Wenn KI Feedback gibt

Eine neue Möglichkeit der Schreibförderung?

Christoph Peschak[1]

DOI: https://doi.org/10.53349/re-source.2025.i3.a1440

Zusammenfassung

Der differenzierte Lehrer*innenkommentar zu Schüler*innentexten ist für die Entwicklung der Schreibkompetenz unerlässlich. Abgesehen von der formalen Sprachrichtigkeit geht es vor allem darum, jene Kriterien in Texten zu identifizieren, die Auskunft über die Textqualität und die Schreibkompetenz der Schüler*innen geben (vgl. dazu Ossner, 2006; Becker-Mrotzek & Böttcher, 2015). Genau hier können large language models (kurz LLM) eine entscheidende Rolle für den Unterricht spielen. In diesem Beitrag werden Rückmeldungen von ChatGPT auf authentische Erzähltexte von Schüler*innen aus dem Gesamtkorpus einer Longitudinalstudie zur Schreibentwicklung (Augst et al., 2007) analysiert. Es soll gezeigt werden, in welchen Bereichen die Rückmeldungen bereits den Kriterien des kompetenzorientierten Schreibunterrichts und des lernförderlichen Feedbacks entsprechen (Becker-Mrotzek & Böttcher, 2014; Festman et al., 2023), aber auch, wo die Grenzen und Probleme dieser Form der Rückmeldung zu verorten sind.

Stichwörter: KI und Schule, Kompetenzorientierter Schreibunterricht, Textfeedback

1 Ausgangslage

Für die Entwicklung von Schreibkompetenz ist eine kritische Rückmeldung über die ästhetische und funktionale Dimension von Textprodukten unerlässlich (Becker-Mrotzek & Böttcher, 2015, S. 132). Eine mäeutische Haltung zum Textentwurf, im Gegensatz zu einer sanktionierenden normorientierten Korrekturpraxis, stellt hierbei einen bedeutenden Aspekt des kooperativen und dialogischen Rückmeldeprozesses durch die Lehrperson dar. Doch im Unterricht der Primarstufe stehen manchmal auch die Beurteilungsdimensionen Sprachrichtigkeit und Lesbarkeit der Handschrift im Vordergrund, die allerdings nicht viel Aussagekraft über die individuelle Schreibkompetenz in Hinblick auf das Verfassen von Texten haben. Denn ein aus der Sicht der Schreibdidaktik gelungener Text beinhaltet weit mehr als nur korrekte Grammatik und Orthografie. Entscheidend sind vor allem der inhaltliche Aufbau, die kreative

[1] Kirchliche Pädagogische Hochschule Wien/Niederösterreich, Mayerweckstraße 1, 1210 Wien. *E-Mail:* christoph.peschak@kphvie.ac.at

Umsetzung der Aufgabenstellung und die individuelle sowie abwechslungsreiche sprachliche Ausdrucksweise der Schüler*innen und die Erfüllung der Textsortenfunktion. Eine hohe Wirksamkeit zeigen Rückmeldungen auf diese Dimensionen der Texte vor allem dann, wenn sie in den Textproduktionsprozess integriert sind, der Text auf Grundlage der Rückmeldung zeitnah überarbeitet wird und abschließend eine Diskussion über die Qualität des Textentwurfs in Form eines Dialogs erfolgt (Sturm & Weder, 2018).

Da jüngere Schüler*innen erst lernen müssen, ihre eigenen Texte kritisch zu betrachten, profitieren sie besonders von klar strukturierten und nachvollziehbaren formativen Rückmeldungen der Lehrperson, die auf Grundlage von gemeinsam erarbeiteten Bewertungskriterien erfolgen. Zwar erleben die Schüler*innen durch die Schreibberatungen oder auch die verbalen Lernentwicklungsbeschreibungen eine Art Fremdbeurteilung, diese erfolgt allerdings dialogisch, denn ihr Fokus liegt auf der Verbalisierung konkreter Überarbeitungsvorschläge für den Text sowie Strategien und Hinweise für den nächsten Schreibprozess (Becker-Mrotzek & Böttcher, 2015; Philipp, 2019). Doch diese Form der Schreibberatung ist mit hohen Anforderungen an die Selbstkompetenz der Lehrperson verbunden und bedarf zeitlicher Ressourcen, die im schulischen Alltag nicht so einfach freizuspielen sind.

Hier könnten LLM wie ChatGPT eine Unterstützung für die Lehrperson darstellen. Fordert sie mit spezifischen Prompts Rückmeldung auf Texte von Schüler*innen ein, so werden diese binnen weniger Augenblicke generiert und könnten von der Lehrperson im Anschluss an die sprachlichen und inhaltlichen Bedürfnisse der Lernenden angepasst werden, sofern erforderlich. Doch entsprechen die Rückmeldungen durch ChatGPT auf Textprodukte den Kriterien der kompetenzorientierten Schreibdidaktik und des lernförderlichen Feedbacks? Verfügt das LLM überhaupt über die notwendigen Informationen zur Schreibkompetenz von Schüler*innen der Primarstufe? Und haben diese Informationen Einfluss auf die durch ChatGPT generierten Rückmeldungen auf den Text?

2 Kriterien der Textbeurteilung

Die Fokussierung auf formative Rückmeldungen, also Rückmeldungen während des Schreibprozesses, ermöglicht eine kontinuierliche Weiterentwicklung der Schreibkompetenz, während eine vorzeitige summative Rückmeldung eher motivationsmindernd und hemmend sein kann (Becker-Mrotzek & Böttcher, 2015, S. 113–144). Durch die formativen Rückmeldungen begreifen Schüler*innen ihre Texte als Entwürfe im Zuge ihrer Schreibentwicklung und erfahren dadurch, wie wichtig es ist, unfertige Textprodukte mit anderen zu diskutieren. Diese Vorgangsweise hat auch einen positiven Einfluss auf die Entwicklung von Strategien zur Bewältigung von Hürden in zukünftigen Schreibprozessen. Nach Philipp (2017) sollten Lehrpersonen ihre Rückmeldungen inhaltlich an folgenden Leitfragen orientieren:

- Welche Fähigkeiten zeigt der Text bereits?
- Welche nächsten Übungsschritte werden benötigt/sind sinnvoll?

- Welche vorhandenen Aspekte des Textes lassen sich ausbauen?

Kriterienkataloge bilden die Grundlage für diese Form der Rückmeldungen und erfüllen zusätzlich verschiedene weitere Funktionen. So können sie den Schreibprozess unterstützen, indem sie den Schülerinnen und Schülern als Hilfestellung bei der Textproduktion zur Verfügung stehen. Für die Lehrperson sollten sie die Grundlage für die Beurteilung und die Schreibberatung darstellen, denn durch sie können Kriterien gelungener und zu überarbeitender Textpassagen für die Schülerinnen nachvollziehbar dargestellt und transparent gemacht werden (Becker-Mrotzek & Böttcher, 2015). Der Vorteil von Kriterienkatalogen zur Rückmeldung auf Texte liegt darin, dass Teilleistungen in unterschiedlichen Beurteilungsdimensionen von Texten festgehalten und so auch für die Anschlusskommunikation an den Schreibprozess aufbereitet werden können. Das ist besonders nach der bewertenden-prüfenden Textbeurteilung von großer Bedeutung für die Aufrechterhaltung der Schreibmotivation (BIFIE, 2012, S. 133). Auch Eltern können die Leistungen besser einordnen, wenn die Kriterien, die der Beurteilung der Schüler*innen zu Grunde liegen, durch Kriterienkataloge transparent sind (Festman et al, 2023, S. 128).

Damit Schüler*innen Kriterienkataloge als Initiation zur Überarbeitung ihrer Texte verstehen, ist es wichtig, dass sowohl die Anzahl als auch die Art der Kriterien ihrem aktuellen Lern- und Kompetenzstand entsprechen. Becker-Mrotzek und Böttcher (2015) definieren hierfür fünf Basisdimensionen der Beurteilung: Sprachangemessenheit, Inhalt, Aufbau, Sprachrichtigkeit und Schreibprozess. Diese Kriterien, entwickelt anhand des Modells von Baurmann (2008) beinhalten weitere zwölf Subkriterien.

So vereint das Kriterium Sprachangemessenheit die Subkriterien Wortwahl und Satzbau. Sprachangemessenheit bedeutet in diesem Bezug, dass das verwendete Sprachregister und die Komplexität des Satzbaus dem Inhalt und der Funktion der Textsorte angemessen sind und semantische Leerstellen, beispielsweise unspezifische Ausdrücke wie „irgend-“ in Kombination mit Pronomen oder einem Indefinitartikel, vermieden werden (Becker-Mrotzek & Böttcher, 2015, S. 129).

Das Kriterium Inhalt eint in seinen Subkategorien jene Kriterien, die Aufschluss über die Gesamtaussage des Textes und über die inhaltliche Relevanz (nach den Griceschen Relevanzmaximen[1]) einzelner Textpassagen und deren Umfang bietet (Becker-Mrotzek & Böttcher, 2015, S. 130).

Das Kriterium Aufbau beinhaltet die Subkriterien Textmuster und Textaufbau, die Rückmeldungen auf die Angemessenheit der gewählten Textform und den Aufbau und die thematische Entfaltung des Textes geben. Zusätzlich beinhaltet das Kriterium die Rückmeldung auf die Berücksichtigung der Perspektive der Leser*innen. Die Adressatenorientierung stellt einen entscheidenden Aspekt bei der Erfüllung der Textsortenfunktion dar (vgl. Augst, 2007; Behrens 2017). Für die Bildung der deduktiven Kategorien des Kodierleitfadens der Studie wurden die Kriterien Sprachangemessenheit, Inhalt und Aufbau inklusive einiger Subkriterien übernommen.

2.1 Die Schreibabsicht Erzählen

Narrative Texte oder Erzähltexte sollen ein Geschehen möglichst präzise darstellen. Dieses Geschehen wird nach Martinez (2017, S. 2) anhand dreier Kriterien definiert: Konkretheit, Temporalität und Kontiguität. Das Kriterium Konkretheit bezieht sich darauf, dass eine konkrete Situation geschildert wird. Die in dieser Situation stattfinden Ereignisse werden chronologisch und strukturiert beschrieben (Temporalität). Die Kontiguität bezieht sich auf die Kausalität und die Bezüge der erzählten Inhalte zueinander. Diese müssen nicht nur in einer zeitlich nachvollziehbaren Abfolge beschrieben werden, sondern müssen auch in einem sinnvollen Zusammenhang zueinander stehen. Die Funktion von Erzähltexten liegt vor allem in der Unterhaltung der Leser*innen (Festman et al., 2023, S. 84).

Textprodukte, die der Textsortenfunktion entsprechen, kennzeichnen sich sprachlich durch eine abwechslungsreiche Wortwahl bei Inhalts- und Funktionswörtern, die in einer lebendigen Erzählweise resultiert sowie einer genauen Beschreibung der Gedanken, Gefühle und der Sinneseindrücke der handelnden Personen. Figurenreden werden hier passend in den Text integriert. Die Leser*innen werden nicht nur über Sachverhalte informiert, sondern unterhalten. Dafür gestalten die Autor*innen die Texte entsprechend emotional bewegend (z. B. spannend, traurig oder fröhlich) und in einem einheitlichen Erzählton. Die meist fiktionalen Texte enthalten einen nachvollziehbaren Planbruch (vgl. Quasthoff, 1980, S. 54) und eine Pointe mit einem Überraschungsmoment (Festman et al., 2023, S. 103—105).

3 Studiendesign

Ziel der Untersuchung war eine vergleichende Analyse von Rückmeldungen durch ChatGPT auf Texte aus dem Gesamtkorpus der Longitudinalstudie von Augst et al. (2007). Die Texte mit der Schreibabsicht Erzählen wurden von Schüler*innen der zweiten und vierten Klasse verfasst. Die Längsschnittstudie hat untersucht, wie sich die Schreibkompetenzen von Schüler*innen im Laufe der zweiten, dritten und vierten Klasse der Primarstufe entwickeln. Dabei verfassten 39 Schüler*innen in regelmäßigen Abständen Texte mit den Schreibabsichten Erzählen, Instruieren & Erklären, Beschreiben, Berichten und Appellieren. Dabei zeigten textstrukturelle, lexikalische, syntaktische und sprachlich-diskursive Analysen der Texte, dass der Entwicklungsprozess der Schreibkompetenz von Primarstufenschüler*innen durch ein vierstufiges Modell beschrieben werden kann, das eine Wechselbeziehung von Sprachentwicklung, Textsortenentwicklung und Textentwicklung aufweist (Augst et al., 2007; Behrens 2017).

In der vorliegenden Studie wurden Rückmeldungen durch ChatGPT ausgehend von zwei Basisprompts (im weiteren Text mit BP1 und BP2 abgekürzt) evoziert.

- **BP1**: Gib Rückmeldung zum nachfolgenden Text. Gehe auf folgende Kriterien ein: Sprachlicher Ausdruck, Inhalt und Aufbau des Textes.

- **BP2**: Gib Rückmeldung zum nachfolgenden Text. Gehe auf folgende Kriterien ein: Sprachlicher Ausdruck, Inhalt und Aufbau des Textes. Die Rückmeldung ist für ein Kind in der zweiten Klasse/vierten Klasse Grundschule.

BP1 wurde dem „ChatGPT-Guide für Lehrkräfte. Version 4.0“ (Flick, 2025) entnommen. Dort wurde er als einziges Beispiel dafür angeführt, wie Lehrpersonen ChatGPT zu Rückmeldungen auf Texte auffordern können.

BP2 stellt in gewisser Weise den Vergleichsprompt dar, auf dessen Grundlage untersucht wurde, ob und inwiefern sich die Rückmeldungen von ChatGPT durch die Nennung einer konkreten Zielgruppe verändern. BP2 enthält also zusätzlich einen Hinweis auf die Zielgruppe, für die die Rückmeldung gedacht ist.

Die Rückmeldungen von ChatGPT auf BP1 wurden in den Analysekorpora A und C, die Rückmeldungen auf BP2 in den Analysekorpora B und D gesammelt, wobei Analysekorpus A und B die Rückmeldungen auf die Texte aus der zweiten Klasse, Analysekorpus B und C die Rückmeldungen auf die Texte der vierten Klasse beinhalteten. Das Gesamtanalysekorpus für die Untersuchung der Rückmeldungen durch ChatGPT bestand aus insgesamt 78 Texten mit der Schreibabsicht Erzählen ($n_{2.Klasse}$ = 39; $n_{4.Klasse}$ = 39). Auf Basis einer qualitativen Inhaltsanalyse (Kuckartz & Rädiker, 2024) wurden die Rückmeldungen auf die Schüler*innentexte mittels deduktiver Kategorien segmentiert und die Ergebnisse in einem weiteren Schritt quantifiziert und analysiert.

Die Entscheidung für die deduktive Vorgehensweise liegt darin begründet, dass passende Definitionskriterien für die Analysekategorien aus bestehenden Arbeiten zur Beurteilung von Schüler*innentexten abgeleitet werden konnten. Hier dienten die Beurteilungskriterien für Schüler*innentexte nach Becker-Mrotzek & Böttcher (2015) als Grundlage für die Definition der Hauptkategorien K1, K2 und K3 und der Unterkategorien K1.1, K1.2, K2.1, K2.2, K3.1 und K3.3. Auch die Kodierregeln der jeweiligen Unterkategorien orientierten sich inhaltlich an den Definitionen der Basisdimensionen der Textkorrektur nach Becker-Mrotzek & Böttcher (2015).

Die Unterkategorien K1.1.1 bis K3.2.2 wurden ebenfalls deduktiv ermittelt, da im Vorfeld der Analyse Kriterien entwickelt werden mussten, die die Rückmeldungen durch ChatGPT als zielgruppen- bzw. nicht zielgruppenadäquat definieren. Zielgruppenadäquatheit ergab sich in dem Zusammenhang daraus, dass die in den Rückmeldungen verwendeten Wörter und Phrasen aufgrund ihrer sprachlichen Struktur als für die jeweilige Zielgruppe geläufig und daher verständlich angenommen werden konnten. Die Rückmeldungen sollten zur besseren Verständlichkeit auch durch konkrete Beispiele zur Verbesserung von Textteilen ergänzt werden. Wörter und Phrasen, die aufgrund ihrer sprachlichen Komplexität bzw. Ambiguität als nicht zielgruppenadäquat angenommen werden konnten, wurden ebenfalls kodiert und den zielgruppenadäquaten Äußerungen anhand ihrer Anzahl gegenübergestellt. Dies soll einen Vergleich in Hinblick auf die Forschungsfrage ermöglichen, ob ChatGPT in der Lage ist, die Rückmeldungen an die Bedürfnisse einer bestimmten Zielgruppe anzupassen.

K1 Sprachangemessenheit	K1.1 Lexik	K1.1.1 zielgruppenadäquat
		K1.1.2 nicht zielgruppenadäquat
	K1.2 Syntax	K1.2.1 zielgruppenadäquat
		K1.2.2 nicht zielgruppenadäquat
K2 Inhalt	K2.1 Inhaltliche Kohärenz	K2.1.1 zielgruppenadäquat
		K2.1.2 nicht zielgruppenadäquat
	K2.2 Inhaltliche Relevanz	K2.2.1 zielgruppenadäquat
		K2.2.2 nicht zielgruppenadäquat
K3 Aufbau	K3.1 Textmuster & -aufbau	K3.1.1 zielgruppenadäquat
		K3.1.2 nicht zielgruppenadäquat
	K3.2 Adressatenorientierung	K3.2.1 zielgruppenadäquat
		K3.2.2 nicht zielgruppenadäquat

Tabelle 1: Kodierleitfaden und Kategoriensystem

Für die Kategorie K1.1 Lexik wurden jene Textsegmente kodiert, die auf die Verbesserung von Inhalts- und Funktionswörtern, wie Konjunktionen, Präpositionen oder Verben, aufmerksam machten und so Hinweise zur Präzisierung der inhaltlichen Ausgestaltung und zur Verbesserung der sprachlichen Verständlichkeit beinhalteten.

Für die Kategorie K1.2 Syntax wurden jene Textsegmente kodiert, die Rückmeldungen auf die Satzstruktur in Zusammenhang mit der Erfüllung der Textsortenfunktion beinhalteten, also beispielsweise auf Nebensätze mit unterschiedlichen Konjunktionen, Hauptsatzverknüpfungen mit unterschiedlichen Konjunktionen oder die Figurenrede hinwiesen und zur Verbesserung dieser Textelemente anregten.

Für die Kategorie K2.1 Inhaltlicher Kohärenz wurden Textsegmente kodiert, die, unter Berücksichtigung der Erfüllung der Textsortenfunktion, Rückmeldungen auf den inhaltlichen Aufbau des Erzähltextes gaben und Vorschläge zur Verbesserung der inhaltlichen Gliederung beinhalteten. Diese Verbesserungsvorschläge bezogen sich hierbei anhand konkreter Beispiele vor allem auf den Gesamtinhalt und den Aufbau des Ausgangstexts.

Für Kategorie K2.2 Inhaltliche Relevanz wurden jeden Textsegmente kodiert, die Rückmeldungen auf die inhaltliche und sprachliche Ausgestaltung konkreter Passagen des Ausgangstextes und Verbesserungsvorschläge in Hinblick auf die Erfüllung der Textsortenfunktion gaben.

Für Kategorie K3.1 Textmuster & -aufbau wurden Textsegmente kodiert, die Rückmeldungen auf den formalen Aufbau des Textes unter Berücksichtigung der Strukturen eines Erzähltextes gaben. Hier wurden vor allem Äußerungen kodiert, die einen inhaltlichen Gesamtüberblick über den Ausgangstext boten und seinen Aufbau mit dem eines typischen Erzähltextes verglichen.

In Kategorie K3.2 Adressatenorientierung wurden jene Textsegmente kodiert, die sprachlich und inhaltlich explizit auf die Perspektive der Leser*innen hinwiesen und darauf aufbauend Vorschläge zur Verbesserung des Ausgangstextes gaben.

Jeder Kategorie wurden auch passende Textpassagen als Ankerbeispiel zugeordnet. Dabei erfolgte die Entscheidung über die zu kodierenden Segmente anhand der Definition der

Kategorie und der festgelegten Kodierregeln. Bei der Analyse der Daten zeigte sich, dass die Kategorien gut voneinander abzugrenzen waren und lediglich ein Revisionsdurchgang benötigt wurde, um nachvollziehbare Erkenntnisse aus dem Material abzuleiten. Der Kodierleitfaden wurde zusätzlich in einer Lehrveranstaltung zur Analyse von weiteren Rückmeldungen von ChatGPT auf Schüler*innentexte angewendet und im Sinne des konsensuellen Kodierens überprüft.

Kategorie	Subkategorie	Ankerbeispiele
K1.1 Lexik	K1.1.1 zielgruppenadäquat	„das ging in eine Höhle“ – Das „das“ sollte „das Mädchen“ sein, um klarer zu machen, wer gemeint ist. (Analysekorpus A, Pos. 95) Es gibt aber ein paar kleine Wiederholungen, zum Beispiel „Ich war in eine Höhle gegangen“ kommt zweimal vor. Das könntest du abwechslungsreicher machen. Vielleicht so: „Mutig betrat ich eine dunkle Höhle.“ (Analysekorpus B, Pos. 6)
	K1.1.2 nicht zielgruppenadäquat	„da kam etwas mit leuchtenden Augen auf sie zu“ könnte präziser formuliert werden: „Da kam etwas mit leuchtenden Augen auf sie zu.“ (Analysekorpus A, Pos. 724) „In eine gruselige Gewitternacht war ich weit von zu Hause fort.“ → Besser: „An einer gruseligen Gewitternacht war ich weit von zu Hause entfernt.“ (Flüssigere Wortwahl.) (Analysekorpus B, Pos. 197)

Tabelle 2: Ausgewählte Ankerbeispiele aus Analysekorpus A und B für die Kategorie K1.1 Lexik (2. Klasse)

Kategorie	Subkategorie	Ankerbeispiele
K1.1 Lexik	K1.1.1 zielgruppenadäquat	Der Text ist insgesamt verständlich, aber einige Sätze könnten klarer und flüssiger formuliert werden. Zum

		Beispiel: „Weil dort viele Spinnen sind“ könnte präziser sein, etwa: „Weil dort viele große Spinnen lauern.“ (Analysekorpus C, Pos. 110) „Diese Trolle bauen sich unter der Erde Höhlen.“ (Das Wort „machen“ ist hier nicht so passend.) (Analysekorpus D, Pos. 213)
	K1.1.2 nicht zielgruppenadäquat	Der sprachliche Ausdruck ist einfach und für ein Kinderbuch gut geeignet. Es gibt jedoch einige Stellen, an denen der Satzbau und die Wortwahl verbessert werden könnten, um den Text flüssiger und klarer zu machen (Analysekorpus C, Pos. 35) „Als ein Zwerg auf dem Heimweg war, war ein großes Unwetter da.“ (Das „ein“ am Anfang macht den Satz flüssiger.) (Analysekorpus D, Pos. 441)

Tabelle 3: Ausgewählte Ankerbeispiele aus Analysekorpus C und D für die Kategorie K1.1 Lexik (4. Klasse)

3.1 Limitationen

Die Texte, die im Zuge der Studie von Augst et al. (2007) entstanden sind, stellen insgesamt ein sehr heterogenes Textkorpus von Texten unterschiedlicher Qualität dar. Dabei muss die Aufgabenstellung, auf Grundlage derer die Schüler*innentexte entstanden sind, kritisch betrachtet werden. So wurden die Schüler*innen mittels Bildimpuls[2] und dem Schreibauftrag „Für das Geschichtenbuch. Denk dir zu diesem Bild eine interessante Geschichte aus und schreibe sie für das Geschichtenbuch auf.“ zum Verfassen der Texte aufgefordert (Augst et al, 2007, S. 47). In dieser Form entspricht die Aufgabenstellung nicht mehr den Kriterien der kompetenzorientierten Schreibdidaktik. So fehlen in der Aufgabenstellung beispielsweise konkrete Hinweise auf die lebensweltliche Situierung oder auf die Adressaten des Textes (vgl. dazu Schmölzer-Eibinger, 2015). Es könnte daher sein, dass bei einer Adaption des Schreibauftrags die Schüler*innen entsprechend andere, wahrscheinlich komplexere, Texte produziert hätten.

Eine weitere Limitation der Studie liegt im Umgang mit ChatGPT. Die Abfragen der Rückmeldungen auf die Basisprompts fanden aus zeitlichen Gründen nur jeweils einmal zu jedem Basisprompt für die jeweiligen Texte statt. Ob sich ausgehend vom selben Prompt zu einem späteren Abfragezeitpunkt inhaltlich eine andere Rückmeldung auf denselben Text ergeben hätte, wurde nicht überprüft. Offen bleibt also, inwiefern sich die Rückmeldungen bei einer

späteren zweiten Abfrage auf ein und denselben Text inhaltlich von der ursprünglichen Rückmeldung auf den Basisprompt unterschieden hätte.

Zusätzlich konnte nur eingeschränkt auf das aktuelle Modell GPT-4o zugegriffen werden, da kein kostenpflichtiges Abo für die vorliegende Studie abgeschlossen wurde. Das bedeutete, dass nach einigen Abfragen automatisch auf ältere Modelle, wie GPT-3.5, zurückgegriffen wurde, was sich vor allem in der sprachlichen Ausgestaltung der Rückmeldungen bemerkbar machte. Rückmeldungen auf Basis des Modells GPT-4o hatten im Vergleich zu den Rückmeldungen, die auf Basis von GPT-3.5 oder niedriger erstellt wurden, einen sehr hohen Anteil an zielgruppenadäquat kodierten Segmenten. Dieser Aspekt wurde jedoch bewusst in Kauf genommen, da davon ausgegangen werden kann, dass die Mehrheit der Lehrpersonen ebenfalls die kostenfreie Version von ChatGPT für ihre Abfragen verwendet und sich dieser Wechsel in den Modellen somit auch im beruflichen Alltag vollziehen kann.

4 Zusammenschau der Analyseergebnisse

Vergleicht man die Rückmeldungen durch ChatGPT auf Grundlage der jeweiligen Basisprompts und der unterschiedlichen Ausgangstexte, so sind teils markante und Unterschiede und überraschende Ergebnisse in der Qualität der Rückmeldungen auszumachen, vor allem, was die Zielgruppenadäquatheit betrifft. Aber auch in der Quantität an Rückmeldungen bezogen auf die unterschiedlichen Beurteilungsdimensionen der Texte zeigen sich große Unterschiede. Alle Analysekorpora wiesen unterschiedliche Häufigkeiten Rückmeldungen zu den einzelnen Kategorien auf.

Auch die weitläufige Annahme, dass ChatGPT bei wenig konkreten Prompts den Fokus automatisch auf die Korrektur von Formalfehlern in der Orthografie oder der Zeichensetzung legen würde, konnte nicht bestätigt werden. Im Gegenteil: Selbst beim wenig konkreten BP1 bezogen sich nahezu alle Rückmeldungen auf die inhaltlich-ästhetische Dimension des Ausgangstextes und nicht auf Korrekturvorschläge von Formalfehlern. Im Vergleich zu den insgesamt 1.414 nach dem Kodierleitfaden kodierten Segmenten, wurden lediglich 55 Segmente kodiert, die zusätzlich Rückmeldungen auf die Korrektur von Orthografie und Zeichensetzung gaben.

4.1 Analysekorpus A und Analysekorpus B im Vergleich

Von den 343 kodierten Segmenten im Analysekorpus A wurden 103 Segmente (rund 30%) der Kategorie K3.1 Textmuster &-aufbau zugeordnet.

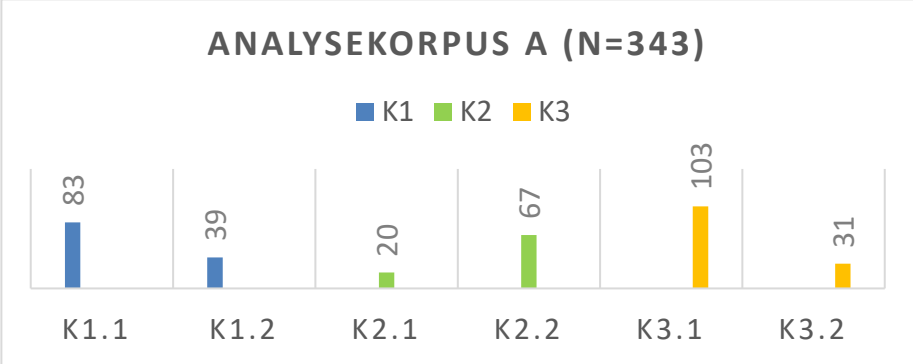

Abbildung 1: Darstellung der kodierten Segmente im Analysekorpus A in absoluten Häufigkeiten

Im Vergleich dazu erfolgten im Analysekorpus B mit 143 (rund 41%) kodierten Segmenten die meisten Rückmeldungen auf die Kategorie K1.1 Lexik. Den markantesten Unterschied zwischen Analysekorpus A und B lässt sich in der Kategorie K3.2 Adressatenorientierung feststellen. Während im Analysekorpus B lediglich 7 Segmente dieser Kategorie zugeordnet werden konnten, fanden sich im Analysekorpus A insgesamt 31 und damit auch mehr kodierbare Segmente als für die Kategorie K 2.1 Inhaltliche Relevanz im selben Korpus.

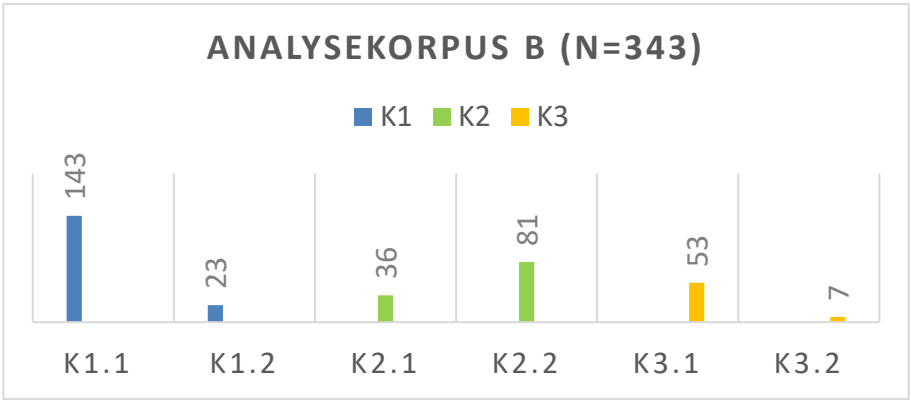

Abbildung 2: Darstellung der kodierten Segmente im Analysekorpus B in absoluten Häufigkeiten

Auch in Hinblick auf den Anteil der zielgruppen- bzw. nicht zielgruppenadäquaten Rückmeldungen sind große Unterschiede zwischen den beiden Analysekorpora auszumachen. Der Anteil an zielgruppen- und nicht zielgruppenadäquaten Aussagen in der Kategorie K1.1 Lexik hält sich in beiden Analysekorpora die Waage (siehe Abbildung 4 und 5). Die Kategorie K2.2 Inhaltliche Relevanz sticht jedoch beim Vergleich der Datensätze im Analysekorpus B durch ihren überwiegenden Anteil an als zielgruppenadäquat kodierten Segmenten heraus. Auffallend in beiden Korpora ist die geringe Anzahl an nicht zielgruppenadäquaten Rückmeldungen in der Kategorie K2.2 Inhaltliche Relevanz.

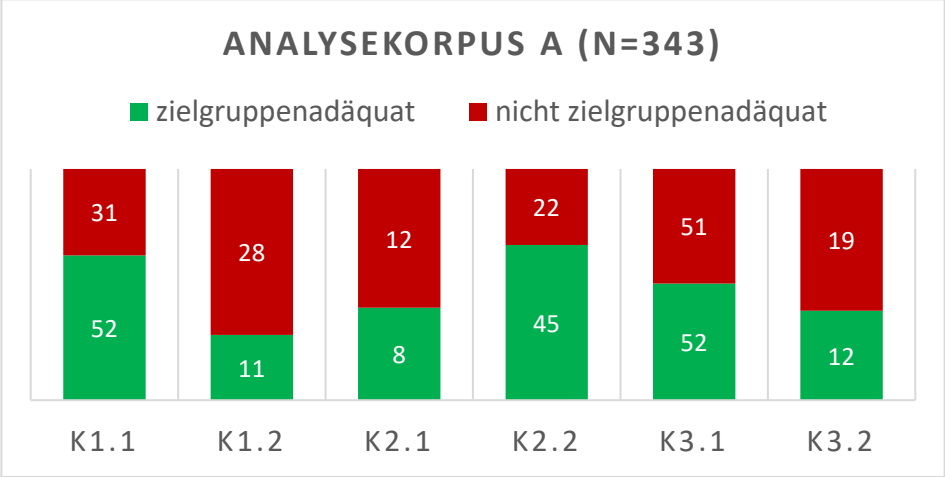

Abbildung 3: Verhältnis zwischen zielgruppenadäquaten und nicht zielgruppenadäquaten Äußerungen im Analysekorpus A

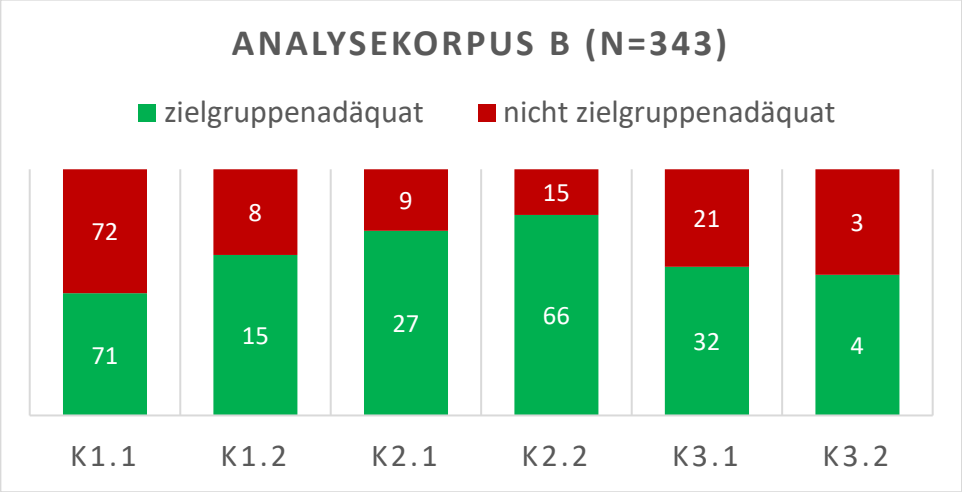

Abbildung 4: Verhältnis zwischen zielgruppenadäquaten und nicht zielgruppenadäquaten Äußerungen im Analysekorpus B

4.2 Analysekorpus C und Analysekorpus D im Vergleich

Bei den kodierten Textsegmenten im Analysekorpus C (N=334) sticht vor allem die große Zahl von 119 (rund 35%) kodierten Segmente heraus, die der Kategorie 2.2 Inhaltliche Relevanz zugeordnet werden konnte. Im Vergleich dazu fanden sich im Analysekorpus C eher wenige Rückmeldungen auf Kategorie K1.1 Lexik. Wie auch schon im Analysekorpus B konnten die wenigsten Textsegmente der Kategorie K3.2 Adressatenorientierung zugeordnet werden.

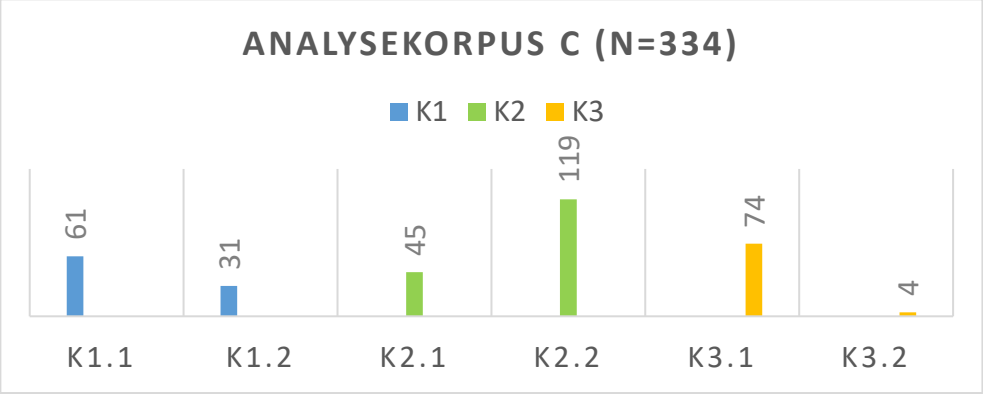

Abbildung 5: Darstellung der kodierten Segmente im Analysekorpus C in absoluten Häufigkeiten

Im Analysekorpus D ergab sich jedoch ein gänzlich anderes Bild. Die Rückmeldungen auf Grundlage von BP2 fokussierten gleich zwei Kategorien gleichermaßen: K1.2 Syntax und K2.2 Inhaltliche Relevanz. Die Kategorie K3.2. Adressatenorientierung stellt allerdings in beiden Analysekorpora mit nur 4 Segmenten in Analysekorpus C und nur 7 Segmenten in Analysekorpus D einen nur geringen Anteil an der Gesamtzahl der segmentierten Elemente dar.

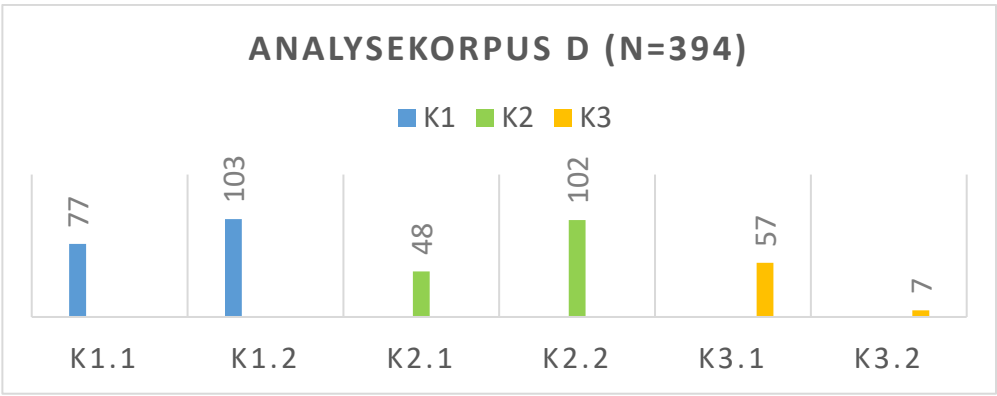

Abbildung 6: Darstellung der kodierten Segmente im Analysekorpus D in absoluten Häufigkeiten

In Hinblick auf die Zielgruppenadäquatheit der Rückmeldungen auf die Texte der Viertklässler*innen lassen sich markante Unterschiede im Vergleich zu den Analysekorpora A und B feststellen. In beiden Analysekorpora C und D ist der Anteil an zielgruppenadäquaten Rückmeldungen wesentlich höher als bei den Analysekorpora A und B. Vor allem Analysekorpus D weist einen äußerst geringen Anteil an nicht zielgruppenadäquaten Rückmeldungen auf. Dies könnte auch auf die Ausgangstexte zurückzuführen sein, die aufgrund ihrer Länge und Struktur mehr Anknüpfungspunkte für die Rückmeldungen durch ChatGPT bietet, als es die kurzen Texte der Zweitklässler*innen im Analysekorpus A und B tun.

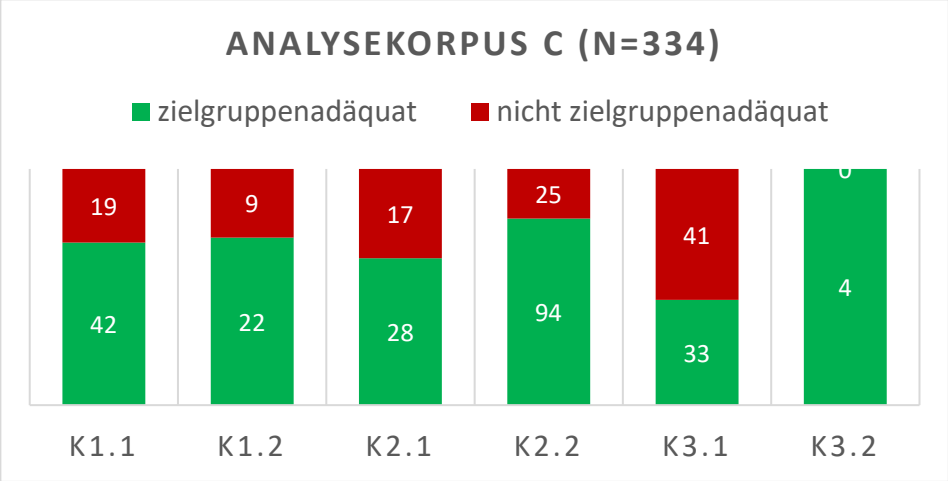

Abbildung 7: Verhältnis zwischen zielgruppenadäquaten und nicht zielgruppenadäquaten Äußerungen im Analysekorpus C

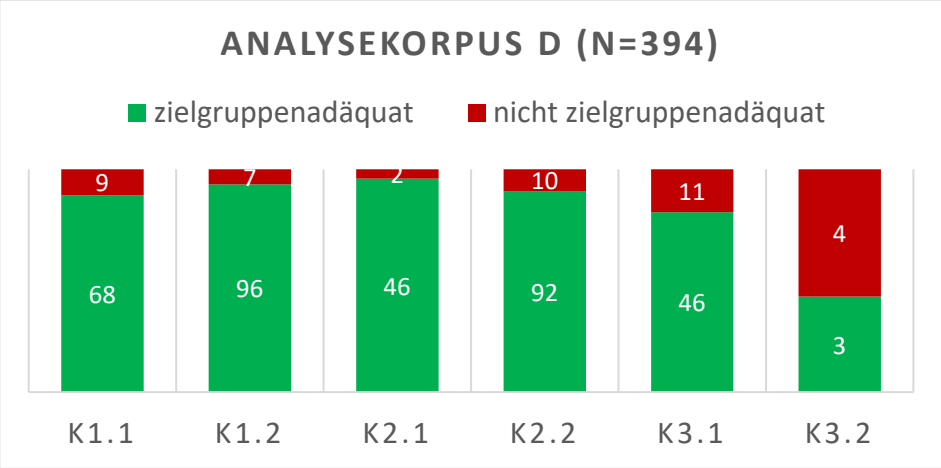

Abbildung 8: Verhältnis zwischen zielgruppenadäquaten und nicht zielgruppenadäquaten Äußerungen im Analysekorpus D

5 Fazit und Ausblick

Die Rückmeldungen durch ChatGPT auf die Texte der Analysekorpora entsprechen in einem Punkt allesamt den Kriterien der kompetenzorientierten Schreibdidaktik: Die Rückmeldungen gehen mit konkreten Nachfragen inhaltlich auf Passagen des Ausgangstextes ein, was auch von der Lehrperson in Hinblick auf die mäeutische Haltung gegenüber Schüler*innentexten gefordert wird. Ebenso beziehen sich die Rückmeldungen nur in geringem Ausmaß auf Verbesserungsvorschläge hinsichtlich der normativen Sprachrichtigkeit. Viel mehr stehen Kategorien wie die inhaltliche Relevanz einzelner Textpassagen, die angemessene Wortwahl oder auch der passende Textaufbau im Fokus der Rückmeldungen.

Die Qualität der Rückmeldungen hängt dabei auch vom Umfang sowie der inhaltlichen Komplexität und Kohärenz der Ausgangstexte ab. So waren Rückmeldungen auf Grundlage von BP2 auf Texte der Schüler*innen der vierten Klasse zielgerichteter und wiesen einen

geringen Anteil an nicht zielgruppenadäquaten Formulierungen auf als die Rückmeldungen auf die Texte der Schüler*innen der zweiten Klasse. Das hängt auch mit dem teilweise geringen Umfang und den eingeschränkten sprachlichen Mitteln der Schüler*innentexte zusammen, auf deren Grundlage die Erstellung von Rückmeldungen tendenziell schwieriger ist. Die Basisprompts sollten daher in einer weiteren Analyse auf Grundlage von anderen Textkorpora (bestehend beispielsweise aus umfangreicheren Texten von Schüler*innen der Sekundarstufe) auf ihre Zielgerichtetheit hin überprüft werden.

Zwar zeigt die Analyse auch, dass die Rückmeldungen durch ChatGPT zum Teil sprachlich zu allgemein und terminologisch unpräzise formuliert sind. Dies lässt sich jedoch durch die Präzisierung des Prompts, wie beispielsweise durch die Nennung einer konkreten Zielgruppe, wesentlich verbessern. Dieser Sachverhalt, in Verbindung mit dem hohen Anteil an zielgruppenadäquaten Rückmeldungen auf die Erzähltexte der vierten Klasse, spricht dafür, dass Lehrpersonen, mit einem konkret formulierten Prompt ausgestattet, für die Rückmeldungen auf unterschiedliche Beurteilungsdimensionen von Texten durchaus auch auf ChatGPT zurückgreifen können. Die Analyse zeigt, dass ChatGPT durchaus Rückmeldungen auf Texte geben kann, die den Kriterien eines lernförderlichen Feedbacks entsprechen. Dafür muss jedoch, wie bei anderen Interaktionen mit ChatGPT auch, der Eingabeprompt entsprechend explizit die Bedürfnisse der jeweiligen Zielgruppe, für die die Rückmeldung gedacht ist, benennen. Ist der Eingabeprompt für die Textrückmeldungen entsprechend konkret, so kann ChatGPT eine Unterstützung für die Lehrperson beim Verfassen von verbalen Rückmeldungen auf Texte von Schüler*innen sein.

Literatur

Augst, G., Disselhoff, K., Henrich, A., Pohl, T., & Völzing P.-L. (2007). *Text – Sorten – Kompetenz. Eine echte Longitudinalstudie zur Entwicklung der Textkompetenz im Grundschulalter.* Peter Lang.

Baurmann, J. (2008). *Schreiben, Überarbeiten, Beurteilen. Ein Arbeitsbuch zur Schreibdidaktik.* Friedrich Verlag.

Becker-Mrotzek, M., & Böttcher, I. (2014). *Schreibkompetenz entwickeln und beurteilen* (6. Aufl.). Cornelsen.

Behrens, U. (2017). Entwicklung der Schreibkompetenz: Vorschule und Primarstufe. In. M. Becker-Mrotzek, J. Grabowski & T. Steinhoff (Hrsg.) *Forschungshandbuch empirische Schreibdidaktik.* Waxmann.

BIFIE (2012). *Themenheft für den Kompetenzbereich „Verfassen von Texten" Deutsch, Lesen, Schreiben. Volksschule Grundstufe I+II.* https://www.iqs.gv.at/downloads/nationale-kompetenzerhebung/materialien-zu-ikm-und-bildungsstandards/publikationen-deutsch

Festman, J., Gerth, S., Mairhofer, M., & Reiter C. (2023). *Texte verfassen in der Primarstufe. Theorie und Praxis für erste Schreibprozesse, Textproduktion und Schreibdidaktik.* Waxmann.

Flick, M. (2025). *ChatGPT-Guide für Lehrkräfte 4.0. ChatGPT für Schule und Unterricht einsetzen.* https://www.manuelflick.de/chatgpt-guide

Kuckartz, U., & Rädiker, S. (2024). *Qualitative Inhaltsanalyse. Methoden, Praxis, Umsetzung mit Software und künstlicher Intelligenz.* Beltz Juventa.

Martinez ,M. (2017). Was ist Erzählen? In M. Martinez (Hrsg.). *Erzählen. Ein interdisziplinäres Handbuch.* J.B. Metzler, S. 2–6.

Ossner, J. (2006). Kompetenzen und Kompetenzmodelle im Deutschunterricht. *Didaktik Deutsch,* (21).

Philipp, M. (2017). Förderung hierarchiehoher Schreibprozesse. In M. Philipp (Hrsg.), *Handbuch Schriftspracherwerb und weiterführendes Lesen und Schreiben.* Beltz, S. 285–299.

Quasthoff, U. M. (1980). *Erzählen in Gesprächen. Linguistische Untersuchungen zu Strukturen und Funktionen am Beispiel einer Kommunikationsform des Alltags.* Narr.

Schmölzer-Eibinger, S. (2015). *Kriterien für „gute" Schreibaufgaben.* https://static.uni-graz.at/fileadmin/gewi-zentren/fachdidaktikzentrum-gewi/Dokumente/Kriterien_Erstellung_von_Schreibaufgaben.pdf

Sturm, A., & Weder, M. (2018). *Schreibkompetenz, Schreibförderung, Schreibmotivation. Grundlagen und Modelle zum Schreiben als soziale Praxis.* (2. Aufl.). Klett Kallmeyer.

[1] Siehe dazu H.P. Grice (1975). Logic and Conversation. In: Studies in the way of words. Cambridge University. In: C. Peter & J. Morgan (Hrsg.). Syntax and Semantics 3: Speech Acts. Academic Press.

[2] Verfügbar unter https://www.uni-koeln.de/phil-fak/deutsch/pohl/tsk/PDFs/Schreibauftraege.pdf.