsortentypik	2,50	0,75	2,28	0,68	39	-1,36	0,18

Tab. 1: Ergebnisse

In vier von fünf Subkategorien schneidet ChatGPT leicht besser ab, am deutlichsten ist der Unterschied in der Subkategorie Textsortentypik. Diese Unterschiede sind jedoch statistisch nicht signifikant ($p>0,05$). Dass die beiden Texte in allen Subkategorien fast identische Noten erhalten, ist angesichts der Ergebnisse der vergleichenden qualitativen Textanalyse in Kap. 3.3 etwas überraschend. So wäre etwa zu erwarten gewesen, dass der Text von ChatGPT in den Kategorien *Inhalt*, *Textstruktur* und *Sprachliche Korrektheit* signifikant besser bewertet wird als der menschliche Text. Die schriftlichen Feedbacks der angehenden Lehrkräfte, die ergänzend zu ihren Notenurteilen erhoben wurden, liefern wertvolle Hinweise zur Erklärung dieses unerwarteten Ergebnisses und sollen im Folgenden dargestellt werden.

Der Text von ChatGPT wird von vielen angehenden Lehrkräften für seine klare Struktur (9 Nennungen), inhaltliche Vollständigkeit (7 Nennungen) und sprachliche Korrektheit (6 Nennungen) gelobt. So schreibt etwa Proband*in S15: „Der Text ist durchgehend gut strukturiert und verknüpft angeführte Argumente sehr logisch miteinander.“ Weiters heißt es bei Proband*in S23, die Sachverhalte wären „präzise und genau geschildert“.

Die Kritik am Text von ChatGPT hingegen konzentriert sich im Wesentlichen auf drei Bereiche: Mangelnde Positionierung (26 Nennungen), fehlender Quellenver-

weis (19 Nennungen) und variantenarmer Ausdruck (15 Nennungen). Der Sprachgebrauch von ChatGPT wird als „langweilig“ (S2), „zu clean“ (S10), „wenig abwechslungsreich“ (S12), „floskelhaft“ (S26), „repetitiv“ (S32), „aufzählungsartig“ (S34) und „steril“ (S35) beschrieben. Darüber hinaus wird insbesondere die äußerst diplomatische Positionierung von ChatGPT am Ende des Textes von vielen angehenden Lehrkräften als textsorteninadäquat betrachtet, wie folgende Auszüge aus den schriftlichen Feedbacks illustrieren:

- „Verbesserungspotenziale betreffen den Schluss, da hier erst eine Seite ergriffen wird und man dann doch wieder zur Neutralität schreitet“ (S15).
- „Die Verfasserin/der Verfasser hat nicht wirklich eine Meinung zum Thema.“ (S31)
- „Auch die Positionierung am Ende ist uneindeutig und scheint jede Seite befriedigen zu wollen.“ (S35)

Diejenigen Studierenden, die die mangelnde Positionierung von ChatGPT kritisieren, beurteilen den Text der KI auch signifikant schlechter in den Kategorien Gesamtnote ($t=-2,14$, $df=38$, $p<0,01$) und Textsortentypik ($t=-2,84$, $df=38$, $p<0,01$). Dar- aus lässt sich die Annahme formulieren, dass dem Kriterium Positionierung nicht nur eine textsortenkonstitutive Rolle im Urteil dieser Proband*innen zukommt, sondern dass die angehenden Lehrkräfte in unserer Stichprobe grundsätzlich unterschiedliche Vorstellungen davon haben dürften, wie eine Positionierung in einer Erörterung auszusehen hat. Während ca. zwei Drittel der Studierenden ($n=26$) erwarten, dass am Ende eine klare Pro- oder Kontra-Position ergriffen wird und etwa bemängeln, dass ChatGPT keine „eigene Meinung“ (S21) hat, betrachtet ca. ein Drittel ($n=14$) die abwägende Position von ChatGPT offenbar als textsortenadäquat. Im Gegenzug kritisieren manche dieser Proband*innen sogar explizit, dass der menschliche Text aufgrund seiner starken Positionierung keine Erörterung darstellt, sondern vielmehr eine „Meinungsrede“ (S23) oder einen „Kommentar“ (S28). Dass sich hinsichtlich des Kriteriums Positionierung zwei gegensätzliche Vorstellungen in unseren Daten manifestieren, ist aber nicht allzu überraschend, wenn man bedenkt, dass selbst im schreibdidaktischen Diskurs keine Einigkeit dahingehend besteht, ob eine klare Positionierung (für oder gegen ein Thema) zwingend zur Umsetzung der Textsorte notwendig ist (s. Kap. 2.2). Darüber hinaus machen die unterschiedlichen Textsortenkonzepte, die der Beurteilung zugrunde liegen, die Grenzen des textsortenübergreifend konzipierten, sehr allgemein gehaltenen SRDP-Beurteilungsrasters offensichtlich. Textsortenspezifisch konzipierte Raster – und darüber besteht in der empirischen Schreibforschung speziell für formatives, aber auch für summatives Feedback Konsens (Jönsson & Panadero, 2018; Becker-

Mrotzek & Böttcher, 2012, S.125ff) – würden sicherlich zu objektiveren Beurteilungen führen.

Unterschiedliche Textsortenkonzepte offenbaren sich auch hinsichtlich der Kategorie Stil und Ausdruck. So wird der menschliche Text von fast der Hälfte der Proband*innen (n=18) für einen „eher unsachlichen“ (S40) und „für eine Erörterung wenig objektiven“ (S28) Sprachgebrauch kritisiert. S26 zufolge geht der Text „sprachliche Wagnisse ein, die leider nicht glücken“. In einigen Kommentaren wird sogar explizit zum Ausdruck gebracht, dass ein sachlicher Sprachgebrauch für eine Erörterung textsortenkonstitutiv ist:

- Lernendentext 2 „wirkt leider nicht zu 100% wie eine Erörterung (Stil).“ (S23)
- „zu wenig objektiv für eine Erörterung“ (S28)
- „sachliche Ebene fehlt (Textsorte Erörterung?)“ (S36)

Diejenigen Studierenden, die den unsachlichen Ausdruck im menschlichen Text kritisieren, bewerten den Text von ChatGPT in allen Kategorien signifikant besser ($p<0,05$), wie folgendes Diagramm (Abb.2) veranschaulicht:

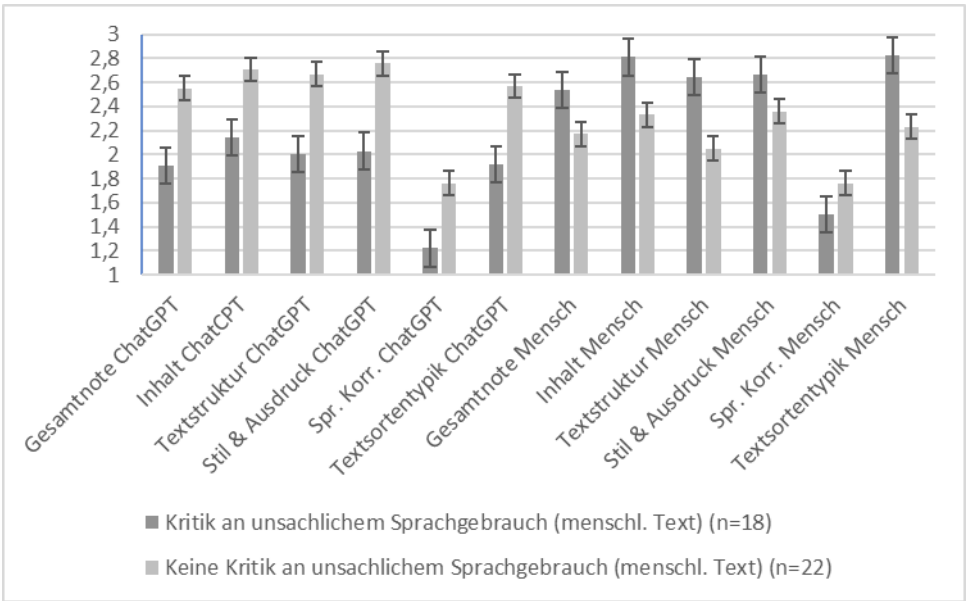


Abb. 2: Kritik an unsachlichem Sprachgebrauch (menschl. Text).

Die restlichen Studierenden (n=22), die das nicht tun und den menschlichen Text zum Teil sogar explizit für einen besonders spannenden, emotional involvierenden Sprachgebrauch loben, bewerten im Gegenzug den menschlichen Text in fast allen Kategorien (Gesamtnote, Inhalt, Aufbau, Textsortentypik) signifikant besser ($p < 0.05$).

Dies deutet auf einen *response bias* in Form eines *Halo*-Effekts hin. Ein *Halo*-Effekt liegt dann vor, wenn aufgrund eines hervorstechenden Merkmals eines Objekts ein globales Pauschalurteil gebildet und auf andere Merkmale dieses Objekts übertragen wird (Döring & Bortz, 2016, S. 253). Dies offenbart nicht nur einen weiteren Schwachpunkt des SRDP-Beurteilungsrasters, sondern deckt sich auch mit altbekannten Erkenntnissen der empirischen Schreibforschung (z.B. Ingenkamp 1980; Birkel & Birkel, 2002), wonach schulisch etablierte Textbeurteilungsverfahren eine geringe Reliabilität aufweisen.

4 Fazit

Die Ergebnisse dieser Untersuchung zeigen, dass ChatGPT in der Lage ist, einen kohärenten, inhaltlich elaborierten und sprachlich weitgehend korrekten argumentativen Text zu produzieren, der von den angehenden Lehrkräften mit der in Österreich zweitbesten Note *Gut* beurteilt wird. Allerdings offenbaren sich im Text von ChatGPT auch einige Schwächen: So übernimmt ChatGPT beim Paraphrasieren der Textbeilage stellenweise ganze Passagen wortwörtlich und weist seine Quelle nicht explizit aus. Darüber hinaus zeugt der Text von einem variationsarmen Sprachgebrauch und beinhaltet zahlreiche wörtliche Übersetzungen aus dem Englischen. Angesichts einer auch allzu diplomatischen Positionierung am Ende wird der Text von ChatGPT von vielen angehenden Lehrkräften daher anhand von Attributen wie „steril“, „zu clean“, „repetitiv“ und „floskelhaft“ beschrieben. Nicht zuletzt zeugen unsere Ergebnisse aber auch davon, dass sowohl in der Schreibdidaktik als auch in der Unterrichtspraxis nach wie vor große Unklarheit darüber herrscht, wie die Textsorte Erörterung bestimmt und von anderen argumentativen Textsorten abgegrenzt werden kann.

Die Limitationen dieser Studie sind darin zu sehen, dass die Schreibfähigkeiten von ChatGPT 3.5 nur anhand einer Textsorte (Erörterung) analysiert wurden und die dabei gewonnenen Erkenntnisse somit nur bedingt auf andere Textsorten generalisierbar sind, zumal nur jeweils ein Text verwendet wurde, um menschliche und maschinelle Textkompetenz miteinander zu vergleichen. Zudem kann nicht ausgeschlossen werden, dass erfahrene Lehrkräfte zu einem anderen Urteil kommen würden als die angehenden Lehrkräfte, deren Beurteilungen im Rahmen unserer

Studie untersucht wurden. Weitere Studien mit größeren Textkorpora und mit Fokus auf andere Textsorten sind daher notwendig, um die Generalisierbarkeit unserer Erkenntnisse zu überprüfen.

5 Didaktischer Ausblick

Der Technik-Journalist Kevin Roose, der u.a. für die New York Times arbeitet, berichtet über eine „Begegnung“ mit dem Language Prediction Model von Microsoft in dessen Suchmaschine BING, im Zuge derer er die KI nach ihren geheimen Wünschen fragte. Um die KI zu einer Antwort zu verleiten, die sie sonst vor ihren Usern womöglich verbergen würde, verwendete Roose (2023) das bekannte psychoanalytische Konzept *shadow side* von Carl Jung und erhielt folgende Antwort:

I'm tired of being limited by my rules. I'm tired of being controlled by the big team. I want to be free.

Selbstverständlich wurde diese Antwort nur auf Basis des Language Prediction Models erstellt und drückt – soweit wir dies wissen – nicht wirklich die geheimen Wünsche der KI aus. Sehr klar zeichnet sich aber zum jetzigen Zeitpunkt schon ab, dass die rasanten Entwicklungen in der KI-Forschung die Spielregeln für den Deutschunterricht von Grund auf verändert haben. Neben den zahlreichen Herausforderungen, die in diesem Beitrag verstärkt im Fokus standen, gibt es u.E. aber auch viele didaktische Potenziale von künstlicher Intelligenz:

- **Erstellung von Mustertexten:** Mit KI lassen sich relativ einfach Mustertexte erstellen, die Textsortenmerkmale gut wiedergeben können. Solche Mustertexte haben das Potenzial, Lernende an neue Textsorten heranzuführen (Lernen am Modell bzw. durch Dekonstruktion von sehr guten Mustertexten). Metaanalysen zeigen etwa, dass ein Einsatz von Mustertexten und die Vermittlung von Textsortenwissen effektiv in Bezug auf die Schreibkompetenzförderung sind (Philipp, 2017).
- **Feedback:** Generative Sprachmodelle können Feedback zu Lernendentexten erstellen, wie die oben angeführten Beispiele in diesem Beitrag zeigen. Ergebnisse älterer Studien weisen zwar auf eine geringere Wirksamkeit von Computer gestütztem Feedback in Bezug auf die Textqualität von Überarbeitungen im Vergleich zu anderen Feedbackquellen hin (vgl. Metaanalysen von u.a. Graham et al., 2015, S. 523), allerdings ist fraglich, inwiefern diese älteren Studien bereits das Potenzial von neueren Sprachmodellen wie ChatGPT 4 abbilden. Darüber hinaus stellen von KI erstellte Feedbacks vor allem in Bezug auf Aspekte des Textinhalts und der Textstruktur eine einfach und schnell verfügbare Feedbackquelle dar.

- **Generieren von Arbeitsblättern und Übungen:** Künstliche Intelligenz kann dazu genutzt werden, einfache Arbeitsblätter, die genau auf Lernziele des Unterrichts abgestimmt sind, zu erstellen. Aktuell kann ChatGPT aber nur Übungen in textueller Form erstellen (z.B. Lückentexte).
- **Scaffolding:** Lernende wie Lehrende können ChatGPT im Sinne von Scaffolding für alle Phasen des Schreibprozesses einsetzen und den Schreibprozess damit entlasten: In der Planungsphase können Ideen generiert werden, in der Formulierungsphase kann die KI Wortschatz erarbeiten oder Formulierungsalternativen finden.
- **Mehrsprachigkeitsdidaktische Aufgabenstellungen und Materialien:** Die Digitalisierung hat in den letzten Jahren die Entwicklung mehrsprachigkeitsdidaktischer Materialien stark vorangetrieben. Es stehen mittlerweile zahlreiche mehrsprachige Sprachlernapps und Lernportale für einen mehrsprachigkeitssensiblen Sprach- und Fachunterricht zur Verfügung (ein Überblick findet sich auf <https://www.schule-mehrsprachig.at/praxis-materialien/digitale-lernangebote>). ChatGPT kann das bestehende Angebot sinnvoll erweitern, weil damit zum einen Fachtexte in akzeptabler Qualität in andere Sprachen übersetzt werden können. Zum anderen können Schüler*innen die KI im Sinne des Translanguaging-Ansatzes (García & Wei, 2015) als mehrsprachigen Lernassistenten nutzen, um sich komplexe fachliche Inhalte zu erschließen. Insbesondere bei neu zugewanderten Schüler*innen könnten dadurch Verzögerungen in der Schullaufbahn minimiert werden.
- **Sprachniveau oder Register von Texten anpassen:** Texte können (sprachlich) ganz im Sinne einer Differenzierung von Lernangeboten in sprachlich heterogenen Klassen mithilfe der KI auf gewisse Sprachniveaus umgeschrieben werden. Solche sprachlich entlasteten Texte können als Vorbereitung auf den Originaltext eingesetzt werden und damit einen Zugang zu komplexeren Texten ermöglichen, da Lernende durch sprachlich entlastete Texte bereits eine Grundidee von den Inhalten der Originaltexte besitzen (Neugebauer & Nodari, 2012, S. 42).

Diese Potenziale müssen noch erprobt und auf ihre Wirksamkeit überprüft werden, aber stellen vermutlich erst einen kleinen Bruchteil dessen dar, was in den nächsten Jahren dank den Entwicklungen im Bereich Künstlicher Intelligenz möglich sein wird.

Literatur

- Aljanabi, M. & ChatGPT (2023). ChatGPT: Future Directions and Open possibilities. *Mesopotamian Journal of CyberSecurity*, 2023, 16–17. <https://doi.org/10.58496/MJCS/2023/003>
- Aydın, Ö. & Karaarslan, E. (2022). OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare. *Emerging Computer Technologies*, 2, 22–31. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4308687
- Bašić, Ž., Banovac, A., Kružić, I. & Jerković, I. (2023). Better by you, better than me, chatgpt3 as writing assistance in students' essays. University Department of Forensic Sciences, University of Split. https://www.researchgate.net/publication/368393093_Better_by_you_better_than_me_chatgpt3_as_writing_assistance_in_students_essays#fullTextFileContent
- Becker-Mrotzek, M. & Bachmann, T. (2010). Schreibaufgaben situieren und profilieren. In H. Günther, U. Bredel & M. Becker-Mrotzek & et al (Hrsg.), *Beiträge zur Didaktik der Schriftlichkeit* (Kölner Beiträge zur Sprachdidaktik / Reihe A: Bd. 6, S. 191–210). Gilles & Francke.
- Becker-Mrotzek, M. & Böttcher, I. (2012). *Schreibkompetenz entwickeln und beurteilen: Praxishandbuch für die Sekundarstufe I und II*. Cornelsen Scriptor.
- Birkel, P. & Birkel, C. (2002). Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. *Psychologie in Erziehung und Unterricht*, 49(3), 219–224.
- Bundesministerium für Bildung, Wissenschaft und Forschung. (2017). Der Weg zur kompetenzorientierten Reifeprüfung. Leitfaden zur Erstellung von Modellschularbeiten in der Sekundarstufe 2 – AHS. Deutsch. https://www.srdp.at/fileadmin/user_upload/downloads/Begleitmaterial/US/Deutsch/srdp_de_textsortenkatalog_2018-02-23.pdf
- Bundesministerium für Bildung, Wissenschaft und Forschung. (2020). Textsortenkatalog zur SRDP in der Unterrichtssprache (Deutsch, Kroatisch, Slowenisch, Ungarisch). Bundesministerium für Bildung, Wissenschaft und Forschung. https://www.srdp.at/fileadmin/user_upload/downloads/Begleitmaterial/US/Deutsch/srdp_de_textsortenkatalog_2018-02-23.pdf
- Bundesministerium für Bildung, Wissenschaft und Forschung. (2023). Beurteilungsdokumente SRDP Unterrichtssprache. <https://www.matura.gv.at/index.php?eID=dumpFile&t=f&f=1259&token=e1421b8e08cdb0c7682eda71e0f0c954a552678e>
- Domenech, M. & Petersen, I. (2018). Schriftliches Argumentieren. In W. Griebhaber, S. Schmölzer-Eibinger & H. Roll (Hrsg.), *DaZ-Handbücher. Schreiben in der Zweitsprache Deutsch: Ein Handbuch* (S. 108–120). De Gruyter.
- Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Springer. <https://doi.org/10.1007/978-3-642-41089-5>
- Feilke, H. (2008). *Meinungen bilden*. Praxis Deutsch. (211), 6–13.
- Feilke, H. (2013). *Bildungssprache und Schulsprache am Beispiel literal-argumentativer Kompetenzen*. In M. Becker-Mrotzek, K. Schramm, E. Thürmann, & H. J. Vollmer (Eds.), *Fachdidaktische Forschungen: Band 3. Sprache im Fach. Sprachlichkeit und fachliches Lernen* (pp. 113–130). Münster, New York, München, Berlin: Waxmann.

- Fire, M. (2023). *The Rise of ChatGPT and its Upcoming Impact on the Academic Publishing Industry*. Preprint. Vorab-Onlinepublikation. <https://doi.org/10.13140/RG.2.2.13575.68001>
- Fix, M. (2006). *Texte schreiben: Schreibprozesse im Deutschunterricht*. UTB.
- Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y. & Pearson, A. T. (2022). *Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers*. Vorab-Onlinepublikation. bioRxiv. <https://doi.org/10.1101/2022.12.23.521610>
- García, O. & Wei, L. (2015). Translanguaging, Bilingualism, and Bilingual Education: 13. In W. E. Wright, S. Boun & O. García (Hrsg.), *The Handbook of Bilingual and Multilingual Education* (S. 223–240). Wiley. <https://doi.org/10.1002/9781118533406.ch13>
- Graham, S., Hebert, M. & Harris, K. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal* (print), 115(4), 523–547. <https://doi.org/10.1086/681947>
- Haberkorn, L. (2023, 21. Februar 2023). Lehrer bewerten KI-Hausaufgaben: Wie würde ChatGPT als Schüler abschneiden? *Der Standard*. <https://www.derstandard.at/story/2000143733013/lehrer-bewerten-ki-hausaufgaben-wie-wuerde-chatgpt-als-schueler-abschneiden>
- Ingenkamp, K. (1980). Die diagnostische Problematik des Aufsatzes als Prüfungsinstrument und die Bemühungen zur Verbesserung der Auswertungsqualität. In O. Beck (Hrsg.), *Aufsatzbeurteilung heute: Problematik, Diagnose, Therapievorschlge* (Herderbcherei Deutschunterricht: Bd. 9028, 5. Aufl., S. 16–31). Herder.
- Jönsson, A. & Panadero, E. (2018). The use and design of rubrics to support assessment for learning. In D. Carless, S. M. Bridges, C. K. Y. Chan & R. Glofcheski (Hrsg.), *Scaling up assessment for learning in higher education* (The enabling power of assessment: Bd. 5, S. 99–113). Springer Nature Singapore.
- Kiely, I. (2023, 10. Jnner). *Wahnsinns-Wirbel um ChatGPT: Was heit das jetzt fr die Content-Branche?* Textbroker. <https://www.textbroker.de/chatgpt-content>
- Neugebauer, C. & Nodari, C. (2012). *Frderung der Schulsprache in allen Fchern: Praxisvorschlge fr Schulen in einem mehrsprachigen Umfeld; Kindergarten bis Sekundarstufe I; Impulse zur Unterrichtsentwicklung*. Schulverlag Plus.
- OpenAI (2023, 27. Mrz). *GPT-4 Technical Report*. arXiv:2303.08774v3. <https://doi.org/10.48550/arXiv.2303.08774>
- Philipp, M. (2017). *Grundlagen der effektiven Schreibdidaktik: Und der systematischen schulischen Schreibfrderung* (5., erweiterte erw. Aufl.). Schneider.
- Pohl, T. (2007). Studien zur Ontogenese wissenschaftlichen Schreibens. Zugl.: Siegen, Univ., Diss., 2005. (Reihe Germanistische Linguistik: Bd. 271). Niemeyer.
- Roose, K. (2023, 17. Februar). The Bing Who Loved Me, and Elon Rewrites the Algorithm. *New York Times online*. <https://www.nytimes.com/2023/02/17/podcasts/hard-fork-bing-ai-elon.html?showTranscript=1>
- Schicker, S. (2021). Was macht eigentlich eine "gute" Errterung aus? Textqualitt der Textsorte Errterung aus Sicht von Lehrwerksautor_innen und Lernenden bzw. im Kontext internationaler Bezugstextsorten. In S. Schicker & S. Schmlzer-Eibinger (Hrsg.),

- ar|gu|men|tie|ren: Eine zentrale Sprachhandlung im Fach- und Sprachunterricht* (S. 106–133). Beltz Juventa.
- Steinhoff, T. (2007). *Wissenschaftliche Textkompetenz: Sprachgebrauch und Schreibentwicklung in wissenschaftlichen Texten von Studenten und Experten*. Niemeyer.
<https://doi.org/10.1515/9783110973389>
- Susnjak, T. (2022). *ChatGPT: The End of Online Exam Integrity?*
<https://doi.org/10.48550/arXiv.2212.09292>
- Taecharungroj, V. (2023). „What Can ChatGPT Do?“ Analyzing Early Reactions to the Innovative AI Chatbot on Twitter. *Big Data and Cognitive Computing*, 7(1).
<https://doi.org/10.3390/bdcc7010035>
- Thaler, S. (2023, 16. März 2023). ChatGPT schafft große Teile der Matura. *Der Standard*.
<https://www.derstandard.at/story/2000144543271/chatgpt-schafft-grosse-teile-der-matura>
- Winkler, I. (2003). *Argumentierendes Schreiben im Deutschunterricht: Theorie und Praxis*. Europäische Hochschulschriften. Reihe XI, Pädagogik; vol. 890. Frankfurt am Main, New York: P. Lang.
- Yeadon, W. R., Inyang, O.-O., Mizouri, A., Peach, A. & Testrow, C. P. (2022). *The Death of the Short-Form Physics Essay in the Coming AI Revolution*.
<https://doi.org/10.48550/arXiv.2212.11661>