

KI-generierte Wörterbuchartikel bewerten. Ein Beitrag zur Methodik der Wörterbuchkritik

Abstract: In this paper, we report on our development of a multi-stage analysis framework that allows us to assess (AI-generated) lexicographic texts on both a quantitative and qualitative level and compare them with human-written texts. We approach this problem through a systematic and fine-grained evaluation, based on a writing experiment in which human subjects create dictionary articles with the help of ChatGPT. The stages of our framework concern the assessment of individual entries, a comparison with existing dictionary entries written by experts, an analysis of the writing experiment, and the discussion of AI-specific aspects. For the first stage, we propose an elaborate evaluation grid that enables a fine-grained comparison of dictionary entries. While this grid has been developed in the context of a specific writing experiment, it can be adapted by metalexicographical experts for the evaluation of all kinds of dictionary entries and all kinds of dictionary information categories.

Keywords: dialogbasierte KI, automatische Textgenerierung, Wörterbuchkritik, Evaluationsmethoden, semantische Analyse, allgemeinsprachige Wörterbücher

Keywords: conversational AI, natural language generation, dictionary criticism, evaluation methodology, semantic analysis, standard dictionaries

1. Einleitung: LLMs und Wörterbuchkritik

Seit OpenAI im November 2022 mit ChatGPT der Öffentlichkeit eine leicht zugängliche und niederschwellig bedienbare, dialogbasierte KI zur Verfügung gestellt hat, ist KI in aller Munde. Textgenerierende KI auf Basis großer LLMs ist aus dem Alltag längst nicht mehr wegzudenken; der kommunikative Veränderungsprozess betrifft sämtliche Bereiche, in denen mit Texten produktiv und rezeptiv gearbeitet wird. Entsprechend ist auch die Lexikographie betroffen, und mehrere Pilotstudien widmen sich bereits den Einsatzmöglichkeiten von ChatGPT in der Lexikographie. Zusammenfassend kommt de Schryver (2023: 377) zu dem Fazit:

We have seen that, with the right prompts, an LLM like ChatGPT can already be brought in to either compile a dictionary on its own, or, somewhat more safely, to speed up dictionary compilation by providing quality draft material which human lexicographers then assess and improve upon.

Noch wird aus aktuellen Experimenten meist die Schlussfolgerung gezogen, dass ChatGPT infolge seiner abweichenden Funktionalität die professionelle lexikographische Arbeit derzeit nicht ersetzen kann (z.B. Arias-Arias et al. 2024). Und dennoch sind die Ergebnisse (freilich sprachenabhängig) mindestens auf den ersten Blick oft bereits verblüffend gut, selbst wenn man die Vorgängerversion ChatGPT 3.5 zugrunde legt. Dies konnte in einer eigenen kleinen

Untersuchung bestätigt werden: Laien ohne lexikographische Vorkenntnisse waren nicht in der Lage, zuverlässig zu entscheiden, ob ein Wörterbuchartikel (WbA) in einem deutschen allgemeinsprachigen Wörterbuch von einem Menschen oder von einer menschengeleiteten Maschine erstellt wurde. Anders sieht es aus, wenn diesbezüglich lexikographische Experten befragt werden: Ihnen gelingt die korrekte Identifizierung zuverlässiger, da die KI-generierten Wörterbuchartikel eben doch (noch) gewisse Schwächen beinhalten. Ähnlich kommt Lew zu einem differenzierten Urteil, wenn er ChatGPT zwar zugesteht, dass es qualitativ hochwertige Wörterbuchartikel zu generieren vermag: Bedeutungsangaben seien „indistinguishable in quality from those written by highly trained human lexicographers“ (Lew 2023: 8). Bezuglich der Beispielangaben jedoch wird ein niedrigeres Qualitätsniveau konstatiert: „The quality of original examples generated by ChatGPT turned out not to be as impressive, and significantly worse than those crafted by professional human lexicographers“ (Lew 2023: 9).

Dies wirft weitere Fragen auf und zeigt, mit welch differenzierterem Blick die Auswirkungen der KI-Revolution auf Wörterbuchartikel bzw. auf Wörterbücher als Produkte und ebenso auf ihre Planung, Benutzung sowie kritische Beurteilung metalexikographisch (vgl. Wiegand 1998: 79–80) reflektiert werden müssen. Nach wie vor ist viel Aufklärung nötig über das, was LLMs sind und was sie können (vgl. z.B. Müller & Fürstenberg 2023). In diesem Beitrag fokussieren wir uns im Anschluss an de Schryver (2023: 377) darauf, dass es aktuell (noch) unabdingbar ist, dass mit KI generierte Wörterbuchartikel von Menschen überprüft und gegebenenfalls verbessert werden müssen, entweder mittels einer Anpassung des Promptings oder auch manuell. Entscheidend ist dabei, dass der textuelle Output der KI einer differenzierten Überprüfung unterzogen werden muss, und hier ergibt sich ein direkter Anschluss an eine bekannte Frage der Wörterbuchkritik (vgl. Bielińska & Schierholz 2017: 6): Wie können Wörterbuchartikel, seien es menschengeschriebene oder mit maschineller Unterstützung generierte, am besten auf ihre Leistungsfähigkeit hin beurteilt werden?

Die Wörterbuchkritik hat hierzu bereits diverse Konzepte vorgelegt. Svensén (2009: 482–485) beschreibt verschiedene methodische Verfahren, wie welche Daten für eine Wörterbuchevaluation gesammelt und ausgewertet werden können: 1. Analyse des Wörterbuchs durch einen oder mehrere Rezessenten, 2. Wörterbuchvergleich, 3. Benutzung von Wörterbüchern durch Probanden mit sich anschließender Befragung, 4. speziell designete Benutzungssituationen von Wörterbüchern bei einzelnen Benutzergruppen wie etwa Lernern einer Sprache. Bei Tarp (2017) (aber auch bei Engelberg & Lemnitzer 2009: 190–222; Nielsen 2009; Kemmer 2010; Parsons & Nichols 2013) werden dazu verschiedene Kriterien genannt, die sich grundlegend auf den funktionellen Wörterbuchtyp und die Adressaten des Wörterbuchs beziehen. Weitere bewertungsrelevante Kategorien beziehen u.a. die Quantität und Qualität von lexikographischen Angaben sowie die Präsentation dieser und das Design mit ein (vgl. Tarp 2017: 127). Dass zunehmend qualitative Aspekte in die Bewertung einfließen, ist wünschenswert, auch wenn diese schwieriger zu operationalisieren sind als quantitative Kriterien.

Um die verschiedenen erarbeiteten Bewertungskategorien anwendungsbezogen zu modellieren, sind konkrete Bewertungsraster notwendig. Hier dominieren in der Wörterbuchkritikforschung bislang Stufenmodelle, die eine skalare Bewertung des jeweiligen Kriteriums ermöglichen. Ein solches Bewertungsraster entwickelte beispielsweise Ripfel (1989: 57): Sie erarbeitete auf Basis einer Auswertung von 736 Wörterbuchrezensionen Kriterien (z.B. Qualität der Bedeutungserläuterungen), die fünfstufig skalar bewertet werden

sollen (von „sehr verständlich / einfach“ bis „unverständlich / umständlich“). Auch Pearson & Nichols (2013), aufgegriffen von Lew & Szarowska (2017), schlagen für ihre Bewertungskategorien eine fünfstufige Skala vor, jedoch werden hier ausschließlich quantitative Kategorien berücksichtigt. In Lew (2023) bewerten vier Experten von GPT-4 generierte Wörterbuchartikel für 15 Kommunikationsverben. Der Fokus liegt dabei ausschließlich auf drei Kategorien (Bedeutungsangaben, Beispieldangaben, Gesamtbewertung des WbA), die auf einer Fünf-Punkte-Skala (von schlecht bis sehr gut) bewertet werden, ohne dass allerdings detaillierte Bewertungskriterien bereitgestellt würden.

Was man in Bezug auf bisherige lexikographische Bewertungsraster vermisst, sind also zum einen die systematische Einbindung qualitativer Kriterien und zum anderen genaue, operationalisierbare Ausführungen, nach welchen Kriterien eine Bewertung jeweils erfolgen soll – analog zu den unverzichtbaren Annotationsrichtlinien bei computerlinguistischen Aufgabenstellungen.

Einen weiteren Fortschritt bei der Bewertung von Wörterbuchqualitäten kann man unseres Erachtens erzielen, wenn man Forschungsergebnisse der germanistischen Textlinguistik (Nussbaumer 1991; Sieber & Nussbaumer 1994, 1995; Fritz 2017; Abel et al. 2020; Wolfer et al. 2023), empirischen Schreibdidaktik (Becker-Mrotzek 2014; Becker-Mrotzek et al. 2018) und qualitativen Medienforschung und Pädagogik (Mayring & Gläser-Zikuda 2008; Mayring & Hurst 2017) zur Bestimmung und Erfassung von Textqualitäten integriert. Hier gelten Bewertungen dann als zuverlässiger, wenn zunächst kritisch-hermeneutisch verschiedene, in die Bewertung einzubindende Einzelmerkmale textsortenspezifisch eruiert werden und wenn diese Einzelmerkmale anschließend in ein möglichst detailliertes Kategoriensystem bzw. einen Kriterienkatalog überführt werden, bei dem die Bewertung auf einem numerischen Punktesystem basiert (vgl. Neumann 2017: 208–211; Mayring & Hurst 2017: 498; Becker-Mrotzek 2014: 507–510).

In diesem Beitrag möchten wir am Beispiel eines Wörterbuchartikels in einem deutschen allgemeinsprachigen Wörterbuch ein mehrstufiges Bewertungsmodell vorstellen (vgl. Abschnitt 3), das speziell für die Bewertung von KI-generierten Texten konzipiert ist, die im Regelfall im Rahmen von Schreibexperimenten erstellt werden (vgl. Abschnitt 2). Das Modell basiert darauf, einen Text (hier: Wörterbuchartikel) nach verschiedenen quantitativen und qualitativen Aspekten zu untersuchen und mit von Menschen erstellten Texten der gleichen Textsorte zu vergleichen. Den Referenzrahmen bilden hierbei sowohl von Experten geschriebene Standardtexte als auch Laientexte, die ebenfalls mittels Schreibexperimenten erhoben werden. Im Zentrum steht ein differenziertes Bewertungsraster, das wir in Abschnitt 4 skizzieren. Dieses kann (und muss) in Abhängigkeit zum funktionellen Wörterbuchtyp modifiziert werden und kann auch als genereller Bewertungsrahmen für KI-generierte Wörterbuchartikel herangezogen werden. Um die Operationalisierung der Bewertungskriterien zu erproben und zu verbessern, wurde das hier vorgestellte Bewertungsraster im Kontext des in Abschnitt 2 dargestellten Schreibexperiments entwickelt. Ein weiterer Fokus des Beitrags liegt darauf, wie sich mittels des Bewertungsrasters die spezifischen Probleme der KI-basierten Textgenerierung systematisch und passgenau herausarbeiten lassen. In Abschnitt 5 führen wir dazu eine detaillierte Analyse KI-generierter Wörterbuchartikel durch und widmen uns auch der Frage, wie mögliche Lösungen für dabei erkannte spezifische Einschränkungen der KI erarbeitet werden können. Das skizzierte Verfahren kann auf andere Textsorten übertragen werden und liefert damit ein allgemeines

Modell für die Bewertung KI-generierter Texte jenseits von Wörterbuchartikeln. Das Flussdiagramm in Abb. 1 zeigt einen gesamthaften Überblick zu unserem Vorgehen.

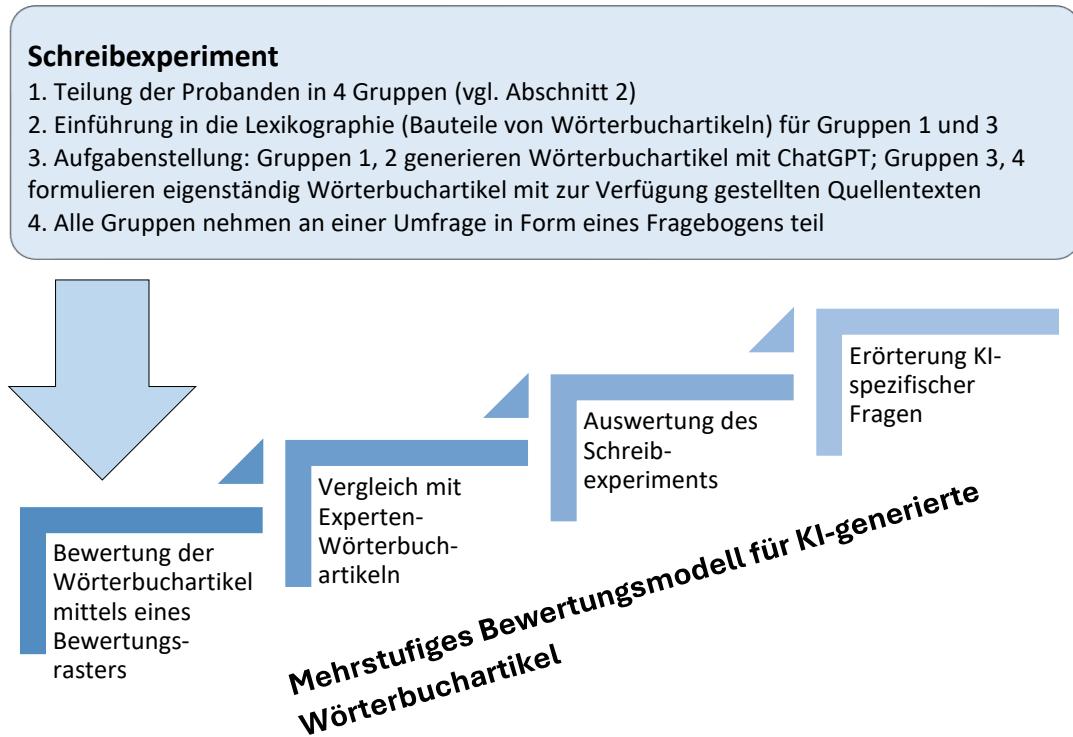


Abb. 1: Methodischer Ansatz des mehrstufigen Bewertungsmodells, dargestellt am Beispiel der Auswertung eines Schreibexperiments zu Wörterbuchartikeln mit und ohne KI-Unterstützung.

2. Schreibexperiment „Wörterbuchartikel“

Eine besondere Herausforderung für die Evaluation KI-generierter Texte liegt darin, dass die Ergebnisse nicht nur von den jeweiligen KI-Modellen, sondern ganz wesentlich von den verwendeten Prompts abhängen (und sogar wiederholte Durchläufe mit identischen Prompts zu sehr unterschiedlichen Texten führen können). Zudem erfolgt die Textgenerierung im praktischen Einsatz meist semiautomatisch, d.h. in einem iterativen Prozess werden die Prompts sukzessive verbessert oder die KI angewiesen, die Textentwürfe im Hinblick auf spezifische Aspekte zu überarbeiten. Aus diesen Gründen müssen unseres Erachtens die zu bewertenden KI-Texte im Rahmen von Schreibexperimenten erhoben werden, in denen menschliche Probanden geeignete Prompts formulieren und Textentwürfe überarbeiten lassen. Auch erscheint es notwendig, eine größere Anzahl von Textvarianten vergleichend zu bewerten, um eine verlässliche Einschätzung spezifischer KI-Modelle zu gewinnen.

Unsere kleine qualitative Pilotstudie ist dementsprechend in Verbindung mit einem Schreibexperiment entstanden, das im Kontext der germanistisch-linguistischen Lehrveranstaltung „Textproduktion mit KI-gestützten Writingtools aus linguistischer Perspektive“ im April 2024 an der FAU Erlangen-Nürnberg durchgeführt wurde. Im Rahmen des Experiments wurden Wörterbuchartikel auf Basis unterschiedlicher Vorkenntnisse zu dieser Textsorte verfasst; sechs der zehn studentischen Teilnehmer (vorwiegend Studierende der Germanistik/Anglistik, auch Computerlinguistik, Informatik und Lexikographie) sollten die Wörterbuchartikel mittels dialogbasierter KI (ChatGPT 3.5) generieren, vier Teilnehmer die Artikel selbstständig ohne Hilfsmittel verfassen. Entsprechend wurden zwei Hauptgruppen

gebildet. Um Textmusterwissen als Variable einbinden zu können, wurde jede der beiden Gruppen noch einmal zweigeteilt: Je die Hälfte erhielt eine 15-minütige lexikographische Kurzeinführung, in deren Rahmen die Bestandteile einer Wörterbuchmikrostruktur vorgestellt und anhand von Beispielen (Artikel „Pferd“ in Duden Bedeutungswörterbuch, Duden online, DWDS) demonstriert wurden. Auch über die gängige lexikographische Praxis, korpusbasiert zu arbeiten und authentische Belege in die Wörterbuchartikel zu integrieren, wurde informiert. Daraus ergeben sich die folgenden Gruppen: mit KI = I, II; ohne KI = III, IV sowie mit lexikographischer Kurzeinführung = I, III und ohne Einführung = II, IV.

Während der 45-minütigen Bearbeitungsphase bearbeiteten sämtliche Gruppen die gleichen Aufgabenstellungen:

Gruppen mit KI (I, II):

Erstellen Sie einen Wörterbuchartikel zu den Stichwörtern *Maus* und *köpfen*. Nutzen Sie dazu ChatGPT 3.5 und dokumentieren Sie alle Ihre Versuche, indem Sie sowohl Ihre Eingabe als auch die Ergebnisse (Chatbot-Output) in ein Word-Dokument kopieren (Reihenfolge Ihres Vorgehens beibehalten). Nutzen Sie ausschließlich die KI. Internet darf keinesfalls benutzt werden, um die Ergebnisse der Studie nicht zu verfälschen.

Gruppen ohne KI (III, IV):

Erstellen Sie einen Wörterbuchartikel in WORD zu den Stichwörtern *Maus* und *köpfen*. Nutzen Sie die dazu beigelegten Quellen. Internet sowie KI dürfen keinesfalls benutzt werden, um die Ergebnisse der Studie nicht zu verfälschen.

Die Teilnehmer, die ohne KI arbeiteten, erhielten pro Lemma sechs Belegzettel, um den lexikographischen Prozess bei der Erarbeitung der Bedeutungsansätze zu simulieren und die semantische Analyse zu unterstützen. Wer ChatGPT 3.5 zur Verfügung hatte, musste sich auf die Trainingsdaten der KI verlassen.

Ziel des Schreibexperiments war es unter anderem, eine vergleichbare Datenbasis für die intendierte Bewertung KI-generierter Wörterbuchartikel zu erhalten und damit die Entwicklung des Bewertungsrasters zu unterstützen.

Ein weiterer Teil des Schreibexperiments bestand im Ausfüllen eines gruppenspezifischen Fragebogens, in dem im Anschluss an das Experiment die Teilnehmer/-innen zu ihren Vorkenntnissen zu Wörterbüchern, zu KI sowie den individuellen Erfahrungen während des Schreibexperiments befragt wurden.

3. Ein mehrstufiges Bewertungsmodell für KI-generierte WbA

Unser Bewertungsmodell für KI-generierte WbA schließt an verschiedene Vorarbeiten an und ergänzt diese: Die Teilaspekte des Bewertungsrahmens orientieren sich an Svensén (2009). Unser Fokus im vorliegenden Beitrag liegt auf der Entwicklung eines Bewertungsrasters für WbA, genauer für WbA in einem deutschen allgemeinsprachigen Wörterbuch. Bei der Erarbeitung der Bewertungskategorien wurden die grundlegenden, bekannten Parameter der Wörterbuchmikrostruktur zugrunde gelegt, und zwar in genauer Adaption für diesen Wörterbuchtyp. Hierfür wurden die Bewertungskriterien von Nielsen (2009) und Tarp (2017) abgeglichen; unser Bewertungsraster greift auch einzelne Aspekte von Ripfel (1989) sowie Pearsons und Nichols (2013) auf. Weiterhin wurden vorhandene Musterartikel (Duden-

Bedeutungswörterbuch, Duden online, DWDS) als Vergleichsgröße herangezogen. Die induktiv ermittelten Bewertungskategorien und -kriterien wurden anschließend mittels Analysen (Bewertung der Wörterbuchartikel aus dem Schreibexperiment) deduktiv erprobt. Da wir davon ausgehen, dass KI-generierte Wörterbuchartikel mit von Menschen generierten Artikeln konkurrieren, wurden im Kontext des Schreibexperiments sowohl Wörterbuchartikel von Menschen ohne KI-Unterstützung als auch von Mensch-KI-Teams erstellt. Beide Artikelkategorien wurden bei der Anwendung des Bewertungsrasters gleichermaßen berücksichtigt. Bei einer Evaluation KI-generierter Wörterbuchartikel sind als weitere Dimension KI-spezifische Fragestellungen zu integrieren. Von besonderem Interesse ist hier, inwiefern beispielsweise die Prompts einen Einfluss auf die Qualität des Outputs haben.

Ziel war es, insgesamt einen Bewertungsrahmen zu entwickeln, der von Experten für die Evaluation von (KI-generierten) Wörterbüchern oder -teilen adaptiert, aber auch von Laien auf eine ad-hoc Bewertung von KI-generieren Wörterbuchartikeln und ähnlichen Texten angewendet werden kann, ohne auf primitive automatisierte Verfahren zurückgreifen zu müssen (z.B. Celikyilmaz et al. 2020).

Für die Evaluation eines einzelnen Wörterbuchartikels wurden ausschließlich Aspekte der Mikrostruktur sowie Adressatenbezug und Wörterbuchtyp als relevant erachtet, nicht sonstige wörterbuchrelevante Kategorien wie etwa Makro-, Medio- oder Zugriffsstrukturen, da der zu beurteilende Text nur einen Wörterbuchausschnitt darstellt. Mit Bezug auf den Wörterbuchtyp (Allgemeinsprachiges Wörterbuch) wurde ein Fokus auf semantische Kategorien als Bewertungsgrundlage gelegt. Weiterhin wurden allgemeine qualitative Aspekte integriert, die textsortenspezifisch die ästhetische, inhaltliche und sprachliche Angemessenheit des Textes berücksichtigten. Die erarbeiteten Kategorien und Kriterien bilden die Basis eines differenzierten Bewertungsrasters (vgl. Abschnitt 4), in dem bei einer bewertenden Analyse die jeweiligen Merkmale mittels numerischer Punktwertangaben erfasst werden können.

Aus den oben beschriebenen Konzepten ergibt sich ein vollständiger systematischer Ansatz zur Bewertung KI-generierter Wörterbuchartikel, der die folgenden vier Ebenen umfasst:

(1) Bewertung der Wörterbuchartikel mittels eines Bewertungsrasters

Auf der ersten Stufe des Bewertungsmodells werden KI-generierte Wörterbucheinträge anhand eines kleinteiligen Bewertungsrasters (vgl. Abschnitt 4), das quantitative und qualitative Kriterien für jede Hauptkategorie im Form- und Semantikkommentar eines Wörterbucheintrags enthält (z.B. Aussprache-, Polysemie- oder Beispiellangabe), evaluiert (vgl. Abschnitt 5.1). Das Raster kann darüber hinaus auch auf von menschlichen Experten verfasste Wörterbucheinträge in verschiedenen Print- und Online-Wörterbüchern angewendet werden, insofern für den jeweiligen Wörterbuchtyp spezifische Anpassungen vorgenommen worden sind.

(2) Vergleich mit anderen Wörterbuchartikeln

Die zweite Stufe des Bewertungsmodells stellt der Vergleich mit anderen WbA dar (vgl. Abschnitt 5.2). Zum einen werden alle auf Basis des Bewertungsrasters evaluierten KI-generierten WbA miteinander verglichen, zum anderen erfolgt ein Vergleich mit existierenden WbA, die von Experten abgefasst sind (hier: Duden.de), sowie mit von Laien erstellten WbA aus Schreibexperimenten (soweit vorhanden).

(3) Auswertung eines Schreibexperiments

Wie in Abschnitt 2 dargelegt, sind die zu bewertenden KI-generierten WbA im Rahmen eines Schreibexperiments (inklusive Fragebogenerhebung) entstanden. Die Auswertung dieses Schreibexperiments bildet die nächste Stufe unseres Bewertungsmodells (vgl. Abschnitt 5.3). Im Mittelpunkt der Auswertung stehen Aspekte wie ein Vergleich der mit bzw. ohne KI-Unterstützung verfassten WbA, ein Vergleich von generierten bzw. menschlich verfassten WbA mit bzw. ohne lexikographische Kurzeinführung sowie die im Fragebogen erfassten Informationen zu den Vorerfahrungen der Probanden mit Wörterbüchern und KI-Tools.

(4) Berücksichtigung KI-spezifischer Fragestellungen

Da unser Fokus auf der Auswertung KI-generierter Wörterbuchartikel liegt, werden auf der vierten Stufe des Bewertungsmodells KI-spezifische Aspekte integriert (5.4). Dazu gehören beispielsweise

- Angaben, die die KI nicht verwirklichen kann;
- die Eingabeaufforderungen (Prompt) bzw. iterative Überarbeitung der WbA;
- ein Vergleich der Mensch-KI-Teams mit einem KI-generierten Wörterbuchartikel, für den ein Prompt durch einen lexikographischen Experten optimiert wurde; und
- die Diskussion KI-spezifischer Fehler.

4. Entwicklung des Bewertungsrasters

Für die bewertende Analyse der Wörterbuchartikel werden der gesamte Artikel sowie die lexikographischen Angaben im Einzelnen unter den beiden Hauptaspekten (a) Quantität und (b) Qualität untersucht. Angaben zur Quantität beziehen sich auf das Vorhandensein von Angaben sowie auf deren Umfang bzw. Vollständigkeit. Angaben zur Qualität beziehen sich auf inhaltliche und sprachliche Korrektheit, inhaltliche und sprachliche Angemessenheit, Authentizität usw. der Angaben, wobei für jede Angabe individuell Qualitäts- und Quantitätskriterien ermittelt werden. Quantitative Kriterien werden zwar berücksichtigt, das Hauptaugenmerk liegt aber (in Abhängigkeit zum Wörterbuchtyp) auf der Qualität der Daten, welche Kriterien dementsprechend viel umfangreicher angesetzt und zudem auch höher bewertet werden. Weiterhin gilt es, Bewertungsmaßstäbe pro Kriterium festzulegen.

Die Bewertungskategorien, die entsprechenden Quantitäts- und Qualitätskriterien sowie bei der konkreten Anwendung des Rasters die Bewertungsmaßstäbe hängen im Detail unmittelbar vom jeweiligen Wörterbuchtyp und damit zusammenhängend den intendierten Adressaten ab.

Der Wörterbuchtyp gibt Aufschluss über ein mögliches Informationsprogramm¹ (*sensu* Engelberg & Lemnitzer 2009: 25) der Mikrostruktur (auch Mikrostrukturenprogramm), sprich über die zu bewertenden Angaben. Als Vergleichsgrundlage für den in Stufe 2 integrierten Wörterbuchvergleich (vgl. Abschnitt 3) ziehen wir hier Lemmata aus dem Onlinewörterbuch Duden.de heran. Duden gibt an, „umfassende Informationen zu Rechtschreibung, Grammatik und Bedeutung eines Wortes“ und Informationen über „den richtigen Gebrauch sowie die Aussprache und Herkunft eines Wortes und [...] dessen Synonyme“ zu bieten.² Wörterbücher mit einem derartigen Umfang an Angaben werden als polyinformative Wörterbücher (*sensu* Wiegand 2010: 85) bezeichnet, welche zu den Allgemeinwörterbüchern (*sensu* Engelberg & Storrer 2016: 41) zählen, da sie im Gegensatz zu informationstyporientierten Spezialwörterbüchern (ebd.: 42) keinen Fokus auf bestimmte Angaben legen (wie z.B. ein Synonymwörterbuch) und in der Auswahl der Lemmata im Gegensatz zu lemmatyporientierten Wörterbüchern (ebd.) keine Begrenzung aufweisen (wie z.B. ein Fremdwörterbuch).³ Der Wörterbuchtyp inkludiert darüber hinaus auch die jeweilige Wörterbuchbenutzergruppe, was ebenfalls nicht nur für die einzelnen Kriterien, sondern auch für deren Bewertung grundlegend ist. Beispielsweise ist das inhaltliche und sprachliche Niveau einer Bedeutungsangabe in einem Schülerwörterbuch ein anderes als in einem Allgemeinwörterbuch oder einem terminologischen Fachwörterbuch. Dies ist sowohl bei der Erstellung als auch bei der Anwendung des Bewertungsrasters durchgängig zu berücksichtigen.

Die für die Bewertung leitenden Hauptkategorien und einzelnen Angaben (*sensu* Wiegand 1989: 468) sind in Tab. 1 dargestellt.

Tab. 1: Hauptkategorien des Bewertungsrasters

Mikrostruktur gesamt	Textstruktur / Design, Gesamtbewertung
Formkommentar	Lemmaziechengestaltangabe, Ausspracheangabe, Grammatikangaben, Rechtschreibungsangabe, Frequenzangabe
Semantikkommentar	Polysemieangabe, Pragmatikangabe, Bedeutungsangabe, Beispielangabe Kollokationsangabe, Wortbildungsangabe, Sprichwortangabe, Synonymieangabe, Etymologieangabe

Für jede Angabe wurden jeweils quantitative und qualitative Kriterien erarbeitet, die hinsichtlich der Gesamtbetrachtung und des Formkommentars hauptsächlich auf metalexikographischen Kategorien beruhen. Konkrete Fragen, die sich daraus ergeben, können wiederum abgeleitet werden (vgl. Tarp 2017: 122–125). Da es in der vorliegenden Pilotstudie um Wörterbuchartikel in einem allgemeinsprachigen Wörterbuch geht, liegt ein Bewertungsschwerpunkt auf dem Semantikkommentar und entsprechenden Bewertungskriterien. Die Erarbeitung dieser Kriterien soll im Folgenden exemplarisch am Beispiel der Bedeutungsangabe genauer erläutert werden. Eine vollständige Herleitung sämtlicher Bewertungskriterien muss hier aus Platzgründen unterbleiben.

¹ Generell gibt es keine festen Informationsprogramme für Mikrostrukturen in Allgemeinwörterbüchern, wobei in einigen empirischen Arbeiten, die Angaben in verschiedenen bestehenden Online- und Printwörterbüchern benannt und aufgelistet werden (vgl. Engelberg & Lemnitzer 2009; Flinz 2011; Geyken & Lemnitzer 2016; Wiegand, 1989 [empirisch für Printlexikographie]).

² <<https://www.duden.de/wörterbuch>> (letzter Zugriff: 31.05.2024).

³ Weiteres zu Wörterbuchtypologien u.a. in Kühn 1989; Hausmann 1989; Engelberg & Lemnitzer 2009; Wiegand 2010; Engelberg & Storrer 2016.

4.1 Exemplarische Herleitung von Quantitäts- und Qualitätskriterien

Quantitative Kriterien bei Bedeutungsangaben beziehen sich nur auf ihr Vorhandensein und ihre Vollständigkeit: Es wird lediglich bewertet, ob generell Bedeutungsangaben vorhanden sind und ob für jedes Semem eine Bedeutungsangabe formuliert worden ist. Quantitative Kriterien werden jeweils mit 0 oder 1 Punkt bewertet.

Die qualitativen Kriterien betreffen Allgemeinsprachlichkeit, Redundanz, Genus proximum, Differentia specifica, sprachliche sowie inhaltliche Korrektheit und die Informationsdichte: Allgemeinsprachlichkeit ist dann gegeben, wenn die Formulierungen nicht fachsprachlich sind oder sich beispielsweise an einem Soziolekt orientieren; Redundanz ist in einem verdichteten Wörterbuchartikel generell zu vermeiden; die Integration von Genus proximum und Differentia specifica dient dazu, Wörter kategorial systematisch zu unterscheiden und semantisch innerhalb einer Kategorie einzuordnen; sprachliche Korrektheit liegt vor, wenn weder orthographische noch grammatische Fehler zu finden sind; inhaltliche Korrektheit setzt voraus, dass die Bedeutungsangaben inhaltlich zutreffend sind; die Informationsdichte bezieht sich auf den Umfang der Bedeutungsangaben, also ob etwa nur ein Wort (Bedeutungäquivalent) angegeben ist oder eine vollständige Bedeutungsparaphrase. Hier werden jeweils 0, 1 oder 2 Punkte vergeben, was einerseits die Qualitätskriterien in der Gesamtbewertung hervorhebt, andererseits aber auch eine feinere Bewertung zulässt: Liegt z.B. Allgemeinsprachlichkeit vor, werden 2 Punkte vergeben; liegt sie nur teilweise vor, nur 1 Punkt und liegt keine vor 0 Punkte.

4.2 Das Bewertungsraster im Überblick

Im Folgenden wird das Gesamtraster für alle Bewertungskategorien mit sämtlichen Kriterien dargestellt (vgl. Tab. 2). Das Raster ist als metalexikographischer Beitrag für eine systematische Bewertung von Wörterbuchartikeln im Allgemeinen zu verstehen und kann sowohl für von Menschen verfasste als auch für KI-generierte WbA herangezogen werden. Generell kann und muss das Raster für die Anwendung auf verschiedene Typen von Wörterbüchern bzw. sogar für unterschiedliche lexikalische Entitäten jeweils angepasst werden. D.h., dass z.B. in bilingualen Wörterbüchern andere Angabetypen und somit andere Kriterien relevant sind (z.B. geringere Differenziertheit der Bedeutungsangabe, aber Einbindung von Äquivalenzangaben, höhere Gewichtung von Kollokationsangaben) oder dass substantivische Lemmata andere Bewertungskategorien erfordern als verbale.

Tab. 2: Bewertungsraster Mikrostrukturen

Kategorien (Bewertungsfelder)	Kriterien (max. 136)	Quantität (0–1 max. 30) Qualität (0–2 max. 106)
Gesamte Mikrostruktur (max. 18)		
Textstruktur / Design (max. 8)	nicht-/typographische Mikrostrukturanzeliger, Absätze, Wörterbuch-struktur erkennbar	
	Korrekttheit Mikrostrukturanzeliger, Verständlichkeit	
Gesamtbewertung (max. 10)	Wörterbuch-funktionen erfüllt, WB-Adressaten berücksichtigt, Prägnanz inhaltlich / sprachlich, Allgemeinsprachlichkeit	
Formkommentar (max. 32)		
Lemmaziechen-	Vorhandensein	

gestaltangabe (max. 3)	Korrekttheit sprachlich
Ausspracheangabe (max. 6)	Vorhandensein, Betonung vorhanden
	Korrekttheit, Betonung korrekt
Grammatikangabe (max. 14)	Wortartangabe vorhanden, Genusangabe vorhanden, Deklinationsangabe vorhanden, Paradigma abdeckend
	Wortartangabe korrekt, Genusangabe korrekt, Deklinationsangabe korrekt, Ohne Redundanz, Paradigma korrekt
Rechtschreibangabe (max. 6)	Vorhandensein, Worttrennung vorhanden
	Korrekttheit, Worttrennung korrekt
Frequenzangabe (max. 3)	Vorhandensein
	Korrekttheit
Semantikkommentar (max. 86)	
Polysemieangabe (max. 4)	Vorhandensein, Vollständigkeit
	Gebräuchlichkeit
Pragmatikangabe (max. 6)	Vorhandensein, Vollständigkeit
	Korrekttheit sprachlich, Differenziertheit
Bedeutungsangabe (max. 16)	Vorhandensein, Vollständigkeit
	Allgemeinsprachlichkeit, Informationsdichte, Korrektheit sprachlich / inhaltlich, Genus proximum, Differentia specifica, ohne Redundanz
Beispielangaben (max. 14)	Vorhandensein, Vollständigkeit
	Verständlichkeit, Korrektheit sprachlich / inhaltlich, Authentizität, Angemessenheit inhaltlich, Quellenangabe vorhanden
Kollokationsangabe (max. 104)	Vorhandensein, Umfang
	Verständlichkeit, Korrektheit sprachlich / inhaltlich, Authentizität, empirisch belegbar, Sortierung nach Lesarten
Sprichwortangabe (max. 5)	Vorhandensein
	Korrekttheit sprachlich, Authentizität
Wortbildungsangabe (max. 10)	Vorhandensein, Umfang
	Korrekttheit sprachlich / inhaltlich, mehrere Wortbildungsarten abgedeckt, Usualitätsgrad
Synonymieangabe (max. 10)	Vorhandensein, Vollständigkeit
	Korrekttheit sprachlich / inhaltlich, Gebräuchlichkeit, empirisch belegbar (Ersatzprobe)
Etymologieangabe (max. 7)	Vorhandensein
	Korrekttheit sprachlich / inhaltlich, Informationsdichte

Die in Tab. 2 dargestellte Matrix für einen substantivischen Wörterbuchartikel in einem allgemeinsprachigen Wörterbuch umfasst insgesamt 136 zu vergebende Punkte, wobei 30 Punkte auf quantitative und 106 Punkte auf qualitative Kriterien entfallen. Auf allgemeine Kategorien fallen insgesamt maximal 18 Punkte, auf Angaben im Formkommentar 32 und auf Angaben im Semantikkommentar 86 Punkte. Für eine Gesamtbewertung schlagen wir (in Anlehnung an Benotungsskalen) die in Tab. 3 dargestellten Prädikate vor, wobei 94-100 % ein sehr gutes Ergebnis (Prädikat I) und alles unter 50 % unzureichende Ergebnisse darstellen. Die Farbgebung orientiert sich für eine leicht zugängliche Visualisierung an den Ampelfarben (grün = exzellent > gelb = durchschnittlich > rot = ungenügend).

Tab. 3: Gesamtbewertung (Prädikate)

98-100 %	133-136 Punkte
----------	----------------

94-97 %	128-132 Punkte	Prädikat I: erfüllt quantitative und qualitative Anforderungen hinsichtlich des Wörterbuchtyps hervorragend
90-93 %	122-127 Punkte	Prädikat II: erfüllt quantitative und qualitative Anforderungen hinsichtlich des Wörterbuchtyps überdurchschnittlich
85-89 %	116-121 Punkte	
80-84 %	109-115 Punkte	
76-79 %	103-108 Punkte	Prädikat III: erfüllt quantitative und qualitative Anforderungen hinsichtlich des Wörterbuchtyps durchschnittlich
71-75 %	97-104 Punkte	
66-70 %	90-96 Punkte	
61-65 %	83-89 Punkte	Prädikat IV: erfüllt quantitative und qualitative Anforderungen hinsichtlich des Wörterbuchtyps unterdurchschnittlich
50-60 %	68-82 Punkte	
0-49 %	00-67 Punkte	Prädikat V: erfüllt quantitative und qualitative Anforderungen hinsichtlich des Wörterbuchtyps nicht

4.3 Exemplarische Anwendung des Bewertungsrasters

Anhand einer Bewertung des Lemmas *Maus* auf Duden.de wird im Folgenden die Anwendung des Bewertungsrasters exemplarisch vorgestellt, wobei das Ergebnis zugleich für den Vergleich in Stufe 2 des Bewertungsmodells herangezogen wird (vgl. Abschnitt 5.2). Die Auswertung der Matrix ergibt, dass der Duden-Wörterbuchartikel zum Lemma *Maus* mit 95 von 136 Punkten insgesamt eine Bewertung von 70 % erreicht und folglich im unteren Bereich des Prädikats III („erfüllt quantitative und qualitative Anforderungen hinsichtlich des Wörterbuchtyps durchschnittlich“) verortet werden kann. Dabei ist zu beachten, dass dieses Prädikat nicht generell für die Wörterbuchartikel auf Duden.de gilt, da hierfür eine repräsentative Stichprobe von WbA zu deutschen Substantiven vorgenommen werden müsste. Der Kontrollartikel zum Lemma *Katze* liegt mit 78 % beispielsweise bereits im oberen Bereich von Prädikat III. Dass der Artikel *Maus* aus Duden.de, obwohl von Experten erstellt und somit als Referenz anzuerkennen, hier keine 100 % erreicht, liegt daran, dass nicht alle möglichen Kategorien umgesetzt sind (z.B. fehlen Quellenangaben). Auch sind teilweise ungebräuchliche Synonymangaben integriert, was in Entsprechung zu den erarbeiteten qualitativen Kriterien (vgl. Abschnitt 4.1) hier negativ ins Gewicht fällt.

Exemplarische Beschreibung der Bewertung der Bedeutungsangabe: Bedeutungsangaben sind im Duden-Artikel zu *Maus* generell vorhanden, aber nicht vollständig, da die Bedeutungsangabe bei Semem „weibliche Scham; Vulva“ fehlt (bzw. fehlt die Verlinkung zum Lemma *Vulva*, wie das etwa bei Semem 2 „Kosewort“ der Fall ist). Die vorhandenen bzw. verlinkten Bedeutungsangaben sind sprachlich und inhaltlich korrekt. Die Angaben sind als allgemeinsprachlich zu bezeichnen, da sie weder fachsprachlich noch umgangssprachlich sind. Angaben zur Einordnung des Wortes in übergeordnete Kategorien wie Arten von Lebewesen liegen ebenfalls vor: „kleines [graues] Nagetier“, wobei die Unterscheidung innerhalb der Kategorie etwas unspezifisch ist und keine ganz klare Abgrenzung beispielsweise gegenüber *Ratte* ermöglicht. Semantisch überdifferenziert und damit redundant erscheint die Differenzierung der Bedeutungsangaben innerhalb von Semem 4 „a) Geld“ und „b) Euro, Mark o.Ä.“. Die Informationsdichte ist für die Bedeutungsangaben von unterschiedlichem Umfang: Während die Angaben zu Semem 1 (Nagetier) und Semem 5 (Computermaus) relativ umfangreich sind und Bedeutungsdefinitionen im eigentlichen Sinne darstellen, werden andere wie Semem 6 (weibliche Scham, Vulva) nur mittels einer synonymen Wortgruppe umschrieben. Die Gebräuchlichkeit der Angaben wird unter der Kategorie Polysemieangabe bewertet, weil z.B. die Frequenz von Semem 6 allgemeinsprachlich zumindest fragwürdig erscheint.

Das insgesamt nicht ideale Abschneiden des Duden-Wörterbuchartikels röhrt weiterhin daher, dass Kompositums- und Kollokationsangaben fehlen. Das bedeutet, dass der Artikel in diesen Kategorien insgesamt 24 mögliche Punkte verliert. Hier zeigt sich bereits, dass die gesamthafte Bewertung nach einem Prädikat möglicherweise zu pauschal sein kann und ein Mehrwert des Bewertungsrasters in seiner Detailhaftigkeit liegt (vgl. unten).

5. Evaluation nach dem mehrstufigen Bewertungsmodell

5.1 Bewertung von Wörterbuchartikeln mittels des Bewertungsrasters

5.1.1 Lemma *Maus*

Die generierten Wörterbuchartikel (WbA) wurden von zwei Experten unabhängig voneinander auf Basis des Bewertungsrasters (vgl. Abschnitt 4) evaluiert. Ein wichtiger Vorteil unseres Rasters ist, dass mit seiner Hilfe die Reliabilität der Bewertungen wesentlich besser überprüft werden kann als durch einen bloßen Vergleich der summarischen Gesamtbewertungen. Dies lässt sich am besten an einem Extrembeispiel verdeutlichen: Wenn wir annehmen, dass beide Experten ihre Punkte für einen WbA (aus Faulheit) völlig zufällig vergeben, so würde in beiden Fällen sehr wahrscheinlich eine Gesamtbewertung zwischen 43 % und 57 % herauskommen, was auf den ersten Blick eine recht gute Übereinstimmung suggeriert. Erst bei Betrachtung der Einzelkriterien wird sichtbar, dass sie nur in etwa bei der Hälfte aller Punkte übereinstimmen – genau der Wert, der durch rein zufällige Übereinstimmungen zu erwarten ist.

Daher berechnen wir Cohen's kappa als Maß für die tatsächliche Übereinstimmung der beiden Experten (Inter-Annotator Agreement, vgl. Artstein & Poesio 2008) und damit die Reliabilität der Annotation. Kappa-Werte liegen üblicherweise im Bereich $0 \leq \kappa \leq 1$, wobei oft $\kappa > .8$ als sehr gute und $\kappa > .67$ als akzeptable Übereinstimmung gewertet wird (Artstein & Poesio 2008: 576). Um kappa auf unsere Daten anwenden zu können, interpretieren wir die Bewertung von qualitativen Kriterien als Kombination von zwei binären Entscheidungen: ein Punkt wird für Mindestanforderungen vergeben, ein weiterer Punkt, wenn das Kriterium in vollem Umfang erfüllt ist. Auf diese Weise erhalten wir für die Artikel aus dem Schreibexperimente zufriedenstellende kappa-Werte zwischen .67 und .80, für die beiden Duden-Artikel sogar .86 und .87. Eine Ausnahme bildet WbA 1.2, mit einem überraschend schlechten Wert von $\kappa = .35$; hier wird noch zu untersuchen sein, warum die beiden Experten zu so unterschiedlichen Einschätzungen kamen.

Die Gesamtbewertungen beider Experten sind – wie erwartet – auch bei WbA 1.2 deutlich ähnlicher, als die kappa-Werte nahelegen. Tab. 4 zeigt die gemittelten Ergebnisse der beiden Experten für die verschiedenen Artikel der Probanden aus dem Schreibexperiment (1.1-4.2) sowie für zwei Experten-Artikel (5.1, 5.2) und einen von einem Experten mit Hilfe der KI generierten Artikel (6.1). Dargestellt sind neben der Gesamtbewertung (Prozentsatz aller erreichbaren Punkte) auch separate Punktzahlen für jeden Bewertungsbereich sowie die einzelnen Kategorien. Auf eine detaillierte Aufschlüsselung aller Kriterien wie in Tab. 2 muss aus Platzgründen verzichtet werden. Die Farbkodierung orientiert sich an den Farben der Tab. 3 und visualisiert, welche Kategorien im Detail mit den Prädikaten I–V bewertet wurden.

Tab. 4: Ergebnisse *Maus*

Quelle		Schreibexperiment										Duden	KI	
Gruppen		I			II			III		IV		Experten		
Artikelnummer		1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	4.1	4.2	5.1	5.2	6.1
Verwendung von KI		j	j	j	j	j	j	n	n	n	n	x	x	j
Lexikogr. Einführung		j	j	j	n	n	n	j	j	n	n	x	x	x
Expertenartikel		n	n	n	n	n	n	n	n	n	n	j	j	j
%	100	45	59	55	41	53	39	56	43	29	41	70	78	90
Kategorien / Punkte	136	61.5	80.5	74.5	56	71.5	53.5	76	58.5	39.5	56	95	106	122.5
Gesamte Mikrostruktur	18	4.5	13	14	12	15.5	10.5	15	14	10	13	18	18	17.5
Textstruktur / Design	8	1.5	6.5	6.5	6.5	7.5	5	8	7	5	7	8	8	8
Gesamtbewertung	10	3	6.5	7.5	5.5	8	5.5	7	7	5	6	10	10	9.5
Formkommentar	32	16.5	17	17.5	9.5	15.5	15.5	13.5	10	14.5	9	30.5	32	30
Lemmazichen-gestaltungangabe	3	2.5	2.5	2.5	2.5	3	3	2.5	3	3	3	3	3	2.5
Ausspracheangabe	6	1.5	0	4	3	2	4	0	2	0	0	4.5	6	4.5
Grammatikangabe	14	12.5	12	11	4	10.5	8.5	11	5	11.5	6	14	14	14
Rechtschreibangabe	6	0	0	0	0	0	0	0	0	0	0	6	6	6
Frequenzangabe	3	0	2.5	0	0	0	0	0	0	0	0	3	3	3
Semantikkommentar	86	40.5	50.5	43	34.5	40.5	27.5	47.5	34.5	15	34	46.5	56	75
Polysemieangabe	4	0.5	2.5	2.5	1	3	2	4	3	3	4	3	3.5	4
Pragmatikangabe	6	1	1	4.5	2.5	0	0	0	0	0	2.5	6	5	5
Bedeutungsangabe	16	14	11	12	8.5	13.5	10.5	13	11.5	12	13.5	12.5	14.5	14
Beispielangaben	14	11.5	10.5	10	9	10	9	14	10.5	0	14	9.5	12	12
Kollokationsangabe	14	0	8.5	0	0	0	0	0	0	0	0	0	0	12.5
Wortbildungsangabe	10	4	4	0	0	0	0	5	5	0	0	0	0	8.5
Sprichwortangabe	5	0	0	0	0	0	0	2.5	4.5	0	0	5	5	4.5
Synonymieangabe	10	9.5	7.5	8.5	7.5	7.5	0	9	0	0	0	4.5	10	8
Etymologieangabe	7	0	5.5	5.5	6	6.5	6	0	0	0	0	6	6	6.5

Bei der Auswertung der Ergebnisse fällt auf, dass bei Betrachtung der Gesamtbewertung lediglich vier der zehn Wörterbuchartikel aus den Schreibexperimenten überhaupt die Mindestanforderungen an die Qualität eines Wörterbuchartikels erfüllen. Diese treten bis auf eine Ausnahme nur bei den Gruppen auf, die eine lexikographische Einführung erhalten haben und über lexikographisches Grundlagenwissen verfügen. Dennoch sind diese jeweils im unteren Bereich des Prädikats IV („erfüllt quantitative und qualitative Anforderungen hinsichtlich des Wörterbuchtyps unterdurchschnittlich“) verortet.

Die feine Granularität des Bewertungsrasters ermöglicht jedoch nicht nur eine pauschale Gesamtbewertung, sondern spezifischere Einblicke. So schneiden die sechs KI-generierten Wörterbuchartikel der Probanden aus dem Schreibexperiment (1.1 bis 2.3) in Bezug auf die Gesamtstruktur noch einigermaßen befriedigend ab – ihr Layout und ihre Struktur erfüllen immerhin die Erwartungen an einen idealtypischen Wörterbuchartikel (durchschnittlich 11.6 von 18 Punkten = 64 % = Prädikat IV+). Probleme zeigen sich jedoch

im Detail: Bei den inhaltlichen Aspekten, sowohl in formaler als auch in semantischer Hinsicht, erreichen die KI-Einträge durchschnittlich jeweils weniger als die Hälfte der möglichen Punkte, wobei der Semantikkommentar (39.4 von 86 Punkten = 46 % = Prädikat V) insgesamt noch etwas schlechter abschneidet als der Formkommentar (15.3 von 32 Punkten = 48 % = Prädikat V). Viele wichtige Kategorien fehlen, und es gibt einige inhaltliche Fehler. Allerdings sind die mit KI-Unterstützung generierten Artikel vergleichsweise noch besser als diejenigen der menschlichen Probanden (Studierende, keine Experten). Hier gelangen die Semantikkommentare (32.8 von 86 Punkten = 38 % = Prädikat V) minimal besser als die Formkommentare (11.8 von 32 Punkten = 37 % = Prädikat V), jedoch sind beide im Durchschnitt ungenügend. Legt man jedoch einen Fokus auf die für allgemeinsprachige Wörterbücher zentrale Bedeutungsangabe, so schneiden die Artikel der menschlichen Probanden besser ab und erreichen mit 78 % (12.5 von 16 Punkten) Prädikat III+, während die mit KI-Unterstützung generierten Artikel nur 72 % (11.6 von 16 Punkten = Prädikat III) erzielen. Die mit 8.5 Punkten hier am schlechtesten bewertete Bedeutungsangabe⁴ lässt einige erwartbare Schwächen der KI hervortreten: Die Anforderung der Allgemeinsprachlichkeit ist nicht gut erfüllt („Familie der Muridae“), die Differentia specifica (im Vergleich zu *Ratte*) ist nicht gut erfasst, und es sind zu viele enzyklopädische Detailinformationen enthalten.

Es folgt eine exemplarische Auswertung des bestbewerteten Wörterbuchartikels aus dem Schreibexperiment: WbA 1.2 wurde durch einen Probanden mit lexikographischer Einführung mittels ChatGPT 3.5 generiert. Im Bereich des Formenkommentars sind weder eine Rechtschreibangabe noch eine Ausspracheangabe vorhanden. Alle berücksichtigten Grammatikangaben sind vorhanden, wobei das Deklinationsparadigma nicht abgedeckt ist. Die doppelte Angabe der Pluralform ist redundant. Sprachlich und inhaltlich sind alle Angaben korrekt. Im Bereich des Semantikkommentars fehlen die Wortbildungsangabe, eine pragmatische Angabe sowie die Sprichwortangabe. Die Bedeutungsangabe zeigt nur drei der vier geforderten Sememe, wobei Semem 3 auch inhaltlich fragwürdig erscheint: „scheuer Mensch“. Die hauptsächlichen Sememe Nagetier und Computereingabegerät zeigen sehr kurze bzw. keine Bedeutungsangabe. Die kategoriale Einordnung liegt zumindest bei Semem 1 vor, wobei diese Elemente der Fachsprache aufweist („Familie Muridae“). Pro Semem liegen sowohl eine Beispielaufgabe als auch mehrere Kollokationsangaben vor. Die Beispiele und Kollokationen bei Semem 1 und 2 erscheinen authentisch und angemessen; die Angaben bei Semem 3 („Angst vor Mäusen haben“) sind empirisch aufgrund mangelnder Salienz fragwürdig. Quellenangaben werden bei der KI generell nicht gemacht. Die Synonymangabe liegt für Semem 1 und 2 vor. Die Etymologieangabe ist ebenfalls vorhanden, bleibt hinsichtlich der Informationsdichte eher gering. Inhaltlich und sprachlich sind bis auf die angesprochenen Aspekte alle Angaben korrekt. Hinsichtlich der Gesamtbewertung finden sich im Artikel Mikrostrukturanziger sowie weitere strukturelle Elemente, die den Eindruck eines Wörterbuchartikels vermitteln; der Artikel ist verständlich geschrieben, erfüllt die Wörterbuchfunktionen und berücksichtigt auch die intendierten Adressaten.

⁴ aus Artikel 2.1, Proband ohne lexikographische Einführung mittels ChatGPT 3.5: „Die Maus ist ein kleines Nagetier aus der Familie der Muridae. Sie zeichnet sich durch ihren schlanken Körper, die großen Ohren und den langen Schwanz aus. Mäuse sind in der Regel nachaktiv und ernähren sich von einer Vielzahl von Nahrungsmitteln, darunter Getreide, Früchte und Insekten. Im Gegensatz zur Ratte haben Mäuse normalerweise eine schlankere Körperform und einen feineren Körperbau.“

5.1.2 Lemma *köpfen*

Für das Lemma *köpfen* wurden einige Änderungen am Raster vorgenommen, die auf die unterschiedlichen Wortarten der Lemmata zurückzuführen sind. Konkret haben sich die Kriterien der Kategorien Grammatik- sowie die Bedeutungsangabe geändert. So wurde etwa innerhalb der Grammatikangabe das Kriterium *Genusangabe vorhanden* gestrichen und *Deklinationsangabe korrekt* durch *Konjugationsangabe korrekt* ersetzt. Das führt zu einer anderen Gesamtpunktzahl als bei *Maus*, die für das Lemma *köpfen* 129 beträgt. Tab. 5 zeigt wiederum den Mittelwert beider Experten:

Tab. 5: Ergebnisse *köpfen*

Quelle		Schreibexperiment										Duden	KI
		I		II			III		IV		Experten		
Gruppen		k1.1	k1.2	k1.3	k2.1	k2.2	k2.3	k3.1	k3.2	k4.1	k4.2	k5.1	k6.1
Artikelnummer		j	j	J	j	j	J	N	n	n	n	x	j
Verwendung von KI		j	j	j	n	n	N	J	j	n	n	x	x
Lex. Einführung		j	j	j	n	n	N	J	j	n	n	x	x
Expertenartikel		n	n	n	n	n	N	N	n	n	n	j	j
%	100	35	51	39	48	40	37	55	35	29	34	66	80
Kategorien / Punkte	129	45	66	50.5	62	52	48	71.5	45	37	44	85.5	103
Gesamte Mikrostruktur	18	6	12.5	10.5	14.5	11	13.5	15	11.5	9.5	11.5	18	15
Textstruktur / Design	8	4	7	5	8	6	7.5	7.5	6	5.5	5.5	8	7.5
Gesamtbewertung	10	2	5.5	5.5	6.5	5	6	7.5	5.5	4	6	10	7.5
Formkommentar	29	9	14	12	16.5	17	16.5	17	11	7.5	6	25.5	27
Lemmazeichen-gestaltangabe	3	3	3	3	3	2	2	3	3	3	3	3	3
Ausspracheangabe	6	1	0	4.5	3.5	4.5	4.5	0	2.5	0	0	6	5.5
Grammatikangabe	11	5	9	3.5	8.5	9.5	9.5	8	4	3	1.5	10.5	9.5
Rechtschreibangabe	6	0	0	1	1.5	1	0.5	6	1.5	1.5	1.5	3	6
Frequenzangabe	3	0	2	0	0	0	0	0	0	0	0	3	3
Semantikkommentar	82	30	39.5	28	31	24	18	39.5	22.5	20	26.5	42	61
Polysemieangabe	4	3	3	3	1	2	1	3	2	3	4	4	4
Pragmatikangabe	6	0	0	0	3.5	0	0	0	0	1.5	0	5.5	5.5
Bedeutungsangabe	12	8.5	9	9.5	8.5	8	8.5	10	9	7.5	11	11.5	10.5
Beispielangaben	14	9.5	6.5	8.5	7.5	4	6	13.5	8	5	11.5	12	11.5
Kollokationsangabe	14	0	9	0	0	0	0	0	0	0	0	0	10.5
Wortbildungsangabe	10	2	0	0	2.5	0	0	5	3.5	0	0	0	3
Sprichwortangabe	5	0	0	0	0	0	0	0	0	0	0	0	3.5
Synonymieangabe	10	7	9	4.5	3	8	0	8	0	0	0	9	8
Etymologieangabe	7	0	3	2.5	5	2	2.5	0	0	3	0	0	4.5

Tab. 6: Gesamtbewertungsskala *köpfen*

98-100 %	133-129 Punkte	Prädikat I: erfüllt quantitative und qualitative Anforderungen
94-97 %	128-132 Punkte	hinsichtlich des Wörterbuchtyps hervorragend
90-93 %	122-127 Punkte	

85-89 %	116-121 Punkte	Prädikat II: erfüllt quantitative und qualitative Anforderungen hinsichtlich des Wörterbuchtyps überdurchschnittlich
80-84 %	109-115 Punkte	
76-79 %	103-108 Punkte	Prädikat III: erfüllt quantitative und qualitative Anforderungen hinsichtlich des Wörterbuchtyps durchschnittlich
71-75 %	97-104 Punkte	
66-70 %	90-96 Punkte	
61-65 %	83-89 Punkte	Prädikat IV: erfüllt quantitative und qualitative Anforderungen hinsichtlich des Wörterbuchtyps unterdurchschnittlich
50-60 %	68-82 Punkte	
0-50 %	00-67 Punkte	Prädikat V: erfüllt quantitative und qualitative Anforderungen hinsichtlich des Wörterbuchtyps nicht

Die Evaluation für das Lemma *köpfen* zeigt generell schlechtere Ergebnisse als für *Maus*. So weisen nur zwei der von den Probanden generierten bzw. verfassten Wörterbuchartikel ein Prädikat besser als V auf. Der beste KI-generierte Wörterbuchartikel (k1.2) wurde wiederum von dem Probanden mit der meisten Erfahrung in Lexikographie produziert, wobei dieser mit 51 % auch nur knapp die Mindestanforderung eines Wörterbuchartikels gemäß der von uns gesetzten Prädikate erfüllt. Im Folgenden wird auch dieser WbA exemplarisch ausgewertet:

Im Bereich des Formkommentars fehlen die Aussprache- sowie die Rechtschreibangabe; innerhalb der Grammatikangabe werden bezüglich des Verbparadigmas lediglich die Verbformen im Präsens sowie die Partizipien angegeben – andere Modi oder Tempora fehlen. Alle vorhandenen Angaben sind korrekt. Innerhalb des Semantikkomentars fehlen Pragmatik-, Wortbildungs- und Sprichwortangabe. Die angegebenen Polysemieangaben stimmen zwar in ihrer Anzahl mit denen im Artikel von Duden.de überein, weichen allerdings inhaltlich stark von diesem ab. Während die erste Bedeutungsangabe „1 jemandem den Kopf abschlagen“ korrekt ist, sind die Angaben „2 Entfernen des oberen Teils von etwas“ bzw. „3 jemandem die Führung nehmen“ fragwürdig bzw. schlicht falsch. Das zweite Semem wird etwa im Duden und im DWDS als übertragende Bedeutung zu 1 geführt; Semem 3 hingegen ist keine Bedeutung des Wortes *köpfen*. Historische bzw. dialektale Sememe wie „den oberen Teil eines Baumes absägen“ (vgl. Wörterbuchnetz: BDO, DRW, DWb) sowie moderne Bedeutungen wie „mit dem Kopf stoßen“ (Duden.de) bzw. „[Fußball] (jmd.) köpft (etw. [einen Ball]) mit dem Kopf weiterbefördern“ (DWDS) werden nicht generiert. Die Beispielaufgaben zu den Bedeutungen sind teilweise ebenfalls als unsinnig zu bezeichnen: „Er köpfte die Flasche mit dem Schwert.“ sowie „Der Kapitän köpfte die Entscheidung und lenkte das Schiff in den Hafen.“: Während erstere immerhin noch denkbar wäre, ist letztere reiner Nonsense. Dies setzt sich dementsprechend in den Kollokationsangaben etwa zu Semem 3 fort: „Entscheidung köpfen, Richtung köpfen, Initiative ergreifen“; letztere vermeintliche Kollokation beinhaltet nicht einmal das Wort *köpfen* und ist dementsprechend keine Kollokationsangabe – auch inhaltlich entbehrt sie jeder Grundlage. Hier wird ein weiterer Schwachpunkt der KI offensichtlich: Die generierten Kollokationsangaben beinhalten häufig wie bei Semem 1 keine Kollokationen, sondern stellen sinnverwandte Wörter „Henker“ oder Synonyme „enthaubern“ dar. Dieser Umstand wird von der KI auch wiedergegeben, indem weitere (redundante) Angaben wie Synonyme und „Andere verwandte Wörter“ gemacht werden, die diese Wörter ebenfalls beinhalten. Das indiziert wiederum, dass die KI den gesamten Artikel nicht redigiert, sondern nur von Wort zu Wort generiert. Die Richtigkeit der Inhalte der Etymologieangabe kann mittels entsprechender Wörterbücher (vgl. Wörterbuchnetz: Lexer, AWB) ebenfalls nicht verifiziert werden: „Mittelhochdeutsch köepfen, althochdeutsch chopfōn, verwandt mit mittelniederländisch copen ‚abschneiden‘“ (WbA k1.2). Die Schreibungen *käpfen* und *copen* sind nicht korrekt; auf Basis von lautgesetzlichen Entwicklungen und Schreibkonventionen wären (mhd.) *köpfen* sowie (mnl.)

koppen zu erwarten. Es ist daher davon auszugehen, dass es sich bei dieser Etymologieangabe um eine ChatGPT-spezifische Halluzination handelt. Hinsichtlich der Gesamtbewertung ähnelt dieser WbA dem von *Maus*, wobei er jedoch inhaltlich mehr Fehler aufweist.

5.2 Vergleich mit anderen Wörterbuchartikeln

Im Vergleich zum Experten-Wörterbuchartikel *Maus* auf Duden.de (Nr. 5.1 in Tab. 4, vgl. auch Abschnitt 4.3) schneiden alle Artikel der Probanden weitaus schlechter ab. In der Gesamtbewertung erreicht der Duden-Artikel 70 % (Prädikat III-), während die Artikel aus dem Schreibexperiment zwischen 59 % (Prädikat IV-) und 29 % (Prädikat V) rangieren und die Anforderungen mehrheitlich nicht erfüllen, der Durchschnittswert liegt bei 46 % (Prädikat V). Dies war auch zu erwarten, da eine kurze lexikographische Einführung die jahrzehntelange Erfahrung von professionellen Lexikographen nicht ersetzen kann. Bei den Unterschieden zu bspw. WbA 1.2 (mit KI, mit lexikographischer Kurzeinführung) fallen folgende Einzelheiten auf: Der Experten-Wörterbuchartikel erreicht bei der Bewertung der Gesamtstruktur die volle Punktzahl (18 von 18 Punkten = 100 % = Prädikat I+), und erzielt auch beim Formkommentar 95 % (30,5 von 32 Punkten = Prädikat I). Dagegen bleibt Artikel 1.2 vor allem hinsichtlich des Formkommentares (17 von 32 Punkten = 53 % = Prädikat V) weit hinter den Anforderungen zurück, da verschiedene Angaben fehlen bzw. nicht vollständig sind. Beim Semantikkommentar fällt zwar der Duden-Artikel hinsichtlich der Punkte (46,5 vs. 50,5) etwas ab. Dies ist aber vordergründig der Tatsache geschuldet, dass hier Kollokations- und Wortbildungsangaben fehlen. Bei den realisierten Kategorien erreicht der Experten-Artikel wiederum deutlich höhere Punktewerte.

Ein entsprechendes Bild zeigt sich bei *köpfen* (vgl. Tab. 5), wenngleich die Punktewerte hier bei sämtlichen Vergleichsgruppen (Probanden des Schreibexperiments, Duden-Artikel, KI-Expert-Artikel) geringer ausfallen als bei *Maus* (vgl. auch Abschnitt 5.1.2): Der Duden-Artikel erreicht hier in der Gesamtbewertung 66 % (wiederum Prädikat III-), die Artikel aus dem Schreibexperiment rangieren zwischen 55 % (wiederum Prädikat IV-) und 29 % (Prädikat V). Auch hier werden bei den meisten Ergebnissen aus dem Schreibexperiment die Anforderungen nicht erfüllt, der Durchschnittswert liegt mit 40 % (Prädikat V) noch niedriger. Auch die Detailauswertung spiegelt die Ergebnisse von *Maus* (vgl. Tab. 7).

Tab. 7: Ergebnisse des Schreibexperiments im Vergleich zu den Duden-Wörterbuchartikeln

	mit KI generierte WbA (Schreibexperiment)	ohne KI generierte WbA (Schreibexperiment)	Duden-WbA
Gesamte Mikrostruktur			
<i>Maus</i>	11.6/18 = 64 % = IV	13.0/18 = 72 % = III	18/18 = 100 % = I
<i>köpfen</i>	11.3/18 = 63 % = IV	11.9/18 = 66 % = III-	18/18 = 100 % = I
Formkommentar			
<i>Maus</i>	15.3/32 = 48 % = V	11.8/32 = 37 % = V	30.5/32 = 95 % = I-
<i>köpfen</i>	14.2/29 = 49 % = V	10.4/29 = 36 % = V	25.5/29 = 88 % = II
Semantikkommentar			
<i>Maus</i>	39.4/86 = 46 % = V	32.8/86 = 38 % = V	46.5/86 = 54 % = IV-
<i>köpfen</i>	28.4/82 = 35 % = V	27.1/82 = 33 % = V	42/82 = 51 % = IV-
Bedeutungsangabe			
<i>Maus</i>	11.6/16 = 72 % = III	12.5/16 = 78 % = III+	12.5/16 = 78 % = III+
<i>köpfen</i>	8.7/12 = 72 % = III	9.4/12 = 78 % = III+	11.5/12 = 96 % = I-

5.3 Auswertung des Schreibexperiments

In diesem Abschnitt wird das Schreibexperiment (vgl. Abschnitt 2) systematisch ausgewertet. Aus Platzgründen kann hier lediglich auf einige ausgewählte Aspekte eingegangen werden. Unser Fokus liegt auf Vergleichen der WbA aus verschiedenen, im Experiment angelegten Szenarien. Eine zentrale Rolle spielt bei den Gruppen mit KI-Nutzung die Auswertung des Promptings in Beziehung zur Artikelgenese. Die Berücksichtigung der Metadaten aus der Fragebogenerhebung bietet Aufschluss über die Auswirkung von KI-Erfahrung und/oder lexikographischen Vorkenntnissen der Probanden.

Als relevant erachten wir gruppen- und artikelorientierte Vergleiche (V): (a) zwischen Gruppen mit und ohne KI-Nutzung sowie (b) zwischen Gruppen mit und ohne lexikographische Einführung sowie (c) kombinierte Vergleiche. Bei gruppenorientierten Vergleichen kann die durchschnittliche Bewertung der jeweiligen Gruppen gegenübergestellt werden, bei artikelorientierten die WbA mit der jeweils besten Bewertung, wobei wir uns auf das Lemma *Maus* beschränken.

Tab. 8: systematische Auswertung des Schreibexperiments

	(a)	(b)	(c)
gruppen-orientiert	V1: <i>mit bzw. ohne KI</i> = I, II vs. III, IV	V3: <i>mit KI</i> <i>mit bzw. ohne Einführung</i> = I vs. II	V7: <i>mit bzw. ohne KI</i> <i>mit Einführung</i> = I vs. III
		V4: <i>ohne KI</i> <i>mit bzw. ohne Einführung</i> = III vs. IV.	V8: <i>mit bzw. ohne KI</i> <i>ohne Einführung</i> = II vs. IV
artikel-orientiert	V2: <i>mit bzw. ohne KI</i> = 1.2 vs. 3.1	V5: <i>mit KI</i> <i>mit bzw. ohne Einführung</i> = 1.2 vs. 2.2	V9: <i>mit bzw. ohne KI</i> <i>mit Einführung</i> = V2
		V6: <i>ohne KI</i> <i>mit bzw. ohne Einführung</i> = 3.1 vs. 4.2	V10: <i>mit bzw. ohne KI</i> <i>ohne Einführung</i> = 2.2 vs. 4.2

Die möglichen Vergleiche von Gruppen oder WbA ohne KI-Nutzung (V4, V6) bleiben im Folgenden unberücksichtigt.

(a) Wie bereits aus Tab. 7 ersichtlich wird, schneiden bei V1 die generierten Artikel in Form- und Semantikkommentar etwas besser, bei der Bewertung der gesamten Mikrostruktur etwas schlechter ab. Bei der Gesamtbewertung allerdings liegen die KI-Artikel mit 3/6 Artikeln (50 %) besser als Prädikat V und die Artikel ohne KI lediglich mit 1/4 (25 %) besser als Prädikat V; die Durchschnittswerte liegen bei 49 % vs. 42 %. Da dies bei so geringen Zahlen statistisch nicht aussagekräftig ist, lohnt sich ein Blick auf die Artikel-orientierten Vergleiche der WbA. So fällt bei V2 auf, dass beide WbA nur im unteren Bereich von Prädikat IV nicht weit voneinander liegen (80.5 vs. 76 / 136 Punkten); während sich WbA 3.1 in der Bewertung der gesamten Mikrostruktur etwas weiter vorne befindet (13 vs. 15), liegt er in Form- (17 vs. 13.4) und Semantikkommentar (50.5 vs. 47.5) leicht zurück. Für die Bewertungskategorie *Textstruktur* erhält WbA 3.1 als einziger WbA aller Studien-TN volle Punktzahl. Auch Polysemie- und der Beispielangabe erhalten volle Punktzahl, was beim Lemma *Maus* nur noch WbA 4.2 schafft. Etymologische oder pragmatische Informationen fehlen in den Belegen für die Gruppen ohne KI, sodass im Formkommentar hier potentiell 13 Punkte nicht

erlangt werden konnten, was auch bei 3.1 zutrifft. Die Kollokationsangaben fehlen in WbA 3.1 ebenfalls, selbst wenn solche aus den Belegen hätten herausgelesen werden können: *Maus huscht, süße Maus*. Beide Probanden in V2 verfügen laut Fragebogen bereits über Vorerfahrungen im Bereich Lexikographie, was ein Grund für das vergleichsweise insgesamt bessere Abschneiden beider WbA sein kann.

(b) Die KI-Artikel der Probanden mit lexikographischer Kurzeinführung bei V3 schneiden mit einem Durchschnitt von 53 % etwas besser ab als diejenigen der Gruppe ohne Einführung mit 44 %. Diese Diskrepanz erscheint angesichts des fehlenden Fachwissens der Gruppe II aber logisch. Auch hier lohnt ein Blick ins Detail: Nur marginal weiter auseinander als bei V2 liegen bei V5 die WbA 1.2 und 2.2 (80.5 vs. 71.5), wobei WbA 1.2 im Formkommentar lediglich ein wenig (17 vs. 15.5) und im Semantikkommentar etwas besser (50.5 vs. 40.5.) als WbA 2.2 abschneidet; in der Gesamtbewertung wird allerdings 2.2 mit 15.5 Punkten etwas besser als 1.2 mit 13 Punkten bewertet. Obschon der Proband im Chatverlauf ChatGPT nach Definition und Struktur von Wörterbuchartikeln fragt, kann das schlechtere Abschneiden von 2.2 im Semantikkommentar auf die fehlende Einführung zurückgeführt werden, da die Antwort von ChatGPT weitestgehend allgemein bleibt (s.u.) und nicht alle in der Einführung vorgestellten Angaben auflistet; der Proband war dementsprechend nicht vollständig über die Bestandteile von WbA informiert. Konkret weist WbA 2.2 somit weder Pragmatik-, noch Kollokations-, Wortbildung- oder Sprichwortangabe auf und verliert hier potentiell 35 Punkte; auch wenn Pragmatik- und Wortbildungsangabe auch bei 1.2 sehr schlecht bewertet werden, fehlt in diesem Artikel nur die Sprichwortangabe vollständig, so dass potentiell lediglich 10 Punkte verloren sind. Die Angaben im Fragebogen, dass 1.2 über „sehr viel Erfahrung mit Wörterbüchern“ verfügt und 2.2 hingegen nur mittlere, gibt hingegen keinen Aufschluss über das ähnliche Abschneiden beider Artikel. Der Chatverlauf von 1.2 offenbart allerdings, dass der Proband lediglich einmal um eine Verbesserung (siehe auch Abschnitt 5.4) gebeten hat (insgesamt 2 Prompts), während für 2.2 der Proband 13 Prompts bis zum Endergebnis brauchte. Hieraus kann spekuliert werden, dass der Proband mit großem Erfahrungsschatz sich zu sehr auf sein vorhandenes Wissen verlässt und dadurch möglicherweise auch nicht besonders motiviert war, die Aufgabe zu lösen.⁵ Bei WbA 2.2 hingegen hat der Proband so lange am Artikel gearbeitet, bis dieser zu seiner Zufriedenheit ausgefallen ist. Zwischen Anzahl der Prompts und Qualität der WbA besteht aber nicht zwingend ein Zusammenhang: so hat für 2.1 der Proband 22 Prompts eingegeben, erreicht in der Bewertung aber lediglich 40 %. Hinsichtlich eines Einflusses des lexikographischen Vorwissens kann generell festgestellt werden, dass bis auf zwei Ausnahmen alle Probanden lediglich über mittlere bis wenige Erfahrungen mit Wörterbüchern (sowie mit KI) verfügten. WbA 1.2 wurde von demjenigen Probanden generiert, der über „sehr viele Erfahrungen“ mit der Lexikographie verfügt; der Artikel eines Probanden mit selbsteingeschätzt hohem Vorwissen bezüglich KI (ebenfalls 2.1) bleibt mit 40 % dagegen weit hinter Artikel 1.2 zurück. Hier zeigt sich also kein positiver Effekt auf die Artikelqualität. Vorwissen im Bereich Lexikographie scheint aber durchaus Einfluss auf die Qualität der Artikel zu nehmen, da die Durchschnittswerte sowohl bei der Gruppe mit KI-Nutzung (53 vs. 44%) als auch bei den Gruppen ohne KI (50 vs. 35 %) relativ deutlich sind, obwohl die Probanden ohne Einführung ChatGPT nach Aufbau und Bestandteilen von WbA fragen konnten. Solche konstituierenden Fragen weisen die Chatverläufe von allen drei Probanden aus Gruppe II auf:

⁵ Der Proband gibt im Fragebogen zudem an, dass er solche Aufgaben schon früher lösen musste.

2.1: „Was sind Kriterien für gute Wörterbuchartikel? – Gute Wörterbuchartikel müssen so fesselnd sein, dass selbst die Wörter Angst haben, aus ihnen herauszuspringen!“

1. Klarheit und Präzision: Vermeide es, in einem Nebel von Wörtern zu ertrinken. Ein guter Artikel sollte klar und präzise sein, damit der Leser nicht das Gefühl hat, in einem Dschungel von Definitionen verloren zu gehen.

2. Relevanz: Der Artikel sollte genau das abdecken, wonach die Leser suchen, und keine Zeit damit verschwenden, über Dinge zu schwadronieren, die sie nicht interessieren.

3. Beispiele: Nichts verdeutlicht die Bedeutung eines Wortes besser als ein gutes Beispiel. Aber lass die Beispiele nicht so lang werden wie ein epischer Roman, sonst verlieren wir uns darin.

4. Etymologie: Ein bisschen Geschichtsstunde schadet nie! Die Etymologie eines Wortes kann faszinierend sein und dem Leser helfen, seine Bedeutung besser zu verstehen.

5. Synonyme und Antonyme: Ein bisschen Farbe im Wörterbuchleben schadet nie. Synonyme und Antonyme helfen dem Leser, das Wort in einen breiteren Kontext zu setzen.

6. Anwendungsbeispiele: Wie ein Handbuch für das Leben sollten gute Wörterbuchartikel praktische Anwendungsbeispiele bieten, damit der Leser weiß, wie er das Wort in der wilden Welt da draußen verwenden kann.

7. Zusätzliche Informationen: Füge interessante Fakten hinzu, die das Wort in einem neuen Licht erscheinen lassen. Ein paar lustige Anekdoten oder kulturelle Verweise können den Leser dazu bringen, das Wort mit anderen Augen zu betrachten.

Wenn du all diese Elemente zu einem Cocktail mixt, der sowohl informativ als auch unterhaltsam ist, hast du einen Wörterbuchartikel, der aus der Masse heraussticht!“

2.2: „Was bedeutet Wörterbuchartikel, geben [sic!] Sie eine [sic!] Beispiel, wie [sic!] so ein Wörterbuchartikel [sic!] aussieht? - Ein Wörterbuchartikel ist eine schriftliche Erklärung oder Definition eines bestimmten Wortes oder Begriffs. Es enthält normalerweise Informationen über die Bedeutung, die Aussprache, die Grammatik, die Herkunft und gegebenenfalls auch Beispiele für die Verwendung des Wortes in einem Satz oder Kontext. [...]“

2.3: „wie lange ist ein Wörterbuchartikel? – Die Länge eines Wörterbuchartikels kann stark variieren, abhängig von verschiedenen Faktoren wie dem Umfang des behandelten Begriffs, der Tiefe der Definitionen, der Anzahl der Beispiele, den verwendeten Quellen und dem Stil des jeweiligen Wörterbuchs. Einige Wörterbuchartikel können nur wenige Sätze oder Absätze umfassen, während andere mehrere Absätze oder sogar ganze Seiten benötigen können, um den Begriff umfassend zu erklären. Insgesamt gibt es also keine feste Regel für die Länge eines Wörterbuchartikels.“

Im Chatverlauf von 2.1 wirft nicht nur der Stil einige Fragen auf, auch bleiben sämtliche Antworten weitestgehend allgemein und tragen nur wenig zum mikrostrukturellen Aufbau eines WbA bei. Grundlegendes Fachwissen kann demnach Prompts positiv beeinflussen und somit als vorteilhaft für die Generierung von Fachtexten gelten.

(c) Bei V7 (mit lexikographischer Einführung) schneidet die Gruppe mit KI-Nutzung durchschnittlich nur wenig besser ab als die Gruppe der beiden Probanden ohne KI-Nutzung (53 vs. 50 %). Analog trifft dies bei V8 (ohne lexikographische Einführung) zu: die KI-Gruppe schneidet mit 44 % nur unwesentlich besser ab als die Gruppe ohne KI mit 35 %. Ohne V9 hier berücksichtigen zu müssen (= V2), sollen zum Abschluss der Artikel-orientierte Vergleich von WbA 2.2 (mit KI, ohne Einführung) und 4.2 (ohne KI, ohne Einführung) kurz verglichen werden. Der Unterschied in der Bewertung ist hier so groß, dass das Ergebnis als aussagekräftig betrachtet werden kann: WbA 2.2 wurde mit 71,5 Punkten mit deutlichem Abstand besser bewertet als WbA 4.2 mit 56 Punkten. Obwohl 4.2 zweimal volle Punktzahl in einer Bewertungskategorie erreicht (Polysemie- und Beispielangabe), bleibt der Artikel in Form- (15,5 vs. 9) und Semantikkommentar (40,5 vs. 34) weit hinter 2.2 zurück. Auch dies deutet darauf hin, dass die KI durch ihre Trainingsdaten zumindest Hinweise auf Struktur und Bestandteile eines WbA geben kann, wodurch die TN die Möglichkeit haben, die KI direkt zu fragen oder es von den generierten Artikeln abzulesen.

5.4 Berücksichtigung KI-spezifischer Fragestellungen

Mit dem folgenden Prompt konnte Wörterbuchartikel 6.1 in Tab. 4 generiert werden, der im Bewertungsrahmen mit 90 % bewertet wurde:⁶

Generate a monolingual dictionary entry in German for the word Maus; including the following categories: Lemma, Aussprache, Betonung, Wortart, Genus, Deklination, Rechtschreibung, Worttrennung, Häufigkeit, Bedeutung (4) and just in brackets (pragmatische Einordnung), Beispiele, Redewendungen und Sprichwörter, Kollokationen (sorted by meanings), Synonyme, Wortbildung and Etymologie. Please highlight the category names using bold type.

Ein genauerer Blick auf das Raster zeigt, dass die Bewertung für die Bereiche 1 und 2 nahezu bei voller Punktzahl liegen. Lediglich im Bereich des Semantikkommentars fehlen dem Artikel 11 Punkte, was u.a. auf die fehlenden Quellenangaben, Ungenauigkeiten bei den Kollokationsangaben („eine Mausfalle aufstellen“ = Kollokation zu *Mausefalle* und nicht *Maus*) sowie untypische Synonyme in der Synonymangabe zurückzuführen ist. Dennoch wird daraus ersichtlich, dass die KI mittels eines optimal abgestimmten Prompts durchaus in die Lage versetzt werden kann, einen adäquaten bzw. hochbewerteten Wörterbuchartikel zu erstellen, der hinsichtlich der Vollständigkeit und Vielfalt der lexikographischen Angaben sogar einen Experten-Artikel übertreffen kann. Zum Vergleich folgt hier das Prompting des besten KI-Artikels (1.2 in Tab. 4) aus dem Schreibexperiment:

ich muss Wörterbuchartikel zu den Stichwörtern das Nomen Maus und das verb köpfen verfassen. kannst du das für mich so machen, wie die Wörterbuchartikel in duden online aussehen. also zuerst information über Grammatik, z.B. Genus, Pluralbildung, Deklination usw. und dann Bedeutung mit jeweils 5 beispiele und auch kollokationen. Es wäre super, wenn du auch verwandte wörter darunter synonyme, antonyme, und andere verwandte wörter angibst. Die Wörterbuchartikel müssen auch mit Angaben über Frequenz und Herkunft versehen werden. Verfasse mir die Artikel für Maus und köpfen.

kannst du die informationen über die grammistik verbessern. nenne die überschriften so dass alle menschen die verstehen können. gib auch informationen darüber wie man genitiv bilden kann. kannst du versuchen die angaben zur herkunft für bestimmte bedeutung zu nennen und nicht für lemma

Der Vergleich legt folgende Schlussfolgerungen nahe: Die größten Unterschiede zwischen den beiden Prompts sind a) die Eingabesprache, b) Struktur der Sprache, c) Stil der Sprache und die Selbstkorrektur. Während sich der Experte bei der Eingabe des Englischen (Befehl) und Deutschen (Inhalte) bedient, nutzt der Proband nur die deutsche Sprache. Sein Dialogstil ist eher anredend und vermenschlichend („kannst du“) und weniger befehlend („Generate“). Der Prompt des Probanden ist in konzeptionell-mündlicher Form verfasst und kann strukturell als linear-assoziativ beschrieben werden; inhaltlich wird die prototypische Struktur eines Wörterbuchartikels nicht vorgegeben. Bei der zweiten, korrigierenden Eingabe wird weiter formuliert und nicht reformuliert. Hingegen hat sich bei der Erstellung des Experten-Artikels herausgestellt, dass die Erweiterung eines funktionierenden Prompts viel bessere Ergebnisse erbringt als eine korrigierende Neueingabe, da dann die KI möglicherweise die gesamte Textstruktur des Artikels verändert bzw. nur einzelne Angaben in korrigierter Form zeigt.

⁶ Dabei ist allerdings auch zu berücksichtigen, dass die Trainingsdaten von GPT-3.5 sehr wahrscheinlich mehrere Artikel zum Lemma *Maus* aus Online-Wörterbüchern umfassen.

Beim Lemma *köpfen* fällt auch der Experten-KI-WbA schlechter aus als der zum Lemma *Maus* (90 vs. 80 %), wobei die größten Unterschiede innerhalb des Semantikkommentars zu finden sind. Der Prompt wurde hinsichtlich der Wortart angepasst:

Generate a monolingual dictionary entry in German for the word "köpfen"; including the following categories: Lemma, Aussprache, Betonung, Wortart, Konjugation, Rechtschreibung, Worttrennung, Häufigkeit, Bedeutung and just in brackets (pragmatische Einordnung), Beispiele, Redewendungen und Sprichwörter, Kollokationen (sorted by meanings), Synonyme, Wortbildung and Etymologie. Please highlight the category names using bold type.

Anzumerken ist hier, dass bei der Generierung des Experten-KI-WbA im Januar 2025 ChatGPT 3.5 nicht mehr zur Verfügung stand und durch ChatGPT 4 ersetzt werden musste. Dies kann auch der Grund dafür sein, dass ChatGPT nun vier Sememe für die Bedeutungsangabe generiert, die mit Experten-WbA aus einschlägigen Wörterbüchern (vgl. 5.2.1) zu größeren Teilen übereinstimmen:

1. Jemandem oder einem Tier den Kopf abschlagen. (Alltagsgebrauch, Historisch)
2. Den oberen Teil einer Pflanze oder eines Objekts entfernen. (Gartenbau, Technik)
3. Den Korken einer Flasche mit einem Schlag entfernen. (Umgangssprache)
4. Einen Ball im Fußball mit dem Kopf spielen. (Sport, Fußball)

Semem 3 wird allerdings erneut als eigenständige Bedeutung angegeben, wohingegen es auf Duden.de oder im DWDS in übertragender Bedeutung von Semem 1 geführt wird (vgl. ebd.). Die Beispielangabe zu diesem Semem ist auch in dieser Version von ChatGPT inhaltlich unsinnig: „Er köpfte die Champagnerflasche mit einem Messer.“ Auch die Sprichwortangabe „Sein eigenes Grab schaufeln und sich selbst köpfen. [...]“ bleibt inhaltlich unklar. Während die Kollokationsangaben jetzt immerhin stets das Lemma selbst beinhalten und semantisch vertretbar sind, sind innerhalb der Synonymieangabe etwa bei Synonymen zu *Ball köpfen* weitere Fehler zu finden: „Mit dem Kopf spielen, nicken (Bedeutung 4)“. In der Wortbildungsangabe finden sich wiederum Angaben, die ein Synonym „Enthaupten“ darstellen oder falsch sind „Köpfstoß“. Innerhalb des Formkommentars beinhaltet die Grammatikangabe eine weitere Zeitform (Präteritum), bleibt aber auch hier den Rest des Paradigmas schuldig. Diese Lücke kann aber mittels eines diesbezüglich angepassten Prompts gefüllt werden.

Das schlechtere Abschneiden der WbA zum Lemma *köpfen* im Schreibexperiment ist eventuell auf die niedrige Frequenz des Wortes *köpfen* zurückzuführen. Es ist anzunehmen, dass die Trainingsdaten von ChatGPT dieses Wort weitaus weniger beinhalten als *Maus*. Sucht man Korpusbelege zu beiden Wörtern in den frei zugänglichen Gegenwartskorpora in DWDS, so finden sich zu *köpfen* in der Zeitspanne von 1897-2025 lediglich 3382 Belege, zu *Maus* jedoch 26.792; auch in den DWDS-Kernkorpora von 1900-2010 tritt *köpfen* lediglich 156- und *Maus* 2334-mal auf. Dieses Bild setzt sich in allen Korpora in DWDS fort und wird sich dementsprechend mutmaßlich auch in den nicht zugänglichen Trainingsdaten von ChatGPT widerspiegeln.

5.5 Gesamtauswertung

Abschließend lassen sich vier zentrale Schlussfolgerungen ziehen (hier exemplifiziert an den Einträgen 1.2 und k1.2):

- (1) Das Vorliegen einer Expertise zu speziellen Textsorten (Wörterbuchartikel) führt zu besseren Ergebnissen, wie der Vergleich der WbA 2.1-2.3/4.1/4.2 zeigt.
- (2) Gut ausgearbeitete Prompts, die lediglich erweitert nach Korrektur erneut eingestellt werden, erzeugen ebenfalls bessere Resultate, wie der Vergleich mit dem Expert-KI-WbA (6.1) erkennen lässt.
- (3) Beim Vergleich mit Experten-Wörterbuchartikeln (5.1, Duden) werden zudem strukturelle und qualitative Unterschiede offenbar.
- (4) Niederfrequente Wörter wie *köpfen* werden von der KI möglicherweise schlechter bearbeitet, da diese in den Trainingsdaten seltener vorkommen.

Aufgrund dieser Schlussfolgerungen können einige spezifische Vorschläge für die Nutzung von KI-Chatbots zur Generierung von speziellen Textsorten wie etwa Wörterbuchartikeln gemacht werden:

- a) Prompting: Prompts sollten komplex und in imperativer Form ohne Anredefloskeln formuliert werden; sie müssen strukturiert sein und spezifische Instruktionen zum sprachlichen Stil und zum Textaufbau enthalten. Sollte das Ergebnis nicht zufriedenstellend ausfallen, sollte der bereits verwendete Prompt angepasst und ggf. umformuliert erneut eingestellt werden, um eine neue Version des gewünschten Texts zu generieren. Bei Verbesserungsanweisungen im selben Thread ohne Wiederbenutzung des ersten angepassten Prompts können die Ergebnisse stark abweichen, da ChatGPT etwa die gewünschte Struktur des Textes wieder verändern kann. Bei Wiederverwendung des alten Prompts werden nur die hinzugefügten Änderungswünsche realisiert.
- b) Output: Wie auch immer die Prompts formuliert werden, empfiehlt es sich nicht, den generierten Ergebnissen unkritisch zu vertrauen. Eine Überprüfung der Ergebnisse durch den Abgleich mit bestehenden Expertentexten (in unserem Falle Wörterbuchartikel) ist unerlässlich, um die Korrektheit des generierten Inhalts zu bewerten. Dies gilt umso mehr für niederfrequente Wörter. Im Bildungskontext hat dies den zusätzlichen Vorteil, dass die spezifische Textkompetenz gefördert wird, wobei Nutzer vergleichen können, wie tatsächlich von Experten erstellte Texte inhaltlich und strukturell aussehen können.
- c) Das Schreibexperiment hat zudem gezeigt, dass für die erfolgreiche Generierung eines fundierten Textes ein hohes Maß an (lexikographischer) Expertise und Textkompetenz auf Seiten des Nutzers notwendig ist.

Nur vor diesem Hintergrund kann eine fundierte Bewertung des Outputs erfolgen, die es dem Nutzer ermöglicht, notwendige Überarbeitungen des Prompts vorzunehmen, um schlussendlich bessere Ergebnisse bei der KI-Verwendung zu erzielen.

6. Ausblick

Generative KI hat einen Punkt erreicht, an dem es mehr als nur einen kurzen, oberflächlichen Blick auf Texte braucht, um zu erkennen, dass sie möglicherweise nicht von einem Menschen verfasst wurden. Eine detaillierte (aber zeitaufwändige) qualitative Bewertung, die sich auf die inhaltlichen und semantischen Aspekte des Textes unter Berücksichtigung texttypspezifischer Kategorien konzentriert, liefert weitaus bessere Erkenntnisse und ist der

in der KI-Forschung verbreiteten oberflächlichen quantitativen Bewertung vorzuziehen. Das in dieser Arbeit vorgeschlagene Bewertungsraster ist ein erster Schritt zur praktischen Umsetzung einer solch gründlichen Evaluierung.

Methodisch steht das hier vorgestellte mehrstufige Bewertungsmodell noch vor einigen Herausforderungen und offenen Fragen. Es ist jedoch bereits ersichtlich, dass es eine relativ umfassende Bewertung der Wörterbuchmikrostruktur ermöglicht, sofern die beabsichtigten Funktionen des Wörterbuchs und die Wörterbuchadressaten berücksichtigt werden. Eines seiner wesentlichen Vorteile besteht darin, dass die verschiedenen Bewertungskategorien und einzelnen Kriterien getrennt aufgeführt und beurteilt werden können. Dies erlaubt einen weitaus detaillierteren und informativeren Vergleich von Wörterbucheinträgen als eine schlichte Gesamtbewertung. Natürlich muss das Bewertungsraster dynamisch an den jeweiligen Wörterbuchtyp und dessen Funktion angepasst werden. Dies wurde bereits für bilinguale chinesisch-deutsche Wörterbuchartikel praktiziert, wodurch verschiedene Fehlerarten, welche die KI produzieren kann, auf systematische Weise identifizieren werden konnten (vgl. Rink et al. 2024). Dies kann etwa den Weg für eine bessere Nutzung von KI-generierten Wörterbucheinträgen im Sprachunterricht ebnen.

Der hier vorgestellte mehrstufige Bewertungsmodell und insbesondere das zentrale Bewertungsraster können bei entsprechender Anpassung auch als allgemeines Modell zur Beurteilung und Optimierung KI-generierter Texte aller Art dienen.

7. Literatur

- Abel, A., Glaznieks, A., Linthe, M. & Wolfer, S. (Hrsg.) (2020): Textqualität im digitalen Zeitalter. *Themenheft Deutsche Sprache*, 48, 97-100. Online: (Letzer Zugriff: 04.03.2025).
- Arias-Arias, I., Domínguez Vázquez, M.J., Riveiro, C.V. (2024): Efficiency and Intelligence In: Lexicography and Artificial Intelligence: Can ChatGPT Recreate the Lexicographical Text Type?. *Lexikos*, 34(1), 51-76. Online: <https://doi.org/10.5788/34-1-1879> (Letzer Zugriff: 04.03.2025).
- Artstein, R., Poesio, M. (2008): Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4), 555–596.
- Becker-Mrotzek, M. (2014): Schreibleistungen bewerten und beurteilen. In: H. Feilke, T. Pohl (Hrsg.), *Deutschunterricht In: Theorie und Praxis. (DTP). Handbuch zur Didaktik der deutschen Sprache und Literatur In elf Bänden*. 4. Schriftlicher Sprachgebrauch - Texte verfassen, 501-513. Baltmannsweiler: Schneider Hohengehren.
- Becker-Mrotzek, M., Böttcher, I., Dreher, J. (2018): *Schreibkompetenz entwickeln und beurteilen*. Berlin: Cornelsen Verlag GmbH.
- Bielińska, M., Schierholz, S. J. (2017): Einleitung. M. Bielińska, S. J. Schierholz (Hrsg.): Wörterbuchkritik – Dictionary Criticism. 1-10. Berlin: Walter de Gruyter.
- Celikyilmaz, A., Clark, E., Gao, J. (2020): Evaluation of text generation: A survey. arXiv:2006.14799. Online: <https://doi.org/10.48550/arXiv.2006.14799> (Letzer Zugriff: 04.03.2025).
- de Schryver, G.-M. (2023): Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387. Online: <https://doi.org/10.1093/ijl/ecad021> (Letzer Zugriff: 04.03.2025).
- Dudenredaktion: *Duden – Das Bedeutungswörterbuch. Bedeutung und Gebrauch von rund 20 000 Wörtern der deutschen Gegenwartssprache*. Berlin: Cornelsen.
- Duden.de: *Duden – mehr als ein Wörterbuch*. Berlin: Cornelsen. Online: <https://www.duden.de/>. (Letzer Zugriff: 31.05.2024).

dwds.de: *DWDS – Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart*. Berlin-Brandenburgische Akademie der Wissenschaften. Online: <https://www.dwds.de/> (Letzer Zugriff: 31.05.2024).

Engelberg, S., Lemnitzer, L. (2009): *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

Engelberg, S., Storrer, A. (2016): Typologie von Internetwörterbüchern und -portalen. A. Klosa, C. Müller-Spitzer (Hrsg.): *Internetlexikografie. Ein Kompendium*. 31-63. Berlin: Walter de Gruyter.

Fritz, G. (2017): *Dynamische Texttheorie*. Gießener Elektronische Bibliothek. Online: <https://core.ac.uk/download/pdf/84117983.pdf> (Letzer Zugriff: 04.03.2025).

Kemmer, K. (2010): *Onlinewörterbücher in der Wörterbuchkritik. Ein Evaluationsraster mit 39 Beurteilungskriterien*. Mannheim. Online: <http://pub.ids-mannheim.de/laufend/opal/opal10-2.html>. (Letzer Zugriff: 04.03.2025).

Lew, R. (2023): ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10(1), 704. Online: <https://doi.org/10.1057/s41599-023-02119-6> (Letzer Zugriff: 04.03.2025).

Lew R, Szarowska A. (2017): Evaluating online bilingual dictionaries: The case of popular free English-Polish dictionaries. *ReCALL* 29(2), 138-159 Online: <https://doi.org/10.1017/S0958344016000252> (Letzer Zugriff: 04.03.2025).

Mayring: Gläser-Zikuda, M. (Hrsg.) (2008): *Die Praxis der Qualitativen Inhaltsanalyse*. Weinheim/Basel: Beltz.

Mayring, P., Hurst, A. (2017): Qualitative Inhaltsanalyse. In: L. Mikos, C. Wegener (Eds): *Qualitative Medienforschung. Ein Handbuch*, 494–502. Konstanz: UVK.

Müller, H.G., Fürstenberg, M. (2023): Der Sprachgebrauchsautomat. Die Funktionsweise von GPT und ihre Folgen für Germanistik und Deutschdidaktik. *Mitteilungen des Deutschen Germanistenverbandes*, 70(4), 327-345.

Neumann, A. (2017): Zugänge zur Bestimmung von Textqualität. In: M. Becker-Mrotzek, J. Grabowski, T. Steinhoff (Hrsg.), *Forschungshandbuch empirische Schreibdidaktik*, 203–219. Münster/New York: Waxmann.

Nussbaumer, M. (1991): Was Texte sind und wie sie sein sollen. Ansätze zu einer sprachwissenschaftlichen Begründung eines Kriterienrasters zur Beurteilung von schriftlichen Schülertexten. Tübingen: Max Niemeyer.

Nielsen, S. (2009): Reviewing printed and electronic dictionaries. A theoretical and practical framework. In: S. Nielsen, S. Tarp (Hrsg.), *Lexicography in the 21st Century. In honour of Henning Bergenholz*, 23-41. Amsterdam/Philadelphia: John Benjamins.

Pearsons, E., Nichols, W. (2013): Toward a Framework for Reviewing Online English Dictionaries. In: *Journal of the Dictionary Society of North America*, 34(1), 201-210.

Rink, C., Ganslmayer, C., & Evert, S. (2024): Towards a comprehensive method for evaluating and utilizing AI-generated bilingual lexicographical data in language learning using the example of Chinese as a foreign language. In: Ai Inoue, Naho Kawamoto, Makoto Sumiyoshi (Hrsg.): *Asian Lexicography – Merging cutting-edge and established approaches*, ASIALEX 2024. Proceedings, 133-142. Online: https://drive.google.com/file/d/11eiezuzidbVhrl-0Cr_k2NxgTg5CEUcP/view (Letzer Zugriff: 04.03.2025).

Ripfel, M. (1989): Wörterbuchkritik. Eine empirische Analyse von Wörterbuchrezensionen. Tübingen: Max Niemeyer.

Sieber, P., Nussbaumer, M. (1994): Texte analysieren mit dem Zürcher Textanalyseraster. In: Sieber: (Ed.). *Sprachfähigkeiten – Besser als ihr Ruf und nötiger denn je!*, 141-186. Aarau: Sauerländer.

Sieber, P., Nussbaumer, M. (1995): Über Textqualitäten reden lernen – z.B. anhand des ‘Zürcher Textanalyserasters’. In: *Diskussion Deutsch*, 141, 15-24.

Svensén, B. (2009): *A Handbook of Lexicography. The Theory and Practice of Dictionary-Making*. Cambridge University Press.

Tarp, S. (2017): Dictionary criticism and lexicographical function theory. In: M. Bielińska, S.J. Schierholz (Hrsg.), *Wörterbuchkritik – Dicitonary Criticism*, 113–132, Berlin: Walter de Gruyter.

Wiegand, H.E. (1998): Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. Berlin: Walter de Gruyter.

Wiegand, H.E. (1989): Arten von Mikrostrukturen im allgemeinen einsprachigen Wörterbuch. In: H.E. Wiegand, H. Steger (Hrsg.), *Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*, 556–568. Berlin: Walter de Gruyter.

Wiegand, H.E. (2010): Systematische Einführung. In: H.E. Wiegand, M. Beißwenger, R.H. Gouws, M. Kammerer, A. Storrer, W. Wolski. (Hrsg.), *Wörterbuch zur Lexikographie und Wörterbuchforschung. Dictionary of Lexicography and Dictionary Research*, Bd.1, 1–121. Berlin: Walter de Gruyter.

Wolfer, S., Abel, A., Glaznieks, A. & Linthe, M. (2023): Ausgangspunkte für die Betrachtung von Textqualität im digitalen Zeitalter. Modellbildung, empirische Fallstudien und methodologische Reflexionen. In: M. Meiler, M. Sieffkes, (Hrsg.), *Linguistische Methodenreflexion im Aufbruch. Beiträge zur aktuellen Diskussion im Schnittpunkt von Ethnographie und Digital Humanities, Multimodalität und Mixed Methods*, 181–202. Berlin: Walter de Gruyter.

woerterbuchnetz.de: *Trierer Initiative zum digitalen Verbund von Wörterbüchern*. Universität Trier - Trier Center for Digital Humanities (TCDH). Online: <https://www.woerterbuchnetz.de/> (Letzer Zugriff: 25.03.2025).