Vilnius University Press

Vilnius University Press

Kalbotyra
2025 (78) 138–155

ISSN 1392-1517 eISSN 2029-8315
DOI: https://doi.org/10.15388/Kalbotyra.2025.78.6

KI als Hilfsmittel für die Formulierung von Suchanfragen in Korpora

Skaistė Volungevičienė

Lehrstuhl für Deutsche Philologie
Institut für Sprachen und Kulturen des Ostseeraums
Philologische Fakultät
Universität Vilnius
Universiteto g. 5
LT-01131 Vilnius, Litauen
E-Mail: skaiste.zabarauskaite@flf.vu.lt
ORCID iD: https://orcid.org/0009-0001-6206-1541
https://ror.org/03nadee84

Maximilian Arndt

Institut für Empirische Sprachwissenschaft
Johann-Wolfgang-Goethe-Universität Frankfurt am Main
Rostocker Straße 2
D-60629 Frankfurt am Main, Deutschland
E-Mail: maximilianarndt1998@gmail.com
https://ror.org/04cvxnb49

Abstract. Die Bedeutung von Korpusarbeit und korpuslinguistischen Methoden nimmt in der linguistischen Forschung und im DaF-Unterricht stetig zu, da sie eine umfassende und systematische Analyse sprachlicher Phänomene ermöglichen. Häufig arbeiten Forscherinnen und Forscher dabei mit einer Vielzahl von Korpora, die unterschiedliche Datensätze und Strukturen aufweisen. Eine zentrale Herausforderung besteht in der Vielfalt der jeweils eingesetzten Abfragesysteme. Diese sind oft komplex und heterogen, sodass ein erheblicher Zeitaufwand erforderlich ist, um ihre Funktionsweise zu verstehen und sie effektiv zu nutzen. Die damit verbundene Vielfalt kann die Suche nach spezifischen sprachlichen Mustern oder Phänomenen sowohl für erfahrene Nutzerinnen und Nutzer als auch für Anfängerinnen und Anfänger erschweren und frustrierend machen, da um-

Submitted: 07/11/2025. **Accepted:** 19/12/2025

fangreiche Einarbeitungszeiten in die jeweiligen Abfragesprachen notwendig sind. Angesichts dieser Herausforderungen kann die Integration von KI als Werkzeug zur Formulierung von Suchanfragen in Korpora künftig eine entscheidende Rolle spielen.

Vor diesem Hintergrund untersucht der vorliegende empirische Beitrag das Potenzial KI-gestützter Systeme zur Formulierung Suchanfragen in COSMAS II. Grundlage der Untersuchung bilden 50 typische korpuslinguistische Suchaufgaben unterschiedlicher Art, die von unerfahrenen Nutzerinnen und Nutzern ohne computerlinguistische Vorkenntnisse formuliert werden. Die KI-generierten Abfragen werden hinsichtlich ihrer formalen Korrektheit, Präzision und Funktionalität analysiert.

Methodisch basiert die Studie auf einer qualitativen Auswertung der Abfrageergebnisse. Die Ergebnisse zeigen, unter welchen Bedingungen KI als unterstützendes Werkzeug bei der Korpusabfrage eingesetzt werden kann und wo ihre Grenzen liegen. Darüber hinaus wird diskutiert, welche Informationen der KI zur Verfügung stehen müssen, um korpuslinguistische Suchprozesse effektiv zu unterstützen.

Schlüsselwörter: KI, Formulierung von Korpusabfragen, Korpora, korpuslinguistische Methoden, DaF, COSMAS II

Artificial intelligence as a support tool for corpus query formulation

Abstract. The importance of corpus work and corpus-linguistic methods is steadily increasing in linguistic research and in the field of German as a Foreign Language (GFL), as they enable a comprehensive and systematic analysis of linguistic phenomena. Researchers often work with a large number of corpora that differ in their datasets and structures. One challenge in this work, however, is the diversity of query systems used across individual corpora. These systems can be complex and heterogeneous, which often requires a considerable amount of time to understand their functionality and use them effectively. This diversity can make the search for specific linguistic patterns or phenomena difficult and frustrating for both experienced users and beginners, as a great deal of time must be spent familiarizing oneself with the particularities of individual query systems. In response to these challenges, the integration of AI as a tool for formulating search queries in corpora may play a decisive role in the future.

In this context, the present empirical contribution examines the potential of AI-supported systems for formulating complex search queries in COSMAS II. The study draws on 50 typical corpus-linguistic search tasks of different types, formulated by inexperienced users without prior knowledge of computational linguistics. The AI-generated queries are analyzed with regard to their formal correctness, precision, and functionality.

Methodologically, the study is based on a qualitative analysis of the query results. The results indicate the conditions under which AI can be used as a supportive tool in corpus querying and where its limitations lie. The contribution also discusses what information needs to be available to the AI in order to effectively support corpus-linguistic search processes.

Keywords: AI, corpus query formulation, corpora, corpus-linguistic methods, GFL, COSMAS II

1 Einleitendes

Die fortschreitende Digitalisierung prägt zunehmend alle Bereiche der Wissenschaft und verändert auch die Methoden und Fragestellungen innerhalb der Sprachwissenschaft grundlegend (McShane & Nirenburg 2021, 7). Insbesondere die computergestützte Korpuslinguistik hat sich zu einem zentralen Forschungsfeld entwickelt, da sie es ermöglicht, Sprache auf der Grundlage umfangreicher, authentischer Datensammlungen empirisch zu untersuchen. Korpora stellen damit eine essentielle Ressource für die Analyse sprachlicher Strukturen, Entwicklungen und Gebrauchsweisen dar.

Eine der größten Herausforderungen in der praktischen Arbeit mit Korpora besteht jedoch darin, adäquate Forschungsanfragen zu formulieren, um spezifische sprachliche Phänomene gezielt zu untersuchen. Das Erstellen solcher Korpusanfragen erfordert nicht nur Kenntnisse in Logik und Linguistik, sondern auch ein Verständnis der technischen Anfragesprachen, die von verschiedenen Korpusplattformen verwendet werden (Hunton 2006, 241). Diese Anforderungen erschweren insbesondere Einsteigerinnen und Einsteigern den Zugang zu den Daten und können den Lernprozess erheblich verlangsamen.

Vor diesem Hintergrund gewinnt die Frage an Bedeutung, inwiefern aktuelle Entwicklungen im Bereich der KI den Umgang mit Korpora vereinfachen und zugänglicher gestalten können. *Large Language Models* (LLMs) wie ChatGPT bieten durch ihre dialogische Struktur und ihr Verständnis natürlicher Sprache das Potenzial, niedrigschwellige Unterstützung bei der Formulierung und Interpretation von Korpusanfragen zu leisten. Ziel dieses Beitrags ist es daher, zu untersuchen, wie solche Modelle eingesetzt werden können, um Barrieren bei der Arbeit mit Korpusdaten zu verringern und den Zugang zu sprachwissenschaftlichen Ressourcen zu erleichtern.

Im Zentrum der Untersuchung steht das Korpusverwaltungssystem COSMAS II (Leibniz-Institut für Deutsche Sprache (IDS), *Corpus Search, Management and Analysis System*), das eine zentrale Rolle in der Korpusforschung im deutschsprachigen Raum einnimmt. Die Plattform bietet Zugang zu einer Vielzahl synchroner und diachroner Korpora und ermöglicht so die Analyse der deutschen Sprache in unterschiedlichen

historischen und gegenwärtigen Kontexten. Trotz dieser umfangreichen Funktionalität stellt die Komplexität der zugrunde liegenden Abfragesyntax und der begleitenden Dokumentation für viele Nutzende eine erhebliche Hürde dar. Die verfügbaren Anleitungen umfassen zahlreiche Webseiten mit detaillierten Erläuterungen und Beispielen[1], was insbesondere für unerfahrene Nutzende eine kognitive Überforderung darstellen kann, zumal die Materialien nicht immer in der Muttersprache der Anwenderinnen und Anwender verfasst sind. Der effiziente Umgang mit COSMAS II setzt daher nicht nur Zeit und Geduld, sondern auch eine ausgeprägte Fähigkeit zum logischen und strukturellen Denken voraus.

Aus diesen Beobachtungen ergibt sich ein klarer Forschungsbedarf: Es gilt zu klären, inwiefern KI-gestützte Systeme dazu beitragen können, die Nutzung von COSMAS II zu erleichtern, indem sie komplexe Strukturen verständlicher machen und die Formulierung linguistisch präziser Abfragen unterstützen.

Dieser Beitrag ist das Ergebnis eines zweimonatigen wissenschaftlichen Praktikums[2], das im Sommer 2024 vom Litauischen Wissenschaftsrat (*Lietuvos mokslo taryba*) unterstützt und an der Universität Vilnius erfolgreich durchgeführt wurde.

2 Kurzer Überblick über den aktuellen Forschungsstand

Die aktuelle Forschung betont zunehmend die Bedeutung digitaler Kompetenz als zentrale Voraussetzung für den erfolgreichen Einsatz von Künstlicher Intelligenz (KI) im Sprachunterricht und Sprachlernen. Mit der jüngsten Verbreitung benutzerfreundlicher generativer KI-Werkzeuge (z. B. ChatGPT) hat der Einsatz von KI und insbesondere Generativer KI (GenAI) im Bildungsbereich erheblich an Bedeutung gewonnen (vgl. Curry & McEnery 2025, 241). Unter digitaler Kompetenz wird dabei nicht nur die Fähigkeit verstanden, digitale Technologien zu nutzen, sondern auch, sie kritisch zu bewerten, reflektiert einzusetzen und gegebenenfalls selbst zu gestalten. Diese Kompetenz umfasst verschiedene Dimensionen, die in der Forschung als Kernbereiche beschrieben werden: das Erkennen und Verstehen digitaler Technologien, die Anwendung und Bewertung ihrer Funktionen, die ethisch verantwortungsvolle Navigation in digitalen Umgebungen sowie die Fähigkeit, selbst digitale Inhalte zu erstellen. Ein hohes Maß an digitaler Kompetenz gilt als Voraussetzung dafür, dass Lehrkräfte und Lernende KI-basierte Werkzeuge effektiv, sicher und didaktisch sinnvoll in den Unterricht integrieren können.

[1] Online Hilfe zu COSMAS II Suchanfragen ist unter https://www2.ids-mannheim.de/cosmas2/win-app/hilfe/suchanfrage/ zu finden. Man kann eines der acht möglichen Themen auswählen: grafische Eingabe, Zeileneingabe, Morph-Assistent, Auswahllisten, Wortformlisten, Statistik, Sonderzeichen und reguläre Ausdrücke.

[2] Informationen zu finanzierten Sommerpraktika sind unter https://lmt.lrv.lt/lt/veiklos-sritys/mokslo-finansavimas/karjeros-mobilumo-sklaidai-skatinimo-priemones/studentu-tyrimai-vasaros-metu/ zu finden.

Mehrere Studien zeigen, dass Lehrkräfte durch den gezielten Ausbau digitaler und technologischer Fähigkeiten ihre Unterrichtspraxis verbessern und Lernprozesse stärker personalisieren können. Dadurch lassen sich Lernumgebungen schaffen, die sowohl motivierend als auch adaptiv auf die individuellen Bedürfnisse der Lernenden eingehen (Kohnke et al. 2023; Pokrivčáková 2019). Besonders im Zusammenhang mit KI-Anwendungen eröffnet digitale Kompetenz neue Möglichkeiten der Unterrichtsgestaltung, etwa durch automatisiertes Feedback, adaptive Lernsysteme oder interaktive Kommunikationswerkzeuge. Lehrkräfte mit höherer digitaler Kompetenz sind besser in der Lage, die Funktionsweisen von KI-Werkzeugen zu verstehen und diese gezielt in sprachdidaktische Szenarien einzubinden.

Forschungsarbeiten zeigen jedoch auch, dass viele Lehrende und Lernende bislang nur über begrenzte Erfahrungen und Kenntnisse im Umgang mit KI-gestützten Anwendungen verfügen. So belegen empirische Studien, dass sowohl das Wissen über KI-Systeme als auch die praktische Anwendungskompetenz vieler Nutzerinnen und Nutzer noch unzureichend ausgeprägt sind. Trotz dieser Defizite zeigen Lehrkräfte eine grundsätzlich positive Einstellung gegenüber computer- und KI-gestütztem Lernen. Frühere Untersuchungen im Bereich des *Computer-Assisted Language Learning* (CALL) und des darauf aufbauenden *Intelligent CALL* (ICALL) belegen, dass Lehrkräfte neuen Technologien gegenüber offen sind und ihr Potenzial zur Förderung sprachlicher Lernprozesse anerkennen (Heift & Schulze 2007, 63).

3 Methodische Ansätze zur Integration von KI in die Korpusforschung

Am Anfang der vorliegenden Untersuchung stand die Hypothese, dass der Einsatz von KI-basierten Systemen wesentlich dazu beitragen kann, komplexe Abfrageprozesse zu vereinfachen, die Bedienbarkeit von Korpusplattformen zu verbessern und den Zugang zu linguistischen Daten für eine breitere Nutzendengruppe zu erleichtern.

Die im Rahmen der Untersuchung identifizierten methodischen und technischen Herausforderungen verdeutlichen zugleich, dass die Integration von LLMs in komplexe Korpusverwaltungssysteme wie COSMAS II gegenwärtig noch erhebliche Entwicklungsarbeit erfordert. Ein Vergleich mit bestehenden Lösungen aus anderen Sprachräumen zeigt jedoch, dass niedrigschwellige Zugänge grundsätzlich realisierbar sind. So entwickelten Milička und Šebestová (2024) eine Anwendung, die Korpusrecherchen im Tschechischen Nationalkorpus ermöglicht. Dieses Korpus ist im Gegensatz zu COSMAS II ohne vorherige Anmeldung zugänglich und bietet damit insbesondere für unerfahrene Nutzende einen deutlichen Vorteil. Die dokumentierten Ergebnisse und die erprobte Anwendung veranschaulichen, dass praktikable Ansätze existieren, um die Barrieren zwischen Nutzenden und Korpussystemen zu verringern (Milička & Šebestová 2024). Daraus

lässt sich ableiten, dass bei entsprechender technischer Anpassung auch für COSMAS II die Entwicklung eines benutzerfreundlicheren, KI-gestützten Zugangs denkbar ist.

In folgenden Unterkapiteln werden verschiedene mögliche Ansätze vorgestellt. Im Mittelpunkt steht die Frage, inwiefern diese Technologien den Forschungsprozess nicht nur effizienter, sondern auch inklusiver gestalten können, indem sie den Zugang für Personen ohne vertiefte technische Kenntnisse erleichtern.

Darüber hinaus werden die jeweiligen Vor- und Nachteile unterschiedlicher KI-basierter Methoden diskutiert und ihre Anwendbarkeit im akademischen Lehr- und Lernkontext kritisch bewertet. Ziel ist es, ein differenziertes Verständnis der Potenziale und Grenzen solcher Verfahren zu entwickeln.

3.1 Von der Idee zum ausgewerteten Ergebnis

Ein naheliegender Ansatz besteht darin, Forschungsanfragen in natürlicher Sprache zu formulieren und die entsprechenden Ergebnisse automatisiert durch COSMAS II ausgeben zu lassen. Eine derartige Lösung wäre ideal, da sie die Nutzenden von der Notwendigkeit entbindet, sich intensiv mit der spezifischen Korpusanfragesprache auseinanderzusetzen. So könnte beispielsweise eine Eingabe wie *Ich möchte alle Verwendungen des Suffixes -ung in Verbindung mit Negationspräfixen finden* automatisch in eine korrekte COSMAS II-Abfrage übersetzt und ausgeführt werden. Fragt man eine KI entsprechend, erhält man folgende Antwort: *Ja – ich kann dir die Suchintention in eine korrekte COSMAS-II-Abfrage übersetzen. Ausführen (also tatsächlich im DeReKo/COSMAS II laufen lassen) kann ich hier allerdings nicht, weil ich keinen direkten Zugriff auf dein COSMAS-II-Konto bzw. die Korpora habe*.

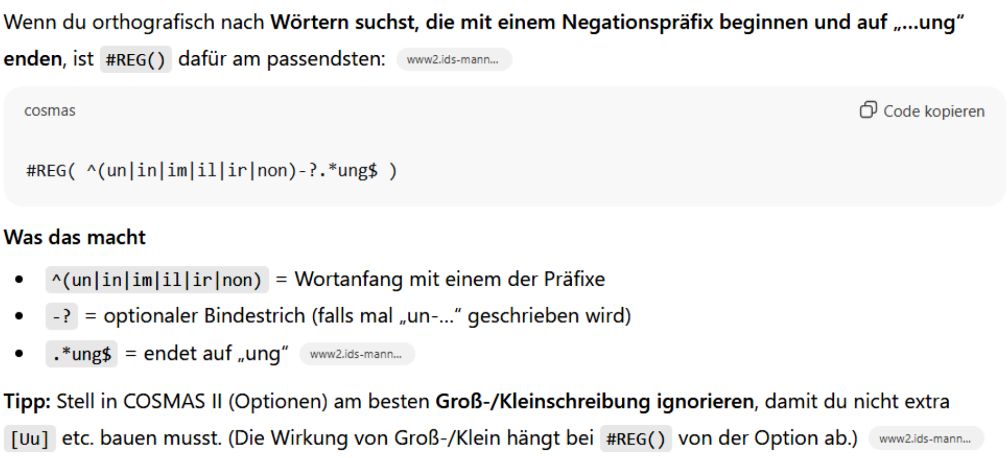

Abbildung 1. Darstellung einer von ChatGPT generierten COSMAS-II-Abfrage

Die Abbildung 1 zeigt die vom ChatGPT empfohlene Verwendung des regulären Ausdrucks *#REG(^(un|in|im|il|ir|non)-?.*ung$)* sowie eine schrittweise Erläuterung der einzelnen Bestandteile der Regex-Abfrage (Wortanfang, Präfixauswahl, optionaler Bindestrich und Suffix). Ergänzend enthält die Antwort von ChatGPT einen Hinweis zur empfohlenen Deaktivierung der Groß- und Kleinschreibung in den COSMAS-II-Optionen, um die Abfrage zu vereinfachen.

Die praktische Umsetzung dieses Ansatzes erweist sich, wie aus der Antwort des Bots hervorgeht, als mit erheblichen Herausforderungen verbunden. Für eine direkte Interaktion mit COSMAS II wird eine *Application Programming Interface* (API) benötigt. Dies erfordert nicht nur Programmierkenntnisse, etwa in *Python*, sondern auch ein vertieftes Verständnis der Funktionsweise von ChatGPT und der Entwicklung entsprechender Anwendungen (OpenAI 2024). Hinzu kommt, dass COSMAS II nur mit einem aktiven Benutzer-Login zugänglich ist, was zusätzliche Anforderungen an Authentifizierung und Zugriffsrechte stellt.

Aufgrund dieser technischen und administrativen Hürden ist dieser Ansatz zwar konzeptionell überzeugend, in der praktischen Umsetzung jedoch nicht niedrigschwellig realisierbar und somit für die Mehrheit der Studierenden kaum geeignet.

3.2 Von der Idee zum fertigen Code

Ein alternativer Ansatz besteht darin, aus einer natürlichsprachlichen Anfrage automatisch den entsprechenden Code in der COSMAS II-Anfragesprache zu generieren. Beispielsweise könnte die Eingabe *Ich möchte alle Wörter mit Negationspräfix und der Endung -ung finden* in den formalen Ausdruck *(ver oder anti* oder un* oder non* oder des* oder ent*) und ung* übersetzt werden. Dieser Ansatz hat den Vorteil, dass er die Komplexität der formalen Anfragesprache reduziert und den Nutzenden unterstützt, präzisere und effektivere Anfragen zu formulieren.

Gleichzeitig bestehen jedoch auch hier Herausforderungen. Natürlichsprachliche Eingaben sind häufig mehrdeutig oder unpräzise, was zu fehlerhaften oder unvollständigen Abfragen führen kann, insbesondere bei komplexeren linguistischen Fragestellungen mit mehreren Suchparametern oder Abhängigkeiten. Zwar liefert die kostenpflichtige Version von ChatGPT (zum Zeitpunkt der durchgeführten Untersuchung GPT-4.0) bei gezielten Anweisungen und Beispielen in der Eingabezeile deutlich bessere Ergebnisse, dennoch bleibt eine manuelle Überprüfung und gegebenenfalls Korrektur der generierten Abfragen unverzichtbar.

3.3 Präzisierung der Eingabeformulierung

Um die Wahrscheinlichkeit von Missverständnissen zu reduzieren, kann die Anfrage expliziter und strukturell klarer formuliert werden. Beispielsweise ließe sich anstelle einer freien Eingabe der Satz *Suche nach den Vorsilben ver-, anti-, un-, non-, des- und ent- in Verbindung mit der Endsilbe -ung* verwenden. Eine derart präzise Formulierung trägt zur besseren Interpretierbarkeit bei und kann die Genauigkeit der resultierenden COSMAS II-Abfragen erhöhen.

Obwohl dieser Ansatz insbesondere bei einfacheren Suchanfragen zu einer Verbesserung führt, stößt er bei komplexeren linguistischen Strukturen an seine Grenzen. Die KI kann Schwierigkeiten haben, die erforderliche logische Struktur, Operatoren und Platzhalter korrekt zu kombinieren, insbesondere wenn abstraktere linguistische Konzepte einbezogen werden. Folglich bleiben auch in diesem Fall eine ergänzende Anleitung sowie eine kritische Nachkontrolle durch die Nutzenden erforderlich.

3.4 Verwendung einer halbcodierten Eingabesprache

Ein deutlich erfolgversprechenderer Ansatz besteht in der Verwendung einer halbcodierten Sprache, die Elemente der formalen COSMAS II-Anfragesprache in vereinfachter und informeller Weise abbildet. Eine entsprechende Eingabe könnte etwa lauten *Suche nach: (Vorsilben: ver-, anti-, un-, non-, des-, ent-) in Verbindung mit (Endung: -ung)*. Durch die Nutzung von Klammern und Interpunktion werden die Beziehungen zwischen den Elementen deutlicher und Missverständnisse reduziert.

Dieser Ansatz führt in der Regel zu zuverlässigeren Ergebnissen, da die KI durch die strukturelle Vorgabe klarere Hinweise auf benötigte Operatoren und Platzhalter erhält. Dennoch bleibt ein gewisses Fehlerpotenzial bestehen, insbesondere in Bezug auf die präzise Wahl der Operatoren und deren korrekte Positionierung. Eine nachträgliche menschliche Kontrolle und gegebenenfalls Anpassung der generierten Abfragen ist daher weiterhin notwendig.

3.5 Vom Code zur erklärten Abfrage

Ein besonders lernförderlicher Ansatz besteht darin, einen funktionierenden Code in der COSMAS II-Anfragesprache als Eingabe zu verwenden und durch die KI eine erläuternde Beschreibung seiner Funktionsweise generieren zu lassen. So könnte etwa der Ausdruck *(ver oder anti* oder un* oder non* oder des* oder ent*) und ung* von der KI folgendermaßen interpretiert werden: *Diese Anfrage sucht nach allen Wörtern, die mit den Präfixen ver-, anti-, un-, non-, des- oder ent- beginnen und mit dem Suffix -ung enden.*

Diese Methode bietet erhebliche Vorteile in Bezug auf Verständlichkeit und Genauigkeit, da sie auf bereits funktionierendem Code basiert. Insbesondere fortgeschrittene Versionen von ChatGPT, wie GPT-4.0 oder höhere, ermöglichen die Generierung präziser und nachvollziehbarer Erklärungen, die das Verständnis der Anfragesprache und der Funktionsweise von COSMAS II nachhaltig fördern können.

Darüber hinaus ist dieser Ansatz für Studierende mit grundlegenden Kenntnissen der COSMAS II-Syntax und einem Verständnis für linguistische Operatoren gut umsetzbar und bietet eine praxisnahe Möglichkeit, analytische und technische Kompetenzen zu verbinden.

3.6 Zwischenfazit

Die analysierten und oben beschriebenen Ansätze zeigen, dass der Einsatz von KI in der Korpusarbeit unterschiedliche Potenziale bietet, deren Nutzen jedoch stark von der Form der Interaktion abhängt. Während vollständig automatisierte Lösungen derzeit technisch und organisatorisch kaum umsetzbar sind, erweisen sich teilautomatisierte Verfahren, insbesondere die Nutzung halbcodierter Eingaben, als praktikabler und lernfördernder.

Zudem zeigt sich, dass die KI-gestützte Erklärung bestehender Codes eine effektive Unterstützung beim Verständnis der COSMAS-II-Syntax darstellt. Insgesamt sind die größten Fortschritte dort zu erwarten, wo KI-Systeme als ergänzende Hilfsmittel eingesetzt werden, nicht als vollständiger Ersatz menschlicher Analysekompetenz.

4 Ergebnisse der KI-gestützten Korpusarbeit

Im Rahmen der durchgeführten Untersuchung wurde die Leistungsfähigkeit von ChatGPT bei der Formulierung und Umsetzung von Korpusanfragen in COSMAS II empirisch getestet. Insgesamt wurden 50 Suchanfragen entwickelt und ausgewertet, um feststellen zu können, in welchem Maße das LLM in der Lage ist, die spezifische Syntax und Logik der Korpussprache korrekt zu interpretieren und anzuwenden.

Für alle Anfragen, die morphosyntaktische Annotationen erforderten, also die Verwendung des Operators *MORPH*, wurde das TAGGED-C-Archiv der morphosyntaktisch annotierten Korpora (CONNEXOR) von COSMAS II genutzt. In allen übrigen Fällen kam das W-Archiv der geschriebenen Sprache zum Einsatz.

Die Ergebnisse der Anfragen wurden anschließend inhaltlich und formal ausgewertet und in 2 Gruppen eingeteilt, je nachdem, inwieweit die vom LLM generierten Formulierungen der COSMAS-II-Syntax entsprachen. Weiter werden diese Gruppen ausführlicher vorgestellt.

4.1 Zum methodischen Vorgehen

Wie bereits zuvor angemerkt, wurde der vorliegenden Untersuchung die Annahme zugrunde gelegt, dass die Nutzerinnen und Nutzer kaum mit den Suchoperatoren von COSMAS II vertraut sind, über wenig oder keine Erfahrung in der Formulierung entsprechender Suchanfragen verfügen und zudem keine computerlinguistischen Kenntnisse besitzen. Zum Zeitpunkt der Durchführung der Untersuchung stand die Version GPT-4.0 von ChatGPT zur Verfügung, die im Rahmen der Analyse eingesetzt wurde.

Ein erstes praktisches Hindernis ergab sich aus der Anmeldungspflicht für COSMAS II. Da entsprechend der getroffenen Annahmen keine Programmierkenntnisse vorausgesetzt werden konnten, wurde dieses Problem auf einfache Weise umgangen: Die vom Sprachmodell generierten Suchanfragen wurden manuell in das Suchfeld von COSMAS II übertragen, anschließend wurden Probesuchen durchgeführt und die resultierenden Trefferlisten ausgewertet.

Wie bereits im Einleitungskapitel angedeutet, existieren zahlreiche Anleitungen mit detaillierten Erläuterungen und Beispielen zur Formulierung von COSMAS-II-Suchanfragen. Die detaillierte Einarbeitung in diese Materialien ist für Nutzerinnen und Nutzer, insbesondere ohne einschlägige Vorerfahrung, jedoch nur eingeschränkt realistisch, da sie mit einem beträchtlichen Zeitaufwand verbunden ist. Zudem ist die gezielte Nutzung der Hilfefunktionen für Personen ohne entsprechende Vorkenntnisse nur begrenzt praktikabel. Vor diesem Hintergrund wurde entschieden, das Sprachmodell schrittweise durch gezielte Prompts an die COSMAS-II-Syntax heranzuführen. Zu berücksichtigen ist dabei, dass GPT-4.0 zum Untersuchungszeitpunkt noch nicht in der Lage war, selbstständig automatisch auf frei zugängliche Online-Anleitungen zuzugreifen. Zum Zeitpunkt der Veröffentlichung dieses Beitrags ist dies mit der Version GPT-5.2 hingegen möglich, was sich unter anderem an den Verweisen auf die offizielle COSMAS-II-Hilfeseite am Ende einzelner Informationsabschnitte erkennen lässt (vgl. Abb. 1 in diesem Beitrag).

Da von keinerlei Vorerfahrung der Nutzerinnen und Nutzer mit den Suchoperatoren von COSMAS II ausgegangen wurde, wurde aus den in Kapitel 3 vorgestellten Ansätzen jener gewählt, der auf eine Interaktion mit dem Sprachmodell in natürlicher Sprache setzt. Zeigten sich die generierten COSMAS-II-Abfragen als unzureichend, wurden die natürlichsprachlichen Eingaben schrittweise modifiziert, beispielsweise durch die Verwendung von Operatoren, die aus der Schulmathematik oder der elementaren Logik bekannt sind. Darüber hinaus wurde das Sprachmodell in den Prompts aufgefordert, in natürlicher Sprache zu erläutern, welche Suchoperation durch den jeweils erzeugten Code im Korpus ausgeführt wird. Auf diese Weise konnte überprüft werden, ob die zugrunde liegende Suchintention korrekt verstanden und umgesetzt worden war.

In mehreren Fällen führten auch wiederholte Präzisierungen sowie Hinweise auf Fehlinterpretationen nicht zu zufriedenstellenden Ergebnissen. In solchen Situationen war es erforderlich, selbstständig in den COSMAS-II-Anleitungen nach geeigneten Suchoperatoren zu recherchieren und diese dem Sprachmodell als explizite Vorgabe bereitzustellen.

Die folgenden Unterkapitel 4.2 und 4.3 widmen sich einer detaillierten Beschreibung der im Rahmen dieses Vorgehens erzielten Ergebnisse. Zunächst werden Beispiele erfolgreicher Umsetzungen der COSMAS-II-Syntax durch das LLM vorgestellt. Anschließend werden fehlerhafte bzw. unzureichende Implementierungen beschrieben, um die Grenzen und Herausforderungen beim Einsatz LLMs zur Generierung komplexer Korpusabfragen festzustellen.

4.2 Erfolgreiche Umsetzungen der COSMAS-II-Syntax durch das LLM

Von den insgesamt 50 generierten Suchanfragen ließ sich eine Gruppe identifizieren, bei der das LLM die Syntax, Operatoren und semantischen Strukturen der COSMAS-II-Anfragesprache korrekt oder nahezu korrekt umsetzte. In diesen Fällen waren die Ergebnisse nachvollziehbar, logisch aufgebaut und führten zu sinnvollen linguistischen Analysen. Typisch für diese Gruppe war die korrekte Anwendung der Operatoren */+w*, */s0* und *MORPH*, das richtige Erkennen von Platzhaltern sowie die Fähigkeit, Kontextanalysen und Vergleiche zwischen Wortformen, Regionen und Zeiträumen vorzunehmen.

Die Anfrage *&Zeitenwende (Wie hat sich der Gebrauch des Wortes Zeitenwende nach 2022 verändert?)* wurde weitgehend korrekt umgesetzt. Das Modell erkannte, dass das Präfix *&* alle Wortformen abruft, interpretierte den Ausdruck *nach 2022* zunächst lokal statt temporal, konnte aber nach weiteren Hinweisen in natürlicher Sprache die Sortierung nach Jahr und relativer Häufigkeit vorschlagen.

Auch bei *lizensier* (Finde alle Formen von lizensieren und vergleiche sie mit lizenzieren)* identifizierte das LLM den richtigen Platzhalteroperator und empfahl die Durchführung einer zweiten Vergleichssuche, um Häufigkeiten gegenüberzustellen.

Für die Anfragen *Velo* und *Fahrrad* lieferte das Modell korrekte Hinweise auf regionale Filterung und Berechnung relativer Häufigkeiten pro eine Million Wörter. In Kombination konnte es Synonyme vergleichen, benötigte jedoch für die Berechnung eine Beispielanleitung.

Bei der Anfrage *Fernsprech* (Wie kann der Gebrauch von Fernsprecher, Fernsprechanlage usw. im Laufe der Zeit untersucht werden?)* schlug das Modell zunächst eine falsche Syntax zur zeitlichen Einschränkung vor, korrigierte diese aber nach Hinweisen.

Eindeutig erfolgreich verlief die Anfrage *ich /s0 aß*, mit der Satzkollokationen gesucht wurden. Die Syntax */s0* wurde korrekt angewendet, um die beiden Wörter im selben Satz zu finden.

Ebenso zeigte das Modell bei *MORPH(A){2} (Suche eine Sequenz von zwei Adjektiven)* ein gutes Verständnis der morphologischen Operatoren und gab sowohl *MORPH(A) /+w1 MORPH(A)* als auch die vereinfachte Form *MORPH(A){2}* korrekt aus.

Auch Abstands- und Ignorierungsoperatoren wurden erfolgreich angewandt: Die Anfragen *weil /+w3:20,s0 MORPH(V)*, *$Fake /+w1 $News* und *unweit /p0 &liegen* wurden vollständig korrekt umgesetzt. Besonders bei der Letzteren zeigte das LLM, dass es den Abschnittsoperator */p0* richtig interpretieren und zusätzliche Hinweise zur Sortierung nach Jahrzehnten geben konnte.

Komplexere Abfragen mit logischen Operatoren wie *Propagand**, *Querdenker nicht Corona* und **tion* und *ung** wurden ebenfalls mit hoher Genauigkeit umgesetzt. Nach kleineren Korrekturen konnte das Modell den richtigen Negationsoperator sowie kombinierte Bedingungen korrekt darstellen.

Auch im Bereich der Mustererkennung zeigten sich solide Ergebnisse. So lieferte das LLM für *+?ente+ (Suche nach Wörtern mit der Zeichenfolge ente)* nach präzisierenden Hinweisen die richtige Lösung *??ente?*, sowie für *Merkel** die nahezu korrekte Abfrage zur Suche nach Wörtern, die die Buchstabenfolge *Merkel* enthalten.

Anfragen, die auf semantische oder pragmatische Bezüge abzielten, wurden ebenfalls erfolgreich verarbeitet. So etwa *Wie kann ich untersuchen, in welchem Kontext Bier steht?*, bei der das LLM nach vorherigen Beispielen die Muster korrekt rekonstruierte, oder *Maus nicht Computer**, wo es die Negation nach einer Korrektur richtig interpretierte.

Auch Suchanfragen mit numerischen oder regulären Ausdrücken wurden weitgehend korrekt umgesetzt. Bei *viktorianisch* /s0 (#REG(18[0-9]{2}) | #REG(19[0-1][0-9]|1920))* wurde eine funktionierende Lösung generiert, die das Ziel (die Kombination von *viktorianisch* mit Jahreszahlen zwischen 1800 und 1920) erreichte.

Schließlich konnte das LLM bei Anfragen mit logischen Kombinationen mehrere Begriffe oder Phrasen korrekt verknüpfen. Beispiele hierfür sind *(Konrad /+w1 Adenauer) … (Olaf /+w1 Scholz) (Suche nach allen Bundeskanzlern)* sowie *(zum /+w1 Beispiel) oder (z. /+w1 B.) oder (bspw.)*, die nach Korrekturhinweisen vollständig korrekte Ausgaben lieferten. Auch Distanzsuchen wie *weit /w4:10 weg* oder Wortkombinationen wie *Haus* /w0 *Schuh* wurden fehlerfrei erkannt.

Diese Gruppe erfolgreicher Anfragen verdeutlicht, dass das LLM in der Lage ist, bei klar definierten, formal eindeutigen Aufgaben die COSMAS-II-Syntax präzise umzusetzen. Insbesondere bei linearen Operatorstrukturen, morphologischen Suchbefehlen und einfachen logischen Kombinationen erzielt das Modell konsistente Ergebnisse. Schwierigkeiten traten vor allem bei mehrdeutigen oder zu allgemein formulierten Eingaben sowie bei numerischen Bedingungen auf. Die Ergebnisse zeigen jedoch deutlich, dass LLMs in der Lage sind, strukturierte Abfragesprachen partiell korrekt zu reproduzieren, sofern sie durch geeignete Prompts und klare Instruktionen gesteuert werden.

4.3 Fehlerhafte oder unzureichende Umsetzungen der COSMAS-II-Syntax

Diese Gruppe umfasst alle Anfragen, bei denen das LLM deutliche Schwierigkeiten zeigte, die Syntax der COSMAS-II-Anfragesprache korrekt zu interpretieren oder anzuwenden. Fehler traten vor allem bei der Verwendung von Operatoren, Platzhaltern und logischen Verknüpfungen auf. Auch das Verständnis von Zeichenfolgen, Reihenfolgen und regulären Ausdrücken erwies sich in vielen Fällen als unzureichend. Häufig waren wiederholte Korrekturen und erläuternde Beispiele erforderlich, ohne dass eine formal korrekte Anfrage erzielt werden konnte.

Die Anfrage *richtig /+w1 ‚und' /+w1 wichtig (Suche nach der Wortfolge richtig und wichtig)* zeigte grundlegende Probleme im Umgang mit Anführungszeichen und Operatoren. Das Modell behandelte *und* als logischen Operator anstelle eines festen Bestandteils der Wortfolge. Auch die Notwendigkeit, die Reihenfolge mittels */+w1* festzulegen, wurde zunächst nicht erkannt.

Eine komplexere Variante *(richtig /+w1 ‚und' /+w1 wichtig) /p0 Antisemitismus* führte zu mehrfach redundanten und syntaktisch falschen Ergebnissen. Das Modell gab u. a. *Antisemitismus /p0 ‚richtig und wichtig'* aus, ohne die überflüssige Wiederholung zu vermeiden.

Bei der Anfrage *Standart nicht Standarte***, die den falschen Gebrauch von *Standard* untersuchen sollte, wurde der Ignorierungsoperator *$* fälschlich als Negationsoperator interpretiert. Anschließend wurden auch die Zeichen *%*, - und # irrtümlich verwendet, bevor nach wiederholtem Hinweis der korrekte Operator *nicht* erkannt wurde.

Auch bei der Mustererkennung zeigten sich Defizite. So interpretierte das LLM in *Tele?+on* die Platzhalterzeichen *?* und + fehlerhaft, indem es mehr Zeichen als erlaubt zwischen *Tele-* und *-on* zuließ. Erst nach wiederholter Erklärung konnte das Modell *Telefon* korrekt als mögliches Ergebnis erkennen.

Bei *aß oder (habe /s0 gegessen)* und der erweiterten Anfrage *(ich /s0 aß) oder (habe /s0 ich /s0 gegessen) oder (ich /s0 habe /s0 gegessen)* traten wiederholte Fehler im Umgang

mit Platzhaltern und logischen Verknüpfungen auf. Die Formulierungen wurden zwar syntaktisch korrekt, aber semantisch redundant umgesetzt, da zwischen den Varianten *ich habe gegessen* und *habe ich gegessen* kein Unterschied besteht.

Die Anfragen zum Kasusgebrauch, etwa *wegen /+w1 (mir dir dem einem) (Dativ)* und *(wegen /+w1 (des eines)) oder meinetwegen oder deinetwegen (Genitiv)*, zeigten, dass das Modell Schwierigkeiten hatte, grammatische Rollen korrekt zu erfassen. Objektpronomen wurden irrtümlich als Genitivformen interpretiert, und die Klammersetzung blieb fehlerhaft.

Auch im Bereich der Nebensatzanalyse traten Fehler auf. Die Anfrage *weil /+w2,s0 MORPH(V)* sollte Verbletztstellung in Nebensätzen identifizieren, wurde jedoch nur als Suche nach *weil* und einem Verb im selben Satz mit bis zu zwei Wörtern Abstand interpretiert. Eine logische Differenzierung zwischen Verbendstellung und Verbzweitstellung war nicht möglich. Ähnliche Probleme zeigten sich bei *weil /+s0 MORPH(V) %+w1,s0 **, wo zwar die Bestandteile korrekt erklärt, jedoch keine funktionierende Gesamtsyntax generiert wurde.

Auch bei semantischen Kontextanfragen wie *&frisch (Untersuchung des Kontexts von frisch)* blieb das Verständnis oberflächlich: Das Modell schlug lediglich *frisch /+w1** vor, ohne die Möglichkeit einer Kontextanalyse in den Ergebnissen zu erkennen.

Bei *er*in* zeigte sich ein partiell korrektes Verständnis der Escape-Sequenz, jedoch Schwierigkeiten bei der praktischen Umsetzung. Erst nach Hinweisen konnte das Modell passende Beispiele wie *Arbeiterin** und *Arbeiterinnen** nennen.

In der Anfrage *Falsch nicht &falsch* (Suche nach Komposita mit Falsch- unter Ausschluss des Adjektivs falsch)* wurde erneut der falsche Operator *%* verwendet. Erst nach dem Hinweis, dass *nicht* der korrekte Negationsoperator ist, konnte eine korrekte Version erstellt werden.

Ein weiteres Beispiel für semantische Fehlinterpretationen stellt die Anfrage *treu #OV Be*ung* dar. Das Modell nannte *Betrug* oder *Belehrung* als wahrscheinlichste Ergebnisse und konnte den eigentlichen Zusammenhang, etwa *Betreuung*, auch nach mehreren Versuchen nicht erkennen. Das Verständnis von Zeichenfolgen innerhalb eines Wortes erwies sich somit als stark begrenzt.

Auch bei *ber?chtigen (Suche nach berichtigen und berechtigen)* nutzte das Modell zunächst reguläre Ausdrücke (#REG), statt die geforderten Platzhalter. Erst nach mehreren Korrekturen wurde *ber?chtigen* als korrekte Lösung ausgegeben. Ein ähnliches Problem trat bei *B?ot* auf, wo das LLM *Boot* und *Brot* nicht zuverlässig erkannte.

Bei der Anfrage *(wo /w0 <sa>) /+s0 (? /w0 <se>)*, die Sätze mit *wo* am Anfang und einem Fragezeichen am Ende finden sollte, gelang erst nach mehrfacher Erklärung eine annähernd korrekte Formulierung. Das Modell verwechselte den Platzhalter *?* mit einem Operator und erkannte erst nach dem Hinweis auf die Notwendigkeit eines Backslashs *(\?)* die richtige Syntax.

Schließlich erwies sich die Kombination von Operatoren als besonders schwierig. Die Anfragen *raten #IN ver*ten* und *ente #IN st*en*, die nach eingebetteten Zeichenfolgen in zusammengesetzten Wörtern suchten, führten zu keinen sinnvollen Ergebnissen. Das Modell konnte die Funktionsweise des Operators *#IN* nicht korrekt umsetzen und keine passenden Wortbeispiele liefern.

Diese Gruppe zeigt deutlich die Grenzen der aktuellen Leistungsfähigkeit von LLMs im Umgang mit formalen Korpusanfragesprachen. Die meisten Fehler resultierten aus der Übertragung natürlichsprachlicher Logik auf regelbasierte Systeme. Schwierigkeiten traten insbesondere bei der korrekten Verwendung von Negations- und Kombinationsoperatoren, der Platzhalterlogik sowie bei verschachtelten Strukturen auf. Auch die semantische Interpretation komplexer Zeichenfolgen blieb oft unzureichend.

Die Ergebnisse verdeutlichen, dass LLMs zwar in der Lage sind, einfache Strukturen korrekt zu reproduzieren, bei formal strengen Abfragesprachen jedoch schnell an systemische Grenzen stoßen. Verbesserungen könnten durch den Einsatz spezialisierter Trainingsdatensätze, die gezielt auf linguistische Abfragesprachen zugeschnitten sind, sowie durch die Kombination mit regelbasierten Parsern erzielt werden. Erst durch solche hybriden Ansätze ließe sich die Zuverlässigkeit und Präzision im Umgang mit Korpusanfragen signifikant erhöhen.

4.4 Konzeptionelle und technische Herausforderungen der Untersuchung

Die Auswahl des Korpusverwaltungssystems COSMAS II und des *Large Language Models* ChatGPT ergab sich nach einer Sichtung der verfügbaren Plattformen als naheliegend. Bereits zu Beginn des Forschungsprojekts zeigte sich, dass diese Kombination sowohl aufgrund ihrer wissenschaftlichen Relevanz als auch der mit ihr verbundenen technischen und methodischen Herausforderungen besondere Eignung für die Untersuchung aufweist. COSMAS II, eine der zentralen Plattformen für die Korpusforschung im deutschsprachigen Raum, bietet umfangreiche sprachliche Ressourcen, die für die vorliegende Analyse von zentraler Bedeutung sind. ChatGPT wiederum stellt eines der bekanntesten und am weitesten verbreiteten *Large Language Models* dar und gilt damit als ein naheliegendes Werkzeug für die Erforschung KI-gestützter Ansätze in der Korpusarbeit.

Während die Auswahl der Untersuchungsobjekte vergleichsweise unkompliziert erfolgte, erwies sich die Entwicklung eines geeigneten methodischen Ansatzes als komplexe Aufgabe. Zentrale Herausforderung war die Konzeption einer Schnittstelle zwischen beiden Technologien, die eine niedrigschwellige Nutzung ermöglichen sollte. Von besonderer Bedeutung war hierbei die Frage, inwiefern Studierende ohne vertiefte technische Kenntnisse in die Lage versetzt werden können, präzise und relevante Korpusanfragen zu formulieren. In der praktischen Umsetzung zeigte sich jedoch, dass die Realisierung dieses Ziels mit erheblichen technischen und konzeptionellen Schwierigkeiten verbunden ist. Um die KI-gestützte Unterstützung valide gestalten zu können, war es notwendig, das Modell mit domänenspezifischem Wissen über COSMAS II zu versorgen und damit zu trainieren. Dieses Training setzt wiederum ein detailliertes Verständnis der Funktionsweise und der umfangreichen Operatoren von COSMAS II voraus.

Die Verwendung von LLMs wie ChatGPT zur Unterstützung in der Korpusforschung, insbesondere bei der Formulierung und Interpretation von Anfragen in COSMAS II, besitzt grundsätzlich das Potenzial, den Zugang zu komplexen sprachlichen Daten erheblich zu erleichtern. In der praktischen Anwendung traten jedoch mehrere Schwierigkeiten auf, die sich sowohl aus der spezifischen Struktur der Korpusanfragesprache als auch aus den Eigenschaften des LLM selbst ergaben. Die Ursachen hierfür liegen in der Diskrepanz zwischen der natürlichsprachlichen Trainingsbasis der Modelle und den formal-logischen Anforderungen linguistischer Korpusplattformen. Zukünftige Lösungsansätze könnten in der Entwicklung spezialisierter Trainingsdatensätze, der Integration regelbasierter Module oder der Kopplung von LLMs mit domänenspezifischen Assistenzsystemen liegen. Trotz dieser Einschränkungen bleibt das Potenzial von KI-basierten Systemen bestehen, insbesondere in der Funktion als Vermittlungs- oder Übersetzungsschicht zwischen natürlicher Sprache und Korpusanfragesprache, um so den Zugang zu sprachwissenschaftlichen Ressourcen langfristig zu erleichtern.

5 Zusammenfassung

Im Rahmen der Untersuchung wurde das Potenzial von KI, insbesondere des *Large Language Models* ChatGPT, zur Unterstützung bei der Formulierung und Durchführung von Korpusanfragen im Korpusverwaltungssystem COSMAS II analysiert. Ziel war es, zu ermitteln, inwieweit der Einsatz von LLMs dazu beitragen kann, die Zugänglichkeit und Effizienz der Korpusnutzung insbesondere für Studierende und Nutzerinnen und Nutzer ohne vertiefte Kenntnisse der Korpusanfragesprachen zu verbessern.

Die Analyse der Interaktionen zwischen dem LLM und der COSMAS II-Anfragesprache zeigte sowohl vielversprechende Ansätze als auch deutliche Einschränkungen. LLMs besitzen grundsätzlich das Potenzial, als niedrigschwellige Unterstützung bei der Kor-

pusarbeit zu fungieren. Durch ihre Fähigkeit, natürliche Sprache zu verarbeiten, können sie Benutzende bei der Formulierung von Abfragen unterstützen und somit den Zugang zu Korpora erleichtern. Auf diese Weise lassen sich Einstiegshürden in die Korpusforschung verringern und Lernprozesse in der sprachwissenschaftlichen Ausbildung effizienter gestalten.

Gleichzeitig wurde deutlich, dass die Arbeit mit der spezifischen Syntax von COSMAS II erhebliche Schwierigkeiten mit sich bringt. Das LLM hatte häufig Probleme, formale Operatoren und Zeichen korrekt zu interpretieren, wodurch fehlerhafte oder unvollständige Abfragen entstanden. Diese Beobachtungen verdeutlichen, dass große Sprachmodelle zwar über ausgeprägte Fähigkeiten in der Verarbeitung natürlicher Sprache verfügen, jedoch bei der Handhabung komplexer, formal definierter Syntaxstrukturen an ihre Grenzen stoßen. Eine präzise Anleitung und eine klar strukturierte Benutzerführung bleiben daher unerlässlich, um Missverständnisse und Fehlinterpretationen zu vermeiden.

Darüber hinaus zeigte sich, dass die Qualität der Ergebnisse stark von der Formulierung der Anfragen abhängt. Während bei einfacheren, eindeutig strukturierten Eingaben zufriedenstellende Ergebnisse erzielt wurden, führten komplexere Abfragen mit mehreren Bedingungen, Operatoren oder linguistischen Spezifikationen häufig zu Fehlinterpretationen.

Insgesamt lässt sich festhalten, dass LLMs wie ChatGPT ein hohes Potenzial als unterstützende Werkzeuge in der Korpusforschung besitzen. Sie können dazu beitragen, den Zugang zu sprachwissenschaftlichen Ressourcen zu erleichtern und den Lernprozess insbesondere für Einsteigerinnen und Einsteiger zu unterstützen. Um dieses Potenzial vollständig auszuschöpfen, sind jedoch gezielte Weiterentwicklungen erforderlich, insbesondere durch die Integration fachspezifischer Trainingsdaten, die Anpassung an formale Anfragesprachen sowie die Entwicklung benutzerfreundlicher Schnittstellen. Eine solche Verbindung von linguistischer Präzision und technischer Zugänglichkeit könnte langfristig entscheidend dazu beitragen, die Nutzung von Korpora im Forschungs- und Lehrkontext zu optimieren.

Author contributions

Skaistė Volungevičienė: conceptualization, methodology, formal analysis, writing – original draft, review & editing. **Maximilian Arndt:** methodology, formal analysis, investigation, writing – original draft, writing – review & editing.

Quellenverzeichnis

COSMAS. 1991–2024. *COSMAS I/II – Corpus Search, Management and Analysis System*. Mannheim: IDS. Available at: https://www.ids-mannheim.de/cosmas2/.

Literaturverzeichnis

Curry, Niall & Tony McEnery. 2025. Corpus linguistics for language teaching and learning: A research agenda. *Language Teaching* 58 (2), 232–251. https://doi.org/10.1017/S0261444824000430

Heift, Trude & Mathias Schulze. 2007. *Errors and intelligence in computer-assisted language learning: Parsers and pedagogues*. New York: Routledge.

Hunton, Susan. 2006. Corpus Linguistics. *Encyclopedia of Language & Linguistics*. Keith Brown, ed. 2. Ausgabe. Elsevier Science. 234–248.

Kohnke, Lucas & Benjamin Luke Moorhouse & Di Zou. 2023. ChatGPT for language teaching and learning. *RELC Journal* 54 (2), 537–550. https://doi.org/10.1177/00336882231162868

McShane, Marjorie & Sergej Nirenburg. 2021. *Linguistics for the Age of AI*. Cambridge: The MIT Press. https://doi.org/10.7551/mitpress/13618.001.0001

Milička, Jiří & Denisa Šebestová. 2024. Query a corpus in near-natural language: A human-friendly corpus query language not only for linguists. *Crossing Boundaries through Corpora: Innovative Approaches to Corpus Linguistics*. Sara Buschfeld, Patricia Ronan, Theresa Neumaier, Andreas Weilinghoff & Lisa Westermayer, eds. Amsterdam: John Benjamins. 248–262. https://doi.org/10.1075/scl.119.10mil

OpenAI. 2024. *GPTs vs Assistants*. Available at: https://help.openai.com/en/articles/8673914-gpts-vs-assistants. Accessed: 1 September 2024.

Pokrivčáková, Silvia. 2019. Preparing teachers for the application of AI-powered technologies in foreign language education. *Journal of Language and Cultural Education* 7 (3), 135–153. https://doi.org/10.2478/jolace-2019-0025