



**Einsatz von künstlicher Intelligenz im
Überarbeitungsprozess von Texten in der
Fremdsprache Deutsch: Fokus auf Wortschatz und
Syntax**

Stéphanie Roussel & Ayaal Herdam, Bordeaux

ISSN 1470 – 9570

Einsatz von künstlicher Intelligenz im Überarbeitungsprozess von Texten in der Fremdsprache Deutsch: Fokus auf Wortschatz und Syntax

Stéphanie Roussel & Ayaal Herdam, Bordeaux

Textproduktion im Fremdsprachenerwerb beinhaltet mehrere Phasen und darunter eine Überarbeitungsphase. Mehrere Studien haben gezeigt, dass die Aufmerksamkeit der Lernenden dabei stärker auf Wortschatz und Syntax als auf Inhalt und Kohärenz gerichtet ist (Hall 1990; Wang & Wen 2002). Diese Studie untersucht, wie sich die Nutzung künstlicher Intelligenz (KI) auf diesen Überarbeitungsprozess auswirkt. Fünfundzwanzig französischsprachige Jurastudierende, die Deutsch als Fremdsprache lernen, schrieben einen Text (Version 1) und überarbeiteten ihn anschließend mit ChatGPT, um eine zweite Version zu schreiben (Version 2). Die Analyse der Texte und der Dialoge mit der KI zeigt, dass ChatGPT zwar oft zuverlässige Übersetzungen und relevante syntaktische Korrekturen bieten, jedoch keine zutreffenden Erklärungen zur grammatikalischen Struktur der deutschen Sprache liefern kann. Diese Ergebnisse eröffnen didaktische Perspektiven.

Text production during second language acquisition involves several stages, including a revision phase. Several studies have shown that learners' attention is more focused on vocabulary and syntax than on content and coherence during this phase (Hall 1990; Wang & Wen 2002). This study investigates how the use of artificial intelligence (AI) affects this revision process. Twenty-five French law students learning German as a foreign language first wrote a text (Version 1) and then revised it using ChatGPT to create an improved version (Version 2). The analysis of the texts and the exchanges with the AI programme show that while ChatGPT can provide reliable translations and relevant syntactic corrections, it cannot offer accurate explanations about the grammatical structure of the German language. These results open didactic perspectives.

1. Einleitung

Da Textproduktion als komplexer Problemlöseprozess (Kertes 2022; Lemke & Hoffmann 2022) verstanden wird, beschäftigt sich die schreibdidaktische Forschung mit der Frage, wie man Lernende dabei unterstützen kann, ihre Texte produktiv zu bearbeiten (Ruhmann & Kruse 2014). So sind in den letzten Jahren im Bereich der Fremdsprachendidaktik vermehrt Studien durchgeführt worden, die zum Beispiel die Vermittlung von Strategien zur Überprüfung, Überarbeitung, oder Selbsteinschätzung einer Textproduktion (Sethuraman & Radhakrishnan 2020), kollaboratives Schreiben (Zhang, Gibbons & Li 2021) oder die Nutzung von Korpora zur Überprüfung der Akzeptabilität bestimmter Konstruktionen (Schaefer-Lacroix 2015) hinterfragen. In

jüngerer Zeit wurde auch die Verwendung von künstlicher Intelligenz (KI) untersucht (Hartmann 2021; Warschauer et al. 2023; Pecorari 2023a), was dazu anregt, über deren Integration in den Fremdsprachenunterricht nachzudenken.

Welche Elemente ziehen bei der Überarbeitung eines Textes in der Fremdsprache die Aufmerksamkeit der Lernenden auf sich und werden in der Interaktion mit der KI thematisiert? Wie integrieren die Lernenden die Rückmeldungen der KI?

Als Beitrag zu diesen von der technologischen Entwicklung aufgeworfenen Fragen wurde eine Studie mit fünfundzwanzig Studierenden im ersten, zweiten und dritten Studienjahr des Jurastudiums an der Universität Bordeaux durchgeführt, die unter anderem einen Kurs für Deutsch als Fremdsprache bzw. als Fachsprache belegt haben. Die Teilnehmer*innen wurden gebeten, einen zweihundert Wörter langen Text auf Deutsch zu schreiben. Anschließend überarbeiteten sie diese erste Version, indem sie gezielt die Konversationsfunktion von ChatGPT nutzten. In der zweiten Version markierten die Studierenden die geänderten, korrigierten, erweiterten oder verbesserten Passagen farblich, was es ermöglichte, die Beiträge der KI deutlich zu isolieren und zu analysieren. Es wurden dann quantitative und qualitative Analysen der Interaktionen der Studierenden mit ChatGPT durchgeführt, wobei in dem vorliegenden Beitrag der Fokus auf den Gesprächsausschnitten zwischen den Studierenden und dem Chatbot lag, die Wortschatz und Syntax thematisieren. Nach einem Überblick über die Forschung zu den in den vergangenen Jahren aufgeworfenen Fragestellungen zur Textproduktion im Fremdsprachenerwerb mithilfe künstlicher Intelligenz, werden die Ergebnisse dieser Studie präsentiert und die Perspektiven für den Deutschunterricht diskutiert.

2. Theoretischer Rahmen

2.1 Merkmale der Textproduktion in der Fremdsprache

Textproduktion ist eine komplexe und kognitiv anspruchsvolle Tätigkeit, die in zahlreichen Modellen beschrieben wird (Hayes & Flower 1980; Chenoweth & Hayes 2001; Lemke & Hoffmann 2022). In der Erstsprache wie in der Fremdsprache umfasst die Textproduktion drei Phasen: Planen (Ideen generieren, Ziele setzen, Strukturieren), Formulieren und Überarbeiten (Lesen und Revidieren). Diese Phasen laufen zugleich parallel und rekursiv ab, bauen aufeinander auf und können sich wiederholen (Hayes & Flower 1980). Die Textproduktion in der Fremdsprache hat einige Besonderheiten

(Börner 1989; Johnson 2020). Bei mehrsprachigen Personen werden die verschiedenen Sprachen während der Textproduktion gleichzeitig aktiviert (Bialystok 2001). Wenn sie die Zielsprache noch nicht ausreichend beherrschen, verbringen die Produzent*innen mehr Zeit damit, von der Erstsprache in die Fremdsprache zu übersetzen, was die Arbeitsgedächtnisbelastung erhöhen kann (Sasaki & Hirose 1996). Die Prozesse der Hemmung oder Aktivierung von Informationen im Gedächtnis während des Schreibens sind verlängert, was das Gefühl der Schwierigkeit bei der Durchführung einer Aufgabe erhöhen kann (Kroll et al. 2015). Darüber hinaus führt die fehlende Automatisierung der Prozesse, die für die Textproduktion erforderlich sind, zu einer erhöhten kognitiven Belastung (Barbier 2012). Diese Erkenntnisse stehen im Einklang mit den spezifischen grundlegenden Merkmalen, die das Modell von Börner (1989) beschrieben hat. Das Modell trägt nämlich den Besonderheiten der Textproduktion im Fremdspracherwerb Rechnung und hebt die Unterschiede zur Muttersprache hervor. Bei der Planung können Unterbrechungen auftreten, wenn das sprachliche oder kulturelle Wissen unzureichend ist. Während der Formulierung kann ein Missverhältnis zwischen dem Wunsch sich auszudrücken und den begrenzten sprachlichen Fähigkeiten entstehen (siehe auch Kertes 2022). Was uns im Rahmen dieser Studie insbesondere interessiert, sind die Merkmale der Überarbeitungsphase. Nach Kertes (2022: 92) „wird den sprachlichen, lexikalischen und grammatikalischen Mitteln des fremdsprachlichen Textes größere Aufmerksamkeit geschenkt als beim sprachlichen Revidieren eines muttersprachlichen Textes“. Dieser auf sprachliche, lexikalische und grammatikalische Elemente gelegte Schwerpunkt sollte sich in der Nutzung Künstlicher Intelligenz durch die Studierenden widerspiegeln, wenn sie ihre fremdsprachlichen Texte überarbeiten.

2.2 Textproduktion mit der KI in dem Fremdsprachenunterricht

Wie Hartmann (2021), Warschauer et al. (2023) oder Pecorari (2023a) beschrieben haben, ist das Interesse für KI-Technologien im Bereich des Fremdsprachenlernens in jüngster Zeit stark gestiegen. Eine Vielzahl von Anwendungen beim Erlernen von Sprachen und im Fremdsprachenunterricht wurde unter die Lupe genommen. Studien zu Duolingo und zu anderen Sprachlernsoftware-Angeboten oder zu maschineller Übersetzung (*Machine Translation, MT*) haben die Relevanz von automatischer Spracherkennung zur Schaffung motivierender Lernumgebungen demonstriert (Bourdais & Guichon 2020, Lee 2019). Dabei werden die positiven Auswirkungen der Verwendung von maschinellen Übersetzungstools auf die Schreibstrategien der Lernenden und deren

Reflexion über die Textproduktion als Prozess betont. In diesem Bereich verbessern intelligente Tutorsysteme und Schreibassistenten, darunter Grammarly und GPT-3 von OpenAI, die Textgenerierung und Textüberarbeitung erheblich (Barrot 2023; Koltovskaia 2020). Durch das schnelle Feedback (z. B. Fehlerberichte) „können die Lernenden sofort Maßnahmen ergreifen, um ihre sprachlichen ‚Insuffizienzen‘ zu korrigieren und sich anschließend zu verbessern“ (Strasser 2020: 3).

Trotz mehrerer Grenzen und Unsicherheitszonen, wie zum Beispiel der geringen Kenntnis der Datenschutzrichtlinien oder des Plagiatsrisikos (Pecorari 2023b), untersuchen immer mehr akademische Arbeiten den Nutzen von Tools wie ChatGPT für die schriftliche Produktion in der Fremdsprache (Imran & Almusharraf 2023; Warschauer et al. 2023). Forscher*innen stellen fest, dass bestimmte Aufgaben während der drei Phasen des Schreibprozesses (Planung, Formulierung und Überarbeitung) in gewisser Weise an die KI delegiert werden können: die Ideenfindung, das Abrufen von Informationen (Chenoweth & Hayes 2001), die Übersetzung, die Identifizierung von sprachlichen Fehlern, lokale Anpassungen (Imran & Almusharraf 2023) oder das Einholen von kontextuellen Informationen. Während der Überarbeitung überprüfen die Schreibenden kritisch ihren Text, führen eine Selbsteinschätzung durch (Athanasopoulos 2023) und nutzen dann die KI rekursiv, um (erneut) bestimmte Aufgaben zu erfüllen, die zur Planung oder Formulierung gehören können. KI könnte also aus der Perspektive der Verteilung der Aufmerksamkeitsressourcen im Arbeitsgedächtnis eine Entlastung der gesamten Schreibprozesse erlauben, indem sie die für das Schreiben benötigte Zeit und den Aufwand reduziert und ein personalisiertes Lernen ermöglicht (Kohnke *et al.* 2023).

Ein weiterer wichtiger Aspekt der Reflexion über Tools wie ChatGPT betrifft die Bedingungen für das Erzielen relevanter Antworten. Die Qualität der Antworten hängt nämlich teilweise von der Genauigkeit der Anfragen an die KI (sogenannte *Prompts*) ab. Da ChatGPT ein unüberwachtes Lernmodell ist, kann es falsche Antworten generieren, die einige Forscher*innen sogar als „beunruhigend“ bezeichnet haben (Las-Vergnas & Rio-Jouet 2023). Diese Verfasser*innen haben beispielsweise in einem Austausch mit Bard-Google Inkonsistenzen oder Widersprüche, Nichtbeachtung kritischer Beobachtungen, Halluzinationen (Erfindung fiktiver Daten zur Untermauerung eines Arguments) und die Anpassung an die Autorität des Gesprächspartners identifiziert.

Warschauer et al. (2023: 2) überprüften auch die scheinbare Widersprüchlichkeit bei der Nutzung der KI, die nicht unüblich ist, wenn es um digitale Technologien geht: „The

‚rich get richer‘ contradiction“. Um die KI zum Schreiben in einer Fremdsprache nutzen zu können, wären eine gute Beherrschung der betreffenden Sprache und die Fähigkeit, effektiv zu schreiben unverzichtbare Voraussetzungen. Hilft die KI dann beim Erlernen der Sprache oder ist die Beherrschung der Fremdsprache eine Voraussetzung, um dank der KI noch mehr zu lernen? In dieser Perspektive zeigen die Studien von Barrot (2023) und Koltovskaia (2020), dass die englischsprachigen Lernenden die Rückmeldungen von Tools wie Grammarly umso relevanter und selektiver integrieren, je höher ihr Kompetenzniveau ist. Weniger kompetente Studierende sind weniger in der Lage, die Schwächen in ihren Texten selbst zu erkennen und zu analysieren, sie vertrauen zu sehr auf automatisches Feedback und integrieren es oft ohne Rücksicht auf den unmittelbaren sprachlichen Kontext zu nehmen oder betreiben das, was Howard (1992) als *Patchwriting* bezeichnet: aus einem Quelltext zu kopieren, dann einige Wörter zu löschen, sie einzeln durch Synonyme zu ersetzen oder die grammatikalischen Strukturen zu ändern. Um die KI vollständig zu nutzen, scheinen die Schreibenden über bestimmte Fähigkeiten verfügen zu müssen. Sie sollten sich der Schwächen ihrer Texte bewusst sein, Stimuli für die Revision liefern, die Grenzen der KI-generierten Antwort erkennen, um so das Output der KI effektiv verarbeiten zu können und dann in ihre eigene Produktion zu integrieren. Wenn es um eine Fremdsprache geht, gibt es aus Gründen, die mit der Verteilung der Aufmerksamkeitsressourcen und der Sättigung des Arbeitsgedächtnisses zusammenhängen (Sweller et al. 2022), gute Anhaltspunkte zu der Annahme, dass diese Bedingungen nicht immer erfüllt sein können.

3. Methodologie und Datenerhebung

Fünfundzwanzig Studierende (weiblich: zwanzig; männlich: fünf) des ersten, zweiten und dritten Jahres des Rechtsstudiums an der Universität Bordeaux, die einen Kurs für Deutsch als Fachsprache belegt haben, nahmen auf freiwilliger Basis an dieser Studie teil. Die Teilnehmer*innen sind im Alter von 18 bis 21 Jahren und ihr Sprachniveau liegt im Durchschnitt bei B2 (Gemeinsamer Europäischer Referenzrahmen für Sprachen, 2001). Elf von ihnen sind im deutsch-französischen *Bachelor of Law* eingeschrieben, zu dessen Zulassungsbedingungen ein B2-Niveau in Deutsch gehört. Für unser Experiment bilden sie eine Gruppe von 13 Studierenden mit guten Deutschkenntnissen (B2 bis C1), die unter anderem einen Luxemburger sowie eine Person umfasst, die ein Jahr in Deutschland verbracht hat. Die anderen zwölf Studierenden haben ein bescheideneres

Deutschniveau (B1 bis B2); sie haben in der Sekundarstufe Deutsch als Fremdsprache gelernt.

Alle teilnehmenden Personen wurden eingeladen, einen zweihundert Wörter langen Text auf Deutsch zu schreiben, in dem sie eine Videoreportage über die Umweltaktivistengruppe „Letzte Generation“ zusammenfassen. Dieses Video mit dem Titel „Geldstrafe für Unterstützer der ‚Letzten Generation‘“, datiert auf den 19. Juli 2023, dauert 2:24 Minuten, wurde von der *Frankfurter Allgemeine Zeitung (FAZ)* online gestellt und ist auf *YouTube* verfügbar.¹ Nach der Erstellung einer ersten Version der Zusammenfassung, wurden sie gebeten, diese unter gezielter Nutzung der Konversationsfunktion von ChatGPT zu einer zweiten verbesserten Version zu überarbeiten und die mit Hilfe der KI modifizierten, korrigierten, verlängerten oder verbesserten Passagen farblich zu markieren. Die teilnehmenden Personen durften zehn Fragen formulieren, ohne dem Chatbot ihren Text vollständig vorzulegen. Aus zeitlichen Gründen haben aber manche von ihnen weniger als zehn Fragen gestellt, und einige haben die Aufgabestellung nicht respektiert und mehr als 10 Fragen gestellt, meist weil es ihnen wichtig war, die Aufgabe so gut wie möglich zu erfüllen. Ziel war es herauszufinden, ob die Fragesteller*innen in der Lage waren, Schwächen in ihrem Text zu erkennen und die KI zur Verbesserung zu nutzen. Insgesamt besteht unser Corpus aus den 50 Textproduktionen der Teilnehmer*innen (25 Versionen ohne KI, 25 Versionen mit KI), 180 Fragen (im Durchschnitt 7,2 Fragen pro Person) der Studierenden an ChatGPT und den entsprechenden Antworten der KI.

Für eine frühere Studie (Roussel & Ochoa, 2025 im Druck), haben wir die Produktionen der Studierenden quantitativ anhand von vier Dimensionen (Wortschatz, Syntax, Inhalt, Verständlichkeit) analysiert und mit insgesamt maximal 20 Punkten bewertet. Als Ergebnisse konnten wir festhalten, dass die Bewertung der Textproduktionen der Studierenden nach der Überarbeitung mit ChatGPT im Durchschnitt um 2 Punkte stiegen und dass Wortschatz und Syntax die am stärksten von der Überarbeitung betroffenen Dimensionen waren. Zwei Deutschlehrende (ein deutscher Muttersprachler und eine französische Muttersprachlerin) haben die beiden Versionen der schriftlichen Produktionen im Doppelblindverfahren (unabhängig und anonym) geprüft und bewertet. Der Cohen-Kappa-Koeffizient zwischen den beiden Bewertungsreihen der beiden Prüfenden

¹ <https://www.youtube.com/watch?v=Fqn3UIk5BY>.

betrug 0,285. Dieser Wert deutet darauf hin, dass eine relativ große Übereinstimmung zwischen den beiden Prüfenden besteht, die über das hinausgeht, was zufällig zu erwarten wäre, auch wenn es Unterschiede in ihren Bewertungen gibt.

Für die vorliegende Studie präsentieren wir einen weiteren Aspekt dieser Daten. Eine quantitative und eine qualitative Analyse der 180 Interaktionen (Frage + Antwort) der Teilnehmer*innen mit ChatGPT wurde durch dieselben beiden Prüfenden durchgeführt. Unser Ziel ist es jetzt, die Fragen der Lernenden und die Antworten der KI zu kategorisieren, um zunächst zu verstehen, auf welche Punkte die Lernenden bei der Überarbeitung ihres Textes achten und welche Art von Antworten die KI auf diese Fragen geben kann. Wir wollten ebenfalls untersuchen, wie der Fokus auf den Wortschatz, den wir in den Textproduktionen der Teilnehmer*innen festgestellt hatten, in der Interaktion mit der KI Gestalt annimmt. Für eine bessere Lesbarkeit unserer Datenanalyse und um die Anonymität zu wahren, haben wir allen teilnehmenden Personen einen Buchstaben von A bis Y zugewiesen. Außerdem haben wir jedem Zitat der Interaktionen mit GPT eine Zahl in Klammern zugeordnet, um leichter auf diese Auszüge verweisen zu können. Da es sich bei den Zitaten um individuelle Äußerungen handelt, verzichteten wir bei der Nennung der zitierten Personen auf genderneutrale Schreibweise. Bei manchen Zitaten geben wir unter a) die Anfrage der teilnehmenden Person, unter b) die Antwort des Chatbots und unter c) die eventuell in der Version 2 des Textes zu beobachtende Umsetzung der Antwort an.

Die insgesamt 180 Anfragen der Studierenden (unten im Original, also auch oft mit Fehlern wiedergegeben) lassen sich in vier verschiedene Kategorien aufteilen: Lexik, Morphosyntax, kontextuelle Informationen, allgemeine Ratschläge für die Redaktion. Rein lexikalische Anfragen, zum Beispiel die Übersetzung ins Französische einzelner Wörter und Wortgruppen ohne explizite Einbettung in einen Kontext und ohne Flexion gehören in die erste Kategorie. Dabei handelt es sich sowohl um Fragen, die das Verstehen des in der Reportage Gehörten und Gesehenen zum Ziel haben, also zum Beispiel die Bitte um Übersetzung eines sprachlichen Elements aus dem Deutschen in die Muttersprache, wie beispielsweise bei den Teilnehmerinnen M und W:

(1) *Que signifie* „Gesellschaftsrat“?

(Was bedeutet „Gesellschaftsrat“?)²

² In Klammern geben wir eine Übersetzung der Anfragen der Studierenden an.

als auch um Anfragen, die die Integration von Wortschatz in den eigenen Text ermöglichen sollen, zum Beispiel Bitten um Übersetzung von lexikalischen Sequenzen ohne ausdrückliche Flexion aus der Muttersprache ins Deutsche oder die Überprüfung von Sequenzen, die im Deutschen existieren könnten, wie bei Teilnehmer T:

(2) *Comment dit-on „condamner en prison“ en allemand?*

(Wie sagt man „condamner en prison“ auf deutsch?)

bzw. bei Teilnehmerin H:

(3) *Est-ce que le mot „klimacrise“ existe en allemand?*

(Existiert das Wort „klimacrise“ in der deutschen Sprache?)

Eine zweite Kategorie stellen die morphosyntaktischen Anfragen dar, die eine Einbettung von lexikalischen Elementen in eine syntaktische Einheit, also in einen Satz oder in einen Textabschnitt beinhalten, z. B. bei Teilnehmerin N:

(4) *“Die Zivilisation hat über den Klimanotstand viele Ängst“, est-ce que tu peux corriger les cas?*

(„Die Zivilisation hat über den Klimanotstand viele Ängst“, kannst du die Fälle korrigieren?)

oder mit denen die Studierenden Auskunft über morphologische, syntaktische oder semantische Eigenschaften bestimmter Wörter und Sequenzen erlangen wollen, mit dem Ziel, sie in ihren eigenen Text einzubringen, wie bei Teilnehmerin J:

(5) *Comment utilise-t-on le verbe „achten“?*

(Wie benutzt man das Verb „achten“?)

Am Beispiel (5) wird deutlich, dass die Antworten auf Anfragen, die in diese zweite Kategorie fallen, komplexer sein können als rein lexikalische Anfragen, denn die Person, die die Frage stellt, möchte möglicherweise nicht nur über Konjugation und Rektion, sondern auch über die Verwendungsmöglichkeiten des Wortes in bestimmten Kommunikationssituationen informiert werden. Eine Bitte um Korrektur, wie im Beispiel (4), wird von den Studierenden manchmal durch eine Bitte um Erklärung der Korrekturen ergänzt, was ebenfalls eine komplexere Antwort erfordert z. B. bei Teilnehmerin N:

(6) *Est-ce que tu peux expliquer pourquoi tu as changé?*

(Kannst du erklären, warum du es verändert hast?)

Die Übergänge zwischen den beiden erstgenannten Kategorien („Lexik“ und „Morphosyntax“) sind natürlich fließend, wenn es sich um die Übersetzung oder die Korrektur von Sequenzen handelt, wie etwa bei Teilnehmerin R

(7) *Comment dit-on „préservation du climat“ en allemand?*

(Wie sagt man „*préservation du climat*“ auf Deutsch?)

In diesen Fällen haben wir die Anfrage der Kategorie 1 (Lexik) zugeordnet, da keine Flexion im Sinne einer Deklination erfragt wurde, auch wenn die Kombination der entsprechenden Lexeme im Deutschen eine morphologische Operation erfordert, in diesem Fall die Aneinanderreihung der Elemente *Klima* und *-schutz* in genau dieser Reihenfolge.

Mit den übrigen Anfragen werden allgemeine Informationen zum Thema oder allgemeine Ratschläge zur Redaktion eines Textes erbeten, z. B. bei Teilnehmer C:

(8) Was ist am 14. März 2023 in Köln passiert?

bzw. bei Teilnehmerin K

(9) Wie würdest du einen Text beginnen, der eine Reportage wiedergibt?

Was die Antworten von ChatGPT angeht, haben wir sie in erster Linie aus sprachdidaktischer Perspektive betrachtet und als zutreffend oder als fehlerhaft bzw. fragwürdig bewertet (Tabelle 2).

4. Ergebnisse

4.1 Sprachbeherrschung als Element der Analyse der Antworten von GPT

Wie im vorherigen Teil gezeigt und in der Tabelle 1 präsentiert, betreffen die meisten Fragen der Studierenden lexikalische oder syntaktische Elemente.

Kategorie	Anteil
Lexik	70%
Morphosyntax	25%
Kontextuelle Informationen	3,5%
Allgemeine Ratschläge für die Redaktion	1,5%

Tabelle 1: Anteil der Anfragen der Teilnehmer an GPT nach Kategorien

Da ihr Deutschniveau relativ homogen ist, im Durchschnitt auf B2 (Gemeinsamer Europäischer Referenzrahmen für Sprachen, 2001), ist es schwierig, einen wirklichen Ein-

fluss des Anfangsniveaus hervorzuheben. Einige Unterschiede in der Reflexion über die Sprache und den Kontext können dennoch in ihren Interaktionen mit der KI beobachtet werden. Aus didaktischer Sicht besteht einer der positiven Effekte der Kommunikation mit dem Chatbot darin, dass der Dialog die Studierenden dazu anregt, ihre Anfragen und somit ihre Überlegungen zu ihrem eigenen Text zu präzisieren. Die strategische Entscheidung über den Umfang der Sequenzen, die die Lernenden dem KI-Programm zur Übersetzung oder zur Korrektur vorlegen konnten, war durch die Aufgabenstellung im Prinzip auf maximal zehn nicht zusammenhängende Sätze begrenzt. Einige beschränkten sich dabei durchweg auf einzelne Wörter, andere erbat die Korrektur oder die Übersetzung ganzer Textabschnitte, ohne Rücksicht auf die Aufgabenstellung. Welche Faktoren bei dieser und ähnlichen Entscheidungen (so beispielsweise bei der Entscheidung über die Anzahl der Anfragen) eine Rolle spielen, ist nicht der Hauptgegenstand dieser Untersuchung, aber wir nehmen an, dass der Grad der Sprachbeherrschung, das Allgemeinwissen und die Erfahrung im Umgang mit KI einen Einfluss auf die Kommunikationsstrategie der Studierenden gegenüber dem Chatbot haben. Ein Teil der studentischen Anfragen sind Rückfragen, die auf eine erste Antwort reagieren und die Anwendbarkeit der Antwort auf den vorgegebenen Kontext und in der eigenen Textproduktion überprüfen, was nach unserer Meinung den Ansatz einer kritischen Haltung zu den KI-generierten Antworten illustriert und ein gewisses Niveau der Sprachbeherrschung voraussetzt. Im Folgenden geben wir einige Beispiele für Dialoge von Studierenden mit relativ hoher Sprachkompetenz (zwischen B2 und C1) an.

Teilnehmer X, als Luxemburger praktisch deutschsprachig, fragt, wie er das Verhängen einer Freiheitsstrafe mit Bewährung formulieren soll. Der Chatbot liefert die Sequenz aus der Sicht des Richters, der die Strafe verhängt. X versteht, dass das nicht zur Textsorte „Zusammenfassung einer Reportage“ passt und fragt nach einer Version, die einem journalistischen Diskurs entspricht.

Teilnehmerin V (mit hoher Sprachkompetenz), fragt nach der Übersetzung von „*bloquer une rue*“. Der KI-generierte, durchaus passende Vorschlag „eine Straße blockieren“ gefällt ihr nicht. Sie fragt, ob sie in diesem Kontext „festnehmen“ verwenden kann. Der Chatbot verneint und schlägt als Alternativen „die Straße sperren“ und „den Verkehr blockieren“ vor. V entscheidet sich schließlich für die letzte Variante, die in der Tat die beste ist.

Teilnehmerin S fragt zuerst nach einem Adjektiv für „*Klimaktivist“. Sie versteht, dass die Antwort des Chatbots („man kann das Wort „klimaaktivist“ als Adjektiv verwenden“) seltsam ist und fragt nach Synonymen, erhält „Klimaengagierter“ und legt dann dem Chatbot ihren ganzen Text mit diesem Wort zur Korrektur vor.

Teilnehmerin O, mit relativ hoher Sprachkompetenz, testet den Satz „Das hat viele andere Probleme kreiert“. Der Chatbot ersetzt „kreiert“ durch „verursacht“. Aber O hat das Gefühl, dass das Wort „kreieren“ im Deutschen existiert und fragt dementsprechend nach. Der Chatbot antwortet, dass im Kontext mit Problemen die Alternative „verursachen“ besser passt. O akzeptiert und verwendet den KI-generierten Vorschlag in ihrer Produktion.

Oft suggeriert der Chatbot, dass die fragestellende Person mehr Kontext liefert, z. B. im Dialog mit Teilnehmerin W, die zuerst fragt, wie man „*trainer un humain*“ übersetzt. Der Chatbot liefert „ziehen“ und „schleifen“, fragt nach dem Kontext, W reagiert darauf („*trainer un humain sur le sol*“) und der Chatbot schlägt dann passend „jemanden über den Boden schleifen“ vor.

In den Dialogen der weniger kompetenten Studierenden mit der KI sind Rückfragen nahezu abwesend. Man kann sagen, dass die Antworten der KI für bare Münze genommen werden und sich die Fragen fast ausschließlich auf den Wortschatz vom Typ „Wie sagt man...?“ beschränken. Wir haben kein Beispiel, in dem sie auf einen Vorschlag von ChatGPT zurückkommen würden oder eine Antwort des Chatbots hinterfragen würden. Eine der wenigen Ausnahmen findet sich bei der Teilnehmerin I:

(10) *Warum machen die Klima Aktivisten Blockade an der Straße?

Der Chatbot antwortet und die Teilnehmerin bittet ihn daraufhin um eine Übersetzung der Antwort.

Diese Beobachtungen geben einen Hinweis darauf, wie das Sprachniveau der Studierenden den Dialog mit der KI beeinflussen kann.

4.2 KI-generierte Antworten und ihre Verwendung in der Textproduktion

4.2.1 Übersicht der KI-Fehlerkategorien

Die Antworten des Chatbots lassen sich ebenfalls kategorisieren. Sie können zum Beispiel mehr oder weniger ausführlich sein und sie können die Realität mehr oder weniger adäquat abbilden. Wie weiter oben ausgeführt, lassen sich die Antworten der

KI aus sprachdidaktischer Perspektive als zutreffend oder als fehlerhaft bzw. fragwürdig bewerten (Tabelle 2).

	Anzahl der Anfragen	zutreffend beantwortet	fehlerhaft oder fragwürdig beantwortet
Lexik	129	107 (82%)	22 (17%)
Morphosyntax	40	17 (42,5%)	23 (57,5%)
kontextuelle Informationen	8	8 (100%)	0 (0%)
allgemeine Ratschläge für die Redaktion	3	3 (100%)	0 (0%)
Gesamt	180	135 (75%)	45 (25%)

Tabelle 2: Anzahl und Natur der KI generierten Antworten

Generell lässt sich sagen, dass die meisten Antworten (75%) richtig oder zumindest nachvollziehbar sind. Das gilt allerdings in besonderem Maße für Antworten auf die rein lexikalischen Anfragen. Hier treffen 82% der Vorschläge des Chatbots eindeutig zu. Für die Morphosyntax bietet sich ein anderes Bild. Von den 40 Antworten des Chatbots in dieser Kategorie sind nur 17 völlig zutreffend (42,5%), 12 sind fehlerhaft und 11 weitere sind zumindest teilweise fragwürdig (insgesamt 23 teilweise problematische Antworten, also 57,5%). Die restlichen Antworten auf Anfragen nach allgemeinen kontextuellen Informationen oder nach allgemeinen Ratschlägen zur Textgestaltung sind im Wesentlichen zutreffend, wobei zu beachten ist, dass der Chatbot darauf hinweist, dass das Programm nur auf Informationen zurückgreift, die in seiner Datenbank vorliegen. Neuere Ereignisse oder Erscheinungen, die nach der letzten Aktualisierung aufgetreten sind, entziehen sich also der Kenntnis des Programms. Die KI-generierte Antwort:

(11) eine Gruppe namens Letzte Generation ist mir nicht bekannt

(auf eine Frage der Teilnehmerin V) wurde folglich nicht als fehlerhaft bewertet, auch wenn sie als einzige verfügbare Information keinen direkten Nutzen für die Textproduktion der Studierenden hervorbringt. Welche KI-generierten Antworten sind also kritikwürdig?

4.2.2. Detaillierte Analyse der Fehlerkategorien

4.2.2.1. Lexikfehler

a) Unkenntnis neuer Begriffe oder Erscheinungen

Unter den relativ wenigen Fehlern bei den Antworten des Chatbots auf rein lexikalische Anfragen finden sich einige unpassende Synonyme oder Bedeutungen, insbesondere bei neueren Begriffen.

Im oben genannten Beispiel (1) ist laut ChatGPT „Gesellschaftsrat“ ein Synonym für „Aufsichtsrat eines Unternehmens“ oder „Vorstand eines Vereins“. Das ist eine Bedeutung, die das Wort im vorgegebenen Kontext sicher nicht hat. Die von manchen Umweltschutzgruppen geforderten Gesellschaftsräte sollen alternative Entscheidungsgremien sein, die nach den Vorstellungen ihrer Initiatorinnen und Initiatoren gesamtgesellschaftliche Interessen bei umweltrelevanten Fragen durchsetzen können. In dieser Bedeutung taucht das Wort allerdings erst seit Anfang 2022 in den Medien auf, offensichtlich nach der zum Test-Zeitpunkt jüngsten Aktualisierung.

Das Fehlen von neueren Informationen in der Datenbank dürfte auch erklären, warum das Programm die Sequenz

(12a) (Anfrage der Teilnehmerin J) *Klimaaktivisten haben sich auf der Straße geklebt

zunächst unerwartet in

(12b) (KI-generierte Antwort) Klimaaktivisten haben sich auf der Straße versammelt

korrigiert. Der politische Aktionsmodus des Sich-auf-die-Straße-Klebens war dem Programm unbekannt und es wählte die nächste wahrscheinliche Lösung unter den in der Datenbank zur Verfügung stehenden Optionen. Dass die Fragestellerin in (12a) nach der Präposition „auf“ den Dativ (und nicht den in dieser Situation zu erwartenden Akkusativ) verwendet hat, mag hier eine Rolle gespielt haben. Die Substitution von „geklebt“ durch „versammelt“ ist zwar eine radikale lexikalische Änderung, aber sie vermeidet eine syntaktische Operation, da der Dativ beibehalten werden kann. Letzteres hätte das Programm im Übrigen auch ohne radikale lexikalische und semantische Veränderung durch die Substitution von „geklebt“ durch „festgeklebt“ erreichen können. Nach Rückfrage durch die Teilnehmerin bestätigt der Chatbot schließlich die fehlerhafte Sequenz (12a) als korrekt. Sie findet sich folgerichtig in dieser Form auch in der studentischen Textproduktion wieder.

Die Unkenntnis des Eigennamens „Letzte Generation“ führt außerdem dazu, dass die in mehreren Anfragen auftretende Sequenz

(13) *letze Generation

vom Programm nicht als fehlerhaft erkannt und folglich in den KI-generierten Korrekturvorschlägen sowie in der studentischen Textproduktion in mehreren Fällen beibehalten wird. Das weiter unten als (24b) angeführte Beispiel ist eine Ausnahme.

b) Fehlerhafte Interpretation durch Beschränkung auf eine einzige oder wenige Optionen

Andere KI-generierte Lexikfehler sind kaum durch die Abwesenheit aktueller Informationen zu erklären. So korrigiert das Programm die Sequenz

(14a) (Teilnehmerin H) *viele Menschen haben zu einige Monaten Haft verurteilen

in

(14b) viele Menschen haben vor einigen Monaten Haftstrafen erhalten.

Menschliche Lehrkräfte würden (14a) wohl auch ohne Kontext überwiegend richtig als „viele Menschen wurden zu einigen Monaten Haft verurteilt“ dekodieren und in diesem Sinne korrigieren. Obwohl das Programm für (14b) eine Übersetzung in die Muttersprache angibt, hat die Fragestellerin die KI-generierte Antwort schließlich *verbatim* in ihrer eigenen Textproduktion verwendet, auch wenn ihre erste, mit sprachlichen Mängeln behaftete Sequenz den Inhalt der Reportage besser widerspiegelt als der KI-generierte Vorschlag.

Die Sequenz (14b) ist zwar nicht absurd, aber sie ist eben nur eine mögliche Interpretation, die in diesem Fall vom Programm für die wahrscheinlichste gehalten und als einzige vorgeschlagen wird. Der Vorschlag von mehreren Optionen wäre hier hilfreich gewesen.

Ähnliches lässt sich über das folgende Beispiel sagen:

(15) *die Aktivisten sind gestoppt bei Polizei.

In diesem Fall hatte die Teilnehmerin I die Übersetzung des Wortes „*arrêtés*“ erfragt, das Programm lieferte die Antworten „gestoppt“ und „angehalten“, aber nicht die Antwort „festgenommen“, die dem Kontext besser entsprochen hätte. Zweifellos wäre es aus Sicht der Fragestellerin effektiver gewesen, dem Programm die gesamte Sequenz (15) zur Korrektur vorzulegen, anstatt lediglich ein Wort zu erfragen.

Das Programm erkennt im Übrigen nicht immer, dass die von ihm vorgeschlagenen lexikalischen Optionen identisch sind, so im Beispiel

(16a) (Teilnehmerin R) *Le maire en allemand*

(16b) (KI-generierte Antwort) Die Übersetzung von „*le maire*“ ist „der Bürgermeister“ oder einfach „der Bürgermeister“.

Die Antwort des Chatbots führt in diesem Fall nicht zu einer fehlerhaften Textproduktion, aber die Präsenz des identischen zweiten und vermeintlich einfacheren Vorschlags ist zumindest verwirrend.

In die Unterkategorie der limitativen Fehlinterpretationen dürften auch Missverständnisse fallen, die durch eine wörtliche Auslegung der studentischen Anfragen entstehen, wie im folgenden Beispiel:

(17a) (Teilnehmer T) *Trouve-moi une image qui illustre les actions de protestation des militants écologistes!*

(Finde ein Bild, das die Protestaktionen der Umweltschützer illustriert!)

(17b) (KI-generierte Antwort) Es tut mir leid, aber als textbasiertes Sprachmodell bin ich nicht fähig, Bilder zu suchen oder zu liefern. Sie können Bilder-Suchmaschinen benutzen. Beachten Sie dabei das Urheberrecht.

Diese Antwort ist formal korrekt, ein globales Verständnis der Kommunikationssituation hätte allerdings dazu geführt, dass das Programm das Wort „Bild“ im Sinne einer sprachlichen Metapher interpretiert. Wir gehen nicht davon aus, dass T hier tatsächlich für den Zweck einer rein sprachlichen Zusammenfassung eine ikonographische Darstellung erwartet hatte. Dass ein Bild dem Teilnehmer geholfen hätte, die Situation besser zu verstehen, ist eher unwahrscheinlich, da das Ausgangsdokument ohnehin ein Video ist und da der besprochene Aktionsmodus zum Zeitpunkt der vorliegenden Studie allgemein bekannt war. Natürlich hätte T den Sinn oder den Kontext seiner Anfrage präzisieren können, aber er hat darauf verzichtet und diese Antwort des Chatbots in seiner Textproduktion nicht weiter beachtet.

4.2.2.2. Morphosyntaktische Fehler

a) Deklination

Problematischer sind die KI-generierten Antworten auf Anfragen, die wir der Kategorie Morphosyntax zugeordnet haben. Im Folgenden sei ein typisches Beispiel für den Dialog

mit dem Chatbot angeführt:

(18a) (Teilnehmerin P) *Est-ce que "in den Reportage" est correct?* (Ist „in den Reportage“ korrekt?)

(18b) (KI-generierte Antwort) In der deutschen Sprache ist das Wort Reportage feminin. Die Präposition „in“ verlangt den Akkusativ. Der korrekte Satz ist also „in die Reportage“.

(18c) (Textproduktion der Fragestellerin) *Es gibt auch mehrere Interviews in die Reportage.

Die KI-generierte Antwort (18b) weist hier mindestens zwei offensichtliche Schwächen auf. Erstens wird nicht erwähnt, dass die Präposition „in“ nicht nur einen Kasus nach sich zieht. Warum das Programm hier nur den Akkusativ suggeriert, bleibt unklar. Möglicherweise tritt die Präposition in der Datenbank vorwiegend in Verbindung mit dem Akkusativ auf. In Sachtexten ist das allerdings nicht der Fall, wovon man sich leicht überzeugen kann, wenn man die beiden letzten Sätze des vorliegenden Artikels und auch die KI-generierte Antwort (18b) selbst beobachtet: „In“ wurde viermal verwendet und leitete in allen vier Fällen eine Nominalgruppe im Dativ ein.

Eine andere mögliche Erklärung ist, dass die Sequenz „in verlangt den Akkusativ“ in dieser oder ähnlicher Form in der Datenbank des Programms oft oder an prominenter Stelle auftaucht. Das würde bedeuten, dass diese Datenbank selbst nicht repräsentativ ist oder dass das Programm die zuerst gefundene Information als die einzig wichtige ansieht. Letztere Hypothese wirft unter anderem die Frage auf, warum das Programm die Option „Akkusativ“ als erste findet. Das könnte mit den unterschiedlichen Traditionen der Präsentation von Deklinationsparadigmen zusammenhängen. Vorstellbar wäre ebenfalls, dass die Anfrage (18a), die ja bereits den Gebrauch des Akkusativs (hier unter der fehlerhaften Annahme, dass das französische Wort Reportage auch im Deutschen männlich ist) suggeriert, die Antwort des Programms beeinflusst hat. An einer anderen Stelle desselben Dialogs der Fragestellerin mit dem Chatbot wird allerdings auch deutlich, dass das Programm nicht mit einer Repräsentation der Konzepte „Akkusativ“ oder „Dativ“ arbeitet, die der einer menschlichen Lehrkraft entsprechen würde:

(19a) (Teilnehmerin P) Ist „in der zweite Prozess“ korrekt?

(19b) „Im“ ist die Kontraktion von „in“ und „dem“ und „zweiten“ ist die korrekte Form für Akkusativ maskulin. Der korrekte Satz ist also „im zweiten Prozess“.

In dieser Situation führte die KI-generierte Antwort nicht zu einer fehlerhaften Textproduktion, aber die Erklärung des Chatbots weist auf eine Verwechslung oder zumindest auf eine partielle Gleichsetzung von Akkusativ und Dativ hin. Analog dazu findet man in den KI-generierten Antworten auf Anfragen der Teilnehmerin V die kuriose Behauptung

(20) Das Demonstrativpronomen „dieses“ ist maskulin neutral.

oder den klaren Fehler

(21) Die Wörter „Auswirkung“ und „Einfluss“ sind im Deutschen feminin.

Das Programm geht also relativ frei mit Begriffen wie Genus und Kasus sowie mit den dazugehörenden Realitäten um.

Nun zur zweiten Schwäche der Antwort (18b), die sich auch in der Antwort (19b) wiederfindet: Der Begriff „Satz“ wird hier für eine Sequenz verwendet, die nicht der herkömmlichen Vorstellung von einem Satz entspricht. Natürlich kann der Begriff „Satz“ unterschiedlich und gegebenenfalls minimal definiert werden, aber der Ausdruck suggeriert immerhin, dass die Fragestellerin die Sequenz so in einem herkömmlichen Satz verwenden kann, was ja in (18c) tatsächlich geschehen ist und zu einem Fehler in der Textproduktion geführt hat.

Eine ähnliche Situation stellt sich in folgenden Beispielen dar:

(22a) (Teilnehmerin O) Verlangt die Präposition „über“ immer den Dativ?

(22b) (KI-generierte Antwort) Ja, das ist im Allgemeinen der Fall. (...) Über verlangt den Dativ, wenn die Relation zum Objekt statisch ist. Wenn eine Bewegung impliziert wird, kann man den Akkusativ verwenden. (...) über dem Tisch = *sur la table*, über den Tisch = *par dessus la table*.

(22c) (Textproduktion der Fragestellerin) *Der Land Hamburg hat denken über dem Problem mit den Klimaaktivisten.

Es wird deutlich, dass der KI-generierte Erklärungsversuch auch in diesem Fall zu einer fehlerhaften Produktion führt. Die vorgeschlagene Übersetzung für das Beispiel „über dem Tisch“ ist außerdem fragwürdig. Das Programm gibt mit „auf dem Tisch“ die richtige Übersetzung für *sur la table* im nächsten Beispiel an, aber es kann dafür den Kasus nicht korrekt bestimmen:

(23a) (Teilnehmerin V) Welchen Fall verlangt die Präposition „auf“?

(23b) (KI-generierte Antwort) Die Präposition „auf“ verlangt den Akkusativ. Wenn Sie zum Beispiel „auf dem Tisch“ sagen, steht „Tisch“ im Akkusativ.

An anderer Stelle unterstützt das Programm seine Behauptung („auf verlangt den Akkusativ“) mit dem fehlerhaften Beispiel

(24) *das Buch liegt auf den Tisch (Antwort auf eine Anfrage der Teilnehmerin D)

welches in dieser Form kaum in den Texten der Datenbank zu finden sein dürfte. Wir schließen daraus, dass das Programm vermutlich zuerst in seiner Datenbank nach passenden Sequenzen sucht und bei Nichterfolg eigene Sequenzen zur Untermauerung seiner Argumentation produziert und dabei das Risiko eingeht, dass diese autoproduzierten Sequenzen nicht korrekt sind. Weitere gescheiterte Erklärungsversuche finden sich beispielsweise in

(25a) (Teilnehmerin O) *Peux tu me corriger sur la syntaxe et sur les accords des prépositions dans ces phrases (Kannst du meine Syntax und die Präpositionen in diesen Sätzen korrigieren?)* *Das ist ein Politiker problem. Die Länder müssen einen Lösung finden.

(25b) (KI-generierte Antwort) Das ist ein Politikerproblem. Die Länder müssen eine Lösung finden. Erklärung: „Einen Lösung“ wurde in „eine Lösung“ korrigiert, um dem neutralen Genus von „Problem“ zu entsprechen.

und in

(26a) (Teilnehmerin O) *Peux-tu me dire si tous mes mots sont bien allemands et les corriger en m'expliquant: (Kannst du mir sagen, ob alle meine Wörter richtiges Deutsch sind und sie korrigieren, indem du es mir erklärst?)* (...) Die Aktivisten sind in der Letzen Generation.

(26b) (KI-generierte Antwort) Die Aktivisten sind in der letzten Generation. Erklärung: Im Deutschen beginnen die Wochentage, die Monate und die Generationen mit einem Großbuchstaben.

Die vom Chatbot vorgeschlagenen Erklärungen in (25b) und (26b) haben faktisch wenig mit den (suboptimalen) Korrekturen zu tun, sie enthalten keinen didaktischen Mehrwert und erwecken eher den Eindruck, eine Erklärung mit Hilfe von Elementen zu simulieren, die zufällig ausgewählt wurden. Zuweilen ist sogar die KI-generierte Korrektur imaginär:

(27a) (Teilnehmerin M) Kannst du den Satz „Es gibt Proteste in ganz Deutschland.“ korrigieren?

(27b) (KI-generierte Antwort) Ihr Satz ist korrekt. Um ihn eleganter zu gestalten, können Sie einen unbestimmten Artikel vor „Proteste“ einfügen. „Es gibt Proteste in ganz Deutschland.“ wird zu „Es gibt Proteste in ganz Deutschland.“

Das Programm ist in dieser Situation nicht in der Lage festzustellen, dass der Ausgangssatz und die Korrektur identisch sind. Eine ähnliche Erscheinung konnte man schon in (16b) beobachten. Die Fragestellerin hat den Satz (27a) in ihrer Textproduktion natürlich nicht verändert, aber die Intervention von ChatGPT ist hier eher nutzlos.

b) Gebrauch von Konjunktionen

KI-generierte syntaktische oder semantische Fehlleistungen ohne Bezug zur Deklination sind ebenfalls ebenfalls im Korpus zu finden. Das Programm schlägt folgendes Beispiel vor, um die vermeintlich korrekte Verwendung von „ob“ zu illustrieren:

(28a) (Teilnehmerin S) *Wenn benutzen Aktivisten die Stark weil die Meinung von Gericht und Regierung über die Klimat probleme nicht verbessern. Soll ich hier wenn oder ob benutzen?

(28b) (KI-generierte Antwort) „Wenn“ wird benutzt, wenn Sie eine Bedingung ausdrücken wollen. Wenn Sie eine Frage formulieren wollen, können Sie ob verwenden: *Ob Aktivisten die Stärke einsetzen, verbessert sich die Meinung von Gericht und Regierung über die Klimaprobleme?

(28c) (finale Textproduktion der Fragestellerin) *Wenn Klimaengagierte die Stärke einsetzen weil die Meinung von Gericht und Regierung sich nicht verbessern.

Die KI-generierte Erklärung in (28b) ist unzureichend, das angegebene Beispiel entspricht keineswegs dem Sprachgebrauch und es wäre erstaunlich, diesen Satz bereits in der Datenbank des Programms zu finden. Die Sequenz (28c) zeigt, dass die unverständliche Antwort (28b) folgerichtig nicht zu einer verbesserten Textproduktion der Fragestellerin beiträgt.

Darüber hinaus kann man beobachten, dass auch korrekte Erklärungen und dazu passende Beispiele nicht zwingend zu einer akzeptablen Textproduktion führen. Das gilt insbesondere für komplexe Konstruktionen, wie „je...desto...“. Der KI-generierte Erklärungsversuch gerät hier in der Regel zu kurz und bleibt lückenhaft, wenn er

überhaupt nachvollziehbar ist, so im folgenden Dialog mit Teilnehmerin W:

(29a) Wie drückt man die Kausalität aus, wie in „*plus il y a ..., plus il y a ...*“?

(29b) Sie können Kausalität in diesem Sinne ausdrücken, indem Sie folgende Struktur benutzen: Je nachdem wie viel..., desto mehr ...

Die Studierenden produzieren dann oft fehlerhafte Sequenzen, wie im Beispiel (28c) oder in

(29c) *Je nachdem wie viel Aktionen, desto mehr Gericht.

Übersetzungs- oder Korrekturanfragen, die bereits die Elemente enthalten, die in der Textproduktion verwendet werden sollen, wären in diesen Fällen wahrscheinlich ziel-führender.

4.2.2.3. Andere Fehlleistungen

Einige andere lexikalische, semantische oder morphosyntaktische Fehlleistungen seien hier nur kurz erwähnt. Das Programm identifiziert falsch geschriebene Nomen als Adjektive, unterschlägt gelegentlich die in den Anfragen verwendeten Hilfsverben, interpretiert Infinitivgruppen als Imperative, toleriert unverständliche Passagen, die zur Korrektur vorgelegt wurden, wechselt in den Antworten manchmal spontan von der Muttersprache der fragstellenden Person ins Englische und erweckt dabei den Eindruck, dass die Antworten des Chatbots zuerst auf Englisch formuliert und dann in die Muttersprache übersetzt werden. Das Programm erklärt einer Teilnehmerin ungefragt ihren Vornamen, missversteht manchmal die in den Anfragen auftauchenden französischen Personalpronomen als Eigennamen und bestätigt einige von den Teilnehmer*innen vorgeschlagene Sequenzen, z. B.

(30) die Störung der öffentlichen Ordnung bedrohen

als korrekt, obwohl sie der beabsichtigten Aussage (in diesem Fall etwa „ein Risiko für die öffentliche Ordnung darstellen“ oder „damit drohen, die öffentliche Ordnung zu stören“) entgegenstehen.

4.2.3. Zur Interpretation der Fehler des Chatbots

Das Programm bezeichnet sich selbst als textbasiertes Sprachmodell (siehe Antwort 17b), es stellt also einen Versuch dar, einige ausgewählte Parameter der Kommunikation zwischen menschlichen Personen künstlich zu reproduzieren. Diesem Anspruch wird es gerecht, denn die Illusion des Dialogs mit einem Individuum entsteht an vielen Stellen

des für diese Studie erarbeiteten Korpus. Künstliche Intelligenz, die der natürlichen Intelligenz einer Lehrkraft nahekäme, scheint aber im Moment noch nicht vorzuliegen. Es war im Übrigen auch nicht unsere Absicht zu testen, ob und inwiefern die KI eine menschliche Lehrkraft ersetzen kann, sondern ob und inwiefern sie zu einer verbesserten Textproduktion der Testpersonen beiträgt.

Von der Aktualisierung der Datenbank in Echtzeit abgesehen, fehlt es dem Programm im Moment an der Möglichkeit, in jedem Fall richtig zu entscheiden, welche Elemente der Anfrage besonders wichtig sind oder ob die Anfrage eventuell selbst mit schwerwiegenden Fehlern behaftet ist, da es die Kommunikationsabsichten der anfragenden Person nicht in ihrer Globalität erkennt. Ähnliches gilt für die Antworten des Chatbots. Die Beschränkung auf wenige Optionen scheint der Strategie von ChatGPT zu entsprechen, den Nutzer*innen sofort verwendbare Lösungen anzubieten und sich auf diese Weise von Suchmaschinen abzuheben. Welche Option der Chatbot dann als einzige oder als optimale Antwort vorschlägt, wird vermutlich durch einen Wahrscheinlichkeitsalgorithmus definiert, der die ihm zur Verfügung stehenden Texte verarbeitet und dabei nur die kontextuellen Informationen berücksichtigt, die ihm geliefert wurden. Das Ergebnis dieser Operation entspricht nicht immer den Erwartungen und trägt nicht zwingend zu einer verbesserten Textproduktion bei. Die beobachteten Schwächen des Programms werden besonders offensichtlich, wenn Erklärungen erfragt werden, in unserem Fall also beispielsweise Erklärungen zur Deklination und zum Gebrauch von Konjunktionen. In diesen Situationen scheint das Programm unter bestimmten Bedingungen die Suche in den Texten der Datenbank aufzugeben und unter Bezugnahme auf autoproduzierte fehlerhafte Sequenzen eine eigene Argumentation zu improvisieren, die zwar die Form einer Erklärung hat, die aber nicht dem Anspruch genügt, der fragenden Person eine Erscheinung verständlicher zu machen. Dazu müsste die zu erklärende Erscheinung zuerst selbst vom Programm verstanden werden. Eine Sprachsimulation, die in einer Datenbank bestimmte Sequenzen nach dem Wahrscheinlichkeitsprinzip auswählt und nach bestimmten Regeln kombiniert oder im Extremfall selbst defekte Beispiele schafft, kann das im Moment noch nicht leisten. Allerdings ist die in dieser Situation korrekte Antwort „Ich kann diese Erscheinung nicht erklären“ im Programm offensichtlich nicht vorgesehen. Sie entspräche wohl auch nicht der Vorstellung, die die Nutzer*innen von der Dienstleistung haben, die eine KI liefern sollte.

5. Schlussbetrachtung

Unsere Studie stimmt mit anderen Forschungsergebnissen überein und bestätigt (Kertes 2022), dass beim Überarbeiten eines fremdsprachlichen Textes hauptsächlich auf lexikalische und grammatische Mittel geachtet wird. Unsere Ergebnisse zeigen, dass die Studierenden hauptsächlich lexikbezogene Anfragen formulieren. In diesem Fall beantwortet das Programm die Anfragen überwiegend richtig, wenn die fragestellende Person genügend Kontext zur Verfügung stellt. Durch morphosyntaktische Anfragen, die die zweite Kategorie der Anfragen in unserer Studie repräsentieren, ist das Programm allerdings relativ schnell überfordert. Der Chatbot antwortet regelmäßig mit nur einem Vorschlag, der oft als wenig hilfreich erscheint, aber von unkritischen Studierenden in dieser Form in der eigenen Textproduktion verwendet wird. Dies steht im Einklang mit den Forschungsarbeiten, die *patchwriting* (Howard 1992) oder die Integration der KI-generierten Antworten ohne Rücksicht auf den unmittelbaren sprachlichen Kontext beschrieben haben (Barrot, 2023). Wenn es darum geht, die grammatikalische Struktur der deutschen Sprache zu erläutern, sind die KI-generierten Erklärungen in der Regel lückenhaft oder unzutreffend und erwecken den Anschein, als Antwort auf eine entsprechende Anfrage lediglich die sprachliche Form einer Erklärung zu reproduzieren. Besonders problematisch sind die gelegentlich zu beobachtenden fehlerhaften Sequenzen, die das Programm offenbar nicht in seiner Datenbank findet, sondern selbst produziert, um seine eigene simulierte Argumentation zu stützen. Dies stimmt mit den Arbeiten von Las-Vergnas & Rio-Jouet (2023) überein, die in Dialogen mit der KI die Erfindung fiktiver Daten zur Untermauerung eines Arguments festgestellt haben.

Aus didaktischer Sicht und im Hinblick auf die Benutzung von ChatGPT im Rahmen der Überarbeitung eigener Textproduktion können wir also folgende Hinweise formulieren: Lexikalische Anfragen sind nützlich und die KI-generierten Antworten sind besonders hilfreich, wenn man genügend kontextuelle Informationen angeben kann. Im Zweifelsfall sollten die Lernenden nachfragen, ob ein Wort oder eine Wortgruppe in einem bestimmten Zusammenhang verwendet werden kann. Unsere Studie zeigt, dass kompetente Lernende eher in der Lage sind, diese Hinterfragung vorzunehmen. Diese Kompetenz kann wahrscheinlich im Unterricht gezielt aufgebaut werden. Es wäre bedauerlich, künstliche Intelligenz nur als Wörterbuch-Ersatz zu verwenden und ihre durchaus vorhandenen syntaktischen und semantischen Kapazitäten

außer Acht zu lassen. KI-generierten Erklärungen zur Morphosyntax muss allerdings mit Misstrauen begegnet werden; die Lernenden sollten die gelieferten Sequenzen in jedem Fall mit ihrem eigenen Vorwissen abgleichen und mit ihrem eigenen Sprachgefühl prüfen. Die als Antwort vorgeschlagenen Elemente müssen oft syntaktisch angepasst werden, wenn die Anfrage nicht schon die Übersetzung oder die Korrektur eines ganzen Satzes zum Inhalt hatte. Die Angabe von kontextuellen Informationen, das nützliche Nachfragen, der kritische Blick auf die Antworten und die syntaktische Einbindung von gelieferten Sequenzen können natürlich selbst Gegenstand von didaktischen Aktivitäten werden, wenn man die Benutzung von KI zur Textproduktion als Ziel definiert.

Bibliographie

- Athanassopoulos, Stavros; Manoli, Polyxeni; Gouvi, Maria; Lavidas, Konstantinos; Komis, Vassilis (2023) The Use of ChatGPT As a Learning Tool to Improve Foreign Language Writing in a Multilingual and Multicultural Classroom. *Advances in Mobile Learning Educational Research* 3 (2), 818–24. <https://doi.org/10.25082/AMLER.2023.02.009>.
- Barbier, Marie-Laure (2012) L'apprentissage bilingue de l'écrit: implications en L2 et en L1. *Le Français dans le monde. Recherches et applications*, CLE International/ Français dans le monde.
- Barrot, Jessie S. (2023) Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, 57, 100745. <https://doi.org/10.1016/j.asw.2023.100745>
- Baskara, Risang; Mukarto, Fransiscus Xaverius (2023) Exploring the Implications of ChatGPT for Language Learning in Higher Education. *Indonesian Journal of English Language Teaching and Applied Linguistics*, 7 (2), 343- 358.
- Bialystok, Ellen. (2001) *Bilingualism in Development: Language, Literacy, and Cognition* (1. Aufl.), Cambridge University Press. <https://doi.org/10.1017/CBO9780511605963>.
- Börner, Wolfgang (1989): Didaktik schriftlicher Textproduktion in der Fremdsprache. In: Antos, Gerd; Krings, Hans P. (Hrsg.) *Textproduktion. Ein interdisziplinärer Forschungsüberblick* (= Konzepte der Sprach- und Literaturwissenschaft 48). Tübingen: Niemeyer, S. 348–375.
- Bourdais, Aurélie; Guichon Nicolas (2023), *Représentations et usages du traducteur en ligne par les lycéens*. *Alsic* 23 (1) <http://journals.openedition.org/alsic/4533>; <https://doi.org/10.4000/alsic.4533>.
- Chenoweth, N. Ann; John Richard Hayes (2001) Fluency in Writing: Generating Text in L1 and L2. *Written Communication* 18 (1), 80–98. <https://doi.org/10.1177/0741088301018001004>.
- Hartmann, Daniela (2021) Künstliche Intelligenz im DaF-Unterricht? Disruptive Technologien als Herausforderung und Chance. *Informationen Deutsch als Fremdsprache* 48 (6), pp. 683–696. <https://doi.org/10.1515/infodaf-2021-0078>.

- Hayes, John Richard; Flower, Linda (1980) Identifying the Organization of Writing Processes. In L. W. Gregg; E. R. Steinberg (Hrsg.) *Cognitive Processes in Writing: An Interdisciplinary Approach*, 3-30. Hillsdale, NJ: Lawrence Erlbaum.
- Howard, Rebecca Moore (1992) A plagiarism pentimento. *Journal of Teaching Writing*, 11, 233–245.
- Imran, Muhammad; Almusharraf, Norah (2023) Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology*, 2023, 15 (4).
<https://doi.org/10.30935/cedtech/13605>
- Johnson, Mark D. (2020) Planning in L1 and L2 writing: Working memory, process, and product. *Language Teaching*, 53 (4), 433–445.
<https://doi.org/10.1017/S0261444820000191>
- Kertes, Patrícia (2022) Prozessorientierung bei der Förderung der schriftlichen Textproduktion in der Fremdsprache. In: Berényi-Nagy, T.; Kertes, P; Márki, H.; Sárvári, T. (Hrsg.) *Deutschunterricht für Ungarn* 32, 88–106.
- Kohnke, Lucas; Moorhouse, Benjamin; Zou, Di (2023) ChatGPT for Language Teaching and Learning. *RELC Journal*, 54 (2), 537–550.
<https://doi.org/10.1177/00336882231162868>
- Koltovskaia, Svetlana (2020) Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing*, 44, 100450. <https://doi.org/10.1016/j.asw.2020.100450>
- Kroll, Judith F.; Gullifer, Jason W.; McClain, Rhonda; Rossi, Eleonora; Martin, Maria Cruz (2015) Selection and control in bilingual comprehension and production. In Schwieter J. W. (Hrsg.) *The Cambridge handbook of bilingual processing* (485–507). Cambridge University Press. <https://doi.org/10.1017/CBO9781107447257.021>.
- Las-Vergnas, Olivier; Rio-Jouet, Antoine (2023) Une IA peut-elle remplacer un professeur? Un exemple de dialogue avec Bard-Google, entre hallucinations et illusions réflexives, *L'esprit d'Archimède*, 2023, 27-34.
- Lee, Sangmin-Michelle (2019) The impact of using machine translation on EFL students' writing. *Computer Assisted Language Learning*, 33 (3), 157–175.
<https://doi.org/10.1080/09588221.2018.1553186>.
- Lemke, Valerie; Hoffmann, Lea (2022) *Schreibdidaktik* (= Literaturhinweise zur Linguistik 12). Heidelberg: Universitätsverlag Winter.
- Pecorari, Diane (2023a) Generative AI: Same same but different? *Journal of Second Language Writing*, 62, 101067. <https://doi.org/10.1016/j.jslw.2023.101067>.
- Pecorari, Diane (2023b) Plagiarism and English for academic purposes: A research agenda. *Language Teaching*, 56 (3), 362–376.
<https://doi.org/10.1017/S0261444821000495>
- Ruhmann, Gabriela; Kruse, Otto (2014) Prozessorientierte Schreibdidaktik: Grundlagen, Arbeitsformen, Perspektiven. In: Dreyfürst, Stephanie; Sennewald, Nadja (Hrsg.) *Schreiben. Grundlagentexte zur Theorie, Didaktik und Beratung*. Opladen, Toronto: Verlag Barbara Budrich, S. 15–34.
- Roussel, Stéphanie; Ochoa, M. (2025 im Druck) «Peux-tu rendre mon texte plus allemand?» Écrire en langue seconde avec l'intelligence artificielle. *Alsic*.
- Sasaki, Miyuki; Hirose, Keikho (1996) Explanatory Variables for EFL Students' Expository Writing. *Language Learning*, 46 (1), 137–168.
<https://doi.org/10.1111/j.1467-1770.1996.tb00643.x>

- Schaeffer-Lacroix, Eva (2015) Impact de discussions métalinguistiques sur l'apprentissage de la production écrite en allemand, langue étrangère. *Linx* 72, 233-237. (Former à l'écrit universitaire, un terrain pour la linguistique?) <https://doi.org/10.4000/linx.1584>.
- Sethuraman, Mekhala; Radhakrishnan, Geetha (2020) Promoting Cognitive Strategies in Second Language Writing. *Eurasian Journal of Educational Research*, 20 (88), 1-17. <https://doi.org/10.14689/ejer.2020.88.5>.
- Strasser, Thomas (2020): „Künstliche Intelligenz im Sprachunterricht: Ein Überblick“. In: *Revista Lengua y Cultura* 1 (1–6). <https://doi.org/10.29057/lc.v1i2.5533>.
- Sweller, John; Roussel, Stéphanie; Tricot, André (2022) Cognitive load theory and instructional design for language learning. In J.W. Schwieter; Z. Wen (Hrsg.) *The Cambridge handbook of working memory and language*, 559-880. Cambridge University Press. <https://doi.org/10.1017/9781108955638.045>.
- Warschauer, Mark; Tseng, Waverly; Yim, Soobin; Webster, Thomas; Jacob, Sharin; Du, Qian; Tate, Tamara (2023) The affordances and contradictions of AI-generated text for writers of english as a second or foreign language. *Journal of Second Language Writing* 62, 101071. <https://doi.org/10.1016/j.jslw.2023.101071>
- Wesołowska, Dorota (2021) Zum Stellenwert der Sprachlernapplikationen am Beispiel des Vokabellernens mit „duolingo“. In: Ehrenmüller, Jürgen; Höppnerová, Věra (Hrsg.) *Experimentierräume in der DaF-Didaktik*
- Zhang, Meixiu; Gibbons, John; Li, Mimi (2021) Computer-mediated collaborative writing in L2 classrooms: A systematic review. *Journal of Second Language Writing* 54, 100854. <https://doi.org/10.1016/j.jslw.2021.100854>

Biografische Angaben

Stéphanie Roussel ist Professorin für Germanistik und Sprachwissenschaften an der Universität Bordeaux. Als Leiterin des Sprachenzentrums, Studiengangsmoderatorin des deutsch-französischen Bachelor of Law und Co-Direktorin des LACES-Labors (ein multidisziplinäres Forschungszentrum für Erziehungswissenschaften) interessiert sie sich besonders für Fremdsprachenlernen. Stéphanie Roussel unterrichtet Deutsch als Fachsprache an der Rechts- und Wirtschaftsfakultät der Universität Bordeaux. Darüber hinaus unterrichtet sie auch Fremdsprachendidaktik an Studierende mit dem Schwerpunkt Linguistik. Die Forschung von Stéphanie Roussel konzentriert sich auf Deutsch für Juristen und auf digitale Technologien für den Fremdsprachenunterricht aus der Perspektive der kognitiven Psychologie.

Ayaal Herdam ist Sprachlehrer an der Universität Bordeaux sowie wissenschaftlicher Mitarbeiter am Institut de Recherche Montesquieu (Universität Bordeaux) und am Laboratoire de Recherche sur le Langage (Universität Clermont Auvergne). Er

unterrichtet hauptsächlich Deutsch als Fachsprache in den Bereichen Jura, Politikwissenschaft, Wirtschaft und Management. Seine wissenschaftlichen Interessen betreffen zum Beispiel linguistische, soziologische und historische Aspekte des interkulturellen Kontakts.

Schlagwörter

Textproduktion, Überarbeitung, Wortschatz, Syntax, Künstliche Intelligenz, ChatGPT, Französischsprachige