Stephanie Evert, Christine Ganslmayer, and Christian Rink

MULTI-LEVEL ANALYSIS AS A SYSTEMATIC APPROACH TO EVALUATING THE QUALITY OF AI-GENERATED DICTIONARY ENTRIES

Abstract In this paper, we report on our development of a multi-level 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, using dictionary articles created by human subjects with the help of ChatGPT as an example. The levels 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 level, we propose an elaborate evaluation grid that enables a fine-grained comparison of dictionary entries. While this grid has been developed for 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 Conversational AI; generated lexicographical data; dictionary criticism; evaluation methodology; semantic analysis; standard dictionaries

1. Introduction: LLMs and Dictionary Criticism

After the November 2022 release of ChatGPT, a freely accessible and easy-to-use conversational AI, text-generating AI products based on large language models (LLMs) have quickly become a part of daily life. The resulting transformation of communicative practice will have a profound impact on all areas of text production and reception, as well as knowledge society more broadly. This includes, of course, the field of lexicography, and several pilot studies have already explored possible applications of ChatGPT in lexicography; de Schryver (2023, p. 377) concludes:

> 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.

While most authors eventually come to the conclusion that ChatGPT cannot replace professional lexicographical work yet (e.g., Arias-Arias et al., 2024), its results are often astonishingly good at first glance, depending on the target language and the particular version of ChatGPT used. This impression was confirmed by an exploratory case study we carried out: Laypersons without prior lexicographical knowledge were not able to tell reliably whether a dictionary entry for a monolingual German general dictionary had been created by a human or by a human-guided AI. Reassuringly,

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

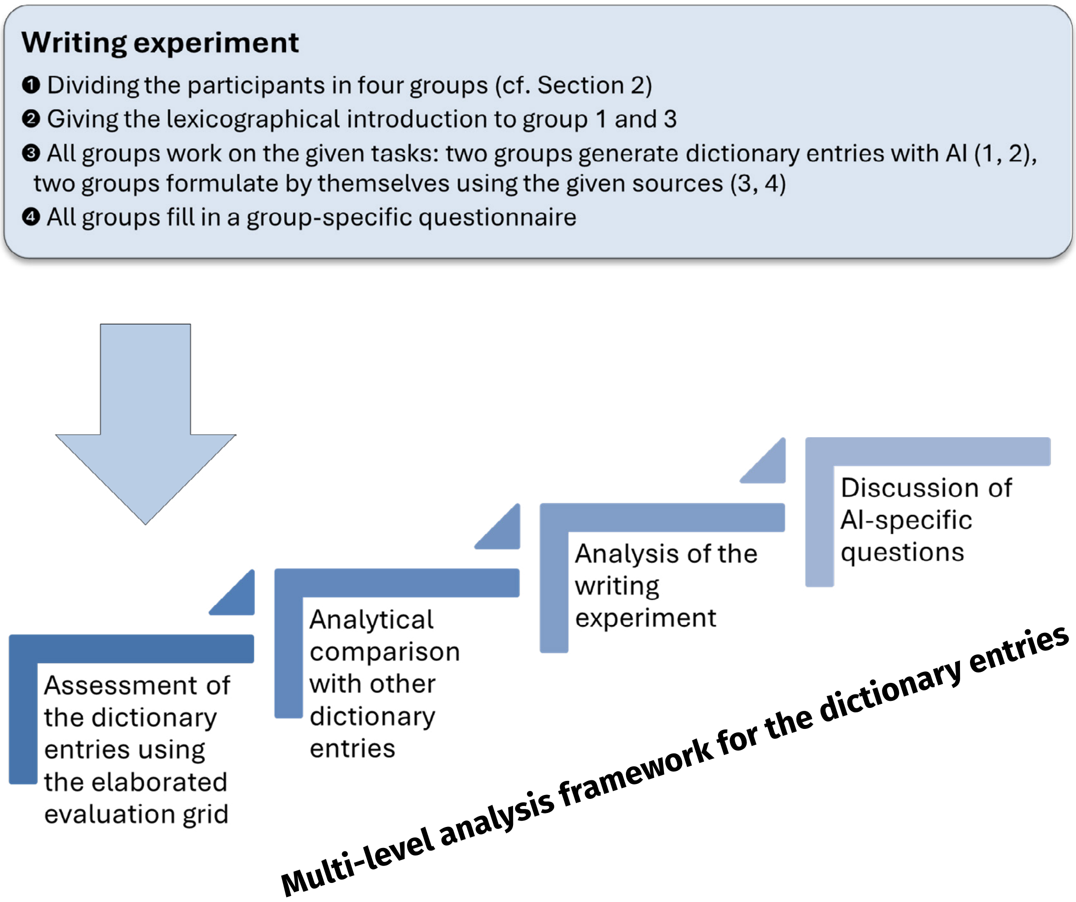
lexicographical experts performed much better at this task. In a similar vein, Lew (2023, pp. 8–9) finds that ChatGPT is at least partially capable of generating quality dictionary entries: whereas definitions seem to be "indistinguishable in quality from those written by highly trained human lexicographers", "examples [...] turned out not to be as impressive, and significantly worse than those crafted by professional human lexicographers."

This raises many questions, and the consequences for dictionaries as products, for their planning and use as well as their critical evaluation call for metalexicographical reflection (cf. Wiegand, 1998, pp. 79–80). Most people still have a very limited understanding of what LLMs are and what they can and cannot do. Our paper aims to shed some light on how good conversational AI has already become at (co-)authoring dictionary entries. Following de Schryver (2023, p. 377), we focus on a semi-automatic approach where AI-generated dictionary entries are reviewed and improved by human writers. The foremost and crucial question in this endeavour is how to achieve a detailed and comprehensive evaluation of the quality of dictionary entries. For this purpose, we developed a multi-level analysis framework for AI-generated dictionary entries and tested it in a writing experiment with non-expert writers assisted by AI.

Various approaches have been proposed in the field of dictionary criticism: Svensén (2009, pp. 482–485) describes several procedures for a dictionary evaluation: 1. analysis of the dictionary by one or more reviewers, 2. dictionary comparison, 3. use of dictionaries by test subjects with subsequent interview, 4. specially designed usage situations with specific user groups such as language learners. Tarp (2017) lists various criteria that relate to the functional type of a dictionary and its target users (see also Engelberg & Lemnitzer, 2009, pp. 190–222; Nielsen, 2009; Pearsons & Nichols, 2013). Further evaluation criteria include the quantity and quality of lexicographical information, their presentation, and the design of entries (cf. Tarp, 2017, p. 127). It is highly desirable for qualitative aspects to play a prominent role in the evaluation, even though it is more difficult to operationalise them than quantitative criteria.

Such criteria have to be developed into a concrete evaluation framework that can be applied successfully and adapted to different use cases. Dictionary criticism research usually applies ordinal scales for the evaluation of each aspect. Ripfel's (1989, p. 57) framework, for example, is based on an analysis of 736 dictionary reviews, from which she derived evaluation categories (e.g., quality of definitions) that are to be judged on a five-point scale (e.g., from "easily comprehensible/simple" to "incomprehensible/cumbersome"). Pearson & Nichols (2013) also propose a five-point scale, but address only quantitative aspects. Lew (2023) asked four experts to evaluate dictionary entries for 15 communication verbs generated by GPT-4, focusing on only three information categories (definitions, examples, entry as a whole) and rating them on a generic five-point scale (from bad to great). However, Lew does not provide precise evaluation criteria.

In our opinion, the evaluation of dictionary entries can be improved substantially if we adopt results and insights on text quality from German text linguistics (Nussbaumer, 1991; Sieber & Nussbaumer, 1994; Fritz, 2017; Abel et al., 2020), empirical writing didactics (Becker-Mrotzek, 2014) and qualitative media research and pedagogy (Mayring & Hurst, 2017). These fields consider an evaluation to be more reliable if, according to the analytical method, individual characteristics of text quality are determined according to text type, and then transformed into an evaluation grid of specific criteria to be judged. The overall evaluation is then based on awarding points for each criterion (cf. Neumann, 2017, pp. 208–211; Mayring & Hurst, 2017, p. 498; Becker-Mrotzek, 2014, pp. 507–510). Our paper outlines the development of such an evaluation grid for entries in a monolingual (German) general dictionary (Section 4). It is based on a general multi-level analysis framework (Section 3), which assesses texts (here: dictionary entries) according to various quantitative and qualitative criteria in comparison to existing reference texts of the same type. Our evaluation grid can be adapted to other functional types of dictionaries and is ideally suited for the evaluation of AI-generated dictionary entries, highlighting the specific strengths, weaknesses, and challenges of AI-assisted text production. Both the framework and the evaluation grid have been developed and tested in the context of a writing experiment, which is briefly described in Section 2. The application of our framework in Section 5 also specifically addresses the question how the specific limitations and challenges of text generation with an AI chatbot can be worked out and potential solutions can be sought. The flowchart in Fig. 1 gives a visual overview of the overall approach and methodology.



Writing experiment

❶ Dividing the participants in four groups (cf. Section 2)
❷ Giving the lexicographical introduction to group 1 and 3
❸ All groups work on the given tasks: two groups generate dictionary entries with AI (1, 2), two groups formulate by themselves using the given sources (3, 4)
❹ All groups fill in a group-specific questionnaire

Assessment of the dictionary entries using the elaborated evaluation grid
Analytical comparison with other dictionary entries
Analysis of the writing experiment
Discussion of AI-specific questions
Multi-level analysis framework for the dictionary entries

Fig. 1: Overall approach and methodology

2. Writing Experiment on Dictionary Entries

Our non-representative exploratory writing experiment was carried out in April 2024 in an advanced seminar in German linguistics at FAU Erlangen-Nürnberg. The task was to write dictionary entries with different levels of prior knowledge of the text type and with or without support from conversational AI. Six of the ten student participants (mainly from German and English studies) were asked to generate dictionary entries using ChatGPT (GPT-3.5), the other four were asked to write the entries independently without computational aid. The experiment was conducted in three phases, as outlined below: (1) To introduce prior knowledge of the text type as an independent variable, half of the students in each group were given a brief lexicographical introduction (15 minutes), presenting the components of dictionary microstructure and illustrating them with examples (entry *Pferḍ* [horse] in *Duden Bedeutungswörterbuch* [Duden Explanatory Dictionary] and *DWDS* [Digital Dictionary of the German Language]). We also introduced the modern corpus-based approach to lexicography, which integrates authentic evidence into the entries. (2) The instructions for the writing phase of 45 minutes were as follows:

Groups with AI:

> Create dictionary entries for the lemmata *Maus* [mouse] and *köpfen* [behead]. Use ChatGPT 3.5 and document all your attempts by copying both your input and the results (chatbot output) into a Word document (in chronological order). Only use the AI assistant. Internet sources must not be used under any circumstances so as not to distort the results of the study.

Groups without AI:

> Create dictionary entries in Word for the lemmata *Maus* [mouse] and *köpfen* [behead]. Use the provided examples as source material. AI and other Internet sources must not be used under any circumstances so as not to distort the results of the study.

Groups without AI were given six sheets of example sentences for each lemma to simulate the lexicographical process of identifying different word senses and subsenses, and to support the semantic analysis. Groups with AI had to rely on the AI training data exclusively. (3) After the writing phase, participants filled in questionnaires about their background and experiences from the experiment (20 minutes). A detailed evaluation of these questionnaires is beyond the scope of the present paper.

3. A Multi-Level Analysis Framework for AI-generated Dictionary Entries

Our multi-level analysis framework for AI-generated dictionary entries builds on and extends prior work by several authors: The general areas of the evaluation framework are inspired by Svensén (2009). Our evaluation categories for an entry in a monolingual German general dictionary are based on the well-known parameters of dictionary microstructure, in combination with the evaluation criteria of Nielsen (2009) and Tarp (2017). The evaluation grid takes up aspects of Ripfel (1989) and Pearsons & Nichols (2013). Evaluation categories and criteria were tested empirically on dictionary articles (e.g., from the writing experiment) and refined if necessary.

Our aim is to develop an analysis framework that can be adapted by experts for the evaluation of different types of (AI-generated) dictionaries, but can also be applied more widely to assess AI-generated dictionary entries and similar texts without having to resort to simplistic automated procedures (e.g., Celikyilmaz et al., 2020).

For the evaluation of an individual dictionary entry, only aspects of the microstructure, target user requirements, and dictionary type are relevant; other metalexicographical categories such as macro, medio or access structure do not apply. A strong focus was placed on semantic categories, complemented by general qualitative aspects such as aesthetic, content-related and linguistic appropriateness of the text, relative to its text type. An additional point of reference was provided by sample entries from existing dictionaries. The resulting categories and criteria form the basis of a finely differentiated evaluation grid (cf. Section 4), which enables judges to award points to each individual criterion.

Since we agree with de Schryver (2023, p. 377) that AI-generated dictionary entries will in most cases have to be checked and improved by humans, a well-founded evaluation of such entries can only be carried out in the context of a writing experiment that includes teams of AI and human as test subjects. The analysis of this experiment also has to address AI-specific questions, e.g., what influence prompts provided to the AI have on the quality of the output.

Our complete systematic approach to the evaluation of AI-generated dictionary entries thus encompasses four levels of increasing breadth and generality.

3.1 Assessment of Dictionary Entries via an Elaborate Evaluation Grid

The first level of our multi-level analysis framework evaluates the AI-generated dictionary entries using an elaborate evaluation grid, which contains quantitative and qualitative criteria for every information category in the form comment and semantic comment of a dictionary entry (such as pronunciation indication, polysemy, or examples). Our evaluation gird and the methodology for its development are presented in detail in Section 4. The grid can be applied equally well to dictionary entries in printed and online dictionaries written by human experts, provided that the necessary adaptations are made to align with the specific dictionary type.

3.2 Comparison With Other Dictionary Entries

The second level of our analysis framework mandates a comparison with other dictionary entries. In our case, entries from writing experiments (cf. Section 3.3) are evaluated according to Section 3.1 and compared to each other. Additional comparisons are made with existing dictionary entries written by professional lexicographers (here: from duden.de). These entries have been written by experts and should therefore be recognised as a reference, even if they do not achieve a perfect score according to our grid (e.g., because collocations are missing or uncommon (sub-)senses are included; see Section 4.3 for details).

3.3 Analysis of Writing Experiments

The next level of the evaluation framework calls for one or more writing experiments to be carried out and analysed. In most cases, test subjects are humans with different levels of expertise, as well as teams of AI and human. The analysis could focus on aspects such as comparing dictionary entries written with and without AI support, the effect of human post-editing on AI-generated dictionary entries, the test subjects' previous experience with dictionaries and AI tools, etc. Of course, writing experiments that target human-written texts exclusively are also possible.

3.4 AI-specific Considerations

Since our evaluation framework has been developed with AI-generated dictionary entries in mind, AI-specific questions should be considered in the fourth level. Typical examples include: (1) information categories that the AI cannot generate, (2) the prompt used to instruct the AI, (3) comparison of writing experiments with purely AI-generated dictionary entries, using a prompt optimised by a lexicographic expert, and (4) the discussion of AI-specific errors and deficiencies.

4. Developing an Elaborate Evaluation Grid

For the evaluative assessment of a dictionary entry, the entire entry and the lexicographical information provided are analysed in regard to the two main aspects of (a) quantity and (b) quality. Quantity refers to the presence of information as well as its scope and completeness. Quality refers to the accuracy and appropriateness of both content and language, as well as the authenticity, etc. of the information provided. Specific qualitative and quantitative criteria are established for each information category. Our main focus (partly depending on dictionary type) is on the qualitative criteria, which are thus more numerous and can be awarded more points. First of all, the targeted dictionary type has to be identified, indicating a possible information programme (*sensu* Engelberg & Lemnitzer, 2009, p. 25) of the microstructure, which in turn determines the information categories to be evaluated. Our dictionary comparison (according to Section 3.2) uses lemma *Maus* [mouse] from the online dictionary duden.de, which claims to offer "umfassende Informationen zu Rechtschreibung, Grammatik und Bedeutung eines Wortes" [comprehensive information on spelling, grammar, and meaning of a word] and "den richtigen Gebrauch sowie die Aussprache und Herkunft eines Wortes und [...] dessen Synonyme" [correct usage, pronunciation and origin of a word and [...] its synonyms] (duden.de).[1] Dictionaries with this amount of information are referred to as poly-informative dictionaries (*sensu* Wiegand, 2010, p. 85), which belong to the class of general dictionaries (*sensu* Engelberg & Storrer, 2016, p. 41) because they do not focus on specific information (unlike e.g., a dictionary of synonyms) and they are not limited in the selection of lemmas (unlike e.g., a dictionary of loanwords). Table 1 shows the main sections and information categories ("Angaben" *sensu* Wiegand, 1989, p. 468) to be analysed.

Table 1: Main sections + information categories of our evaluation grid

Overall microstructure	text structure / design, overall assessment
Comment on form	lemma, pronunciation indication, grammar indication, orthography indication, frequency
Comment on semantics	polysemy, pragmatics, meaning, examples, collocations, word formation, proverbs & idioms, synonymy, etymology

Specific quantitative and qualitative criteria then have to be developed individually for each category. For the overall structure and comment on form, they are mainly based on metalexicographical categories, from which specific questions can in turn be derived (cf. Tarp, 2017, pp. 122–125). The main focus of our evaluation is on the different information categories in the comment on semantics. The development of corresponding criteria is exemplified below for the category of meaning indication.

4.1 Exemplary Derivation of Quantitative and Qualitative Criteria

We illustrate the derivation of criteria for the evaluation grid on the example of the information category meaning indication. Quantitative criteria refer only to its presence and completeness, i.e., whether it is included at all and whether a definition has been provided for each (sub-)sense. Quantitative criteria are each awarded 0 or 1 point.

Qualitative criteria relate to use of standard language, redundancy, genus proximum, differentia specifica, linguistic correctness, accuracy of content and information density. Use of standard language requires meaning indications to be formulated without specialised terminology or, e.g., orientation towards a particular sociolect; redundancy should generally be avoided; the integration of genus proximum and differentia specifica aims to systematically differentiate words and position them within a semantic category; linguistic correctness is satisfied if there are no orthographic or grammatical errors; accuracy of content requires that the provided meaning indication is correct and complete; and information density refers to the length of the meaning indication, e.g., whether only a single word (meaning equivalent) is given or a complete semantic paraphrase. Qualitative criteria can be awarded up to 2 points, which gives them more weight in the overall evaluation and allows for more fine-grained judgments (awarding 1 point if a criterion is only partially satisfied).

[1] https://www.duden.de/woerterbuch (retrieved May 31st, 2024)

4.2 Evaluation Grid

Table 2 shows the complete evaluation grid with all categories and individual criteria. There is no room here for a full derivation of the evaluation criteria (as exemplified in Section 4.1), which will be detailed in a follow-up publication. In addition to providing a framework for evaluating the results of the writing experiment (Section 2), our evaluation grid is to be understood as a metalexicographical contribution to the systematic evaluation of dictionary entries. The grid can (and needs to) be adapted for application to different types of dictionaries: in bilingual dictionaries, e.g., other information categories are relevant and other criteria have to be formulated (less differentiation of the meaning indication, but inclusion of translation equivalents, greater importance of collocation information, etc.).

Table 2: Evaluation grid for the microstructure of a dictionary entry

Categories (Evaluation areas)	**Criteria** (max. 136): Quantity (0–1 each, total max. 30)	**Criteria** (max. 136): Quality (0–2 each, total max. 106)
Whole microstructure (max. 18)		
Textual structure / design (max. 8)	Non/typographical microstructural indicator, paragraphs, entry structure recognisable	Correctness microstructural indicator, understandability
Overall rating (max. 10)		Dictionary functions fulfilled, target users considered, conciseness of content, linguistic conciseness, use of general language
Comment on form (max. 32)		
Lemma (max. 3)	Availability	Linguistic correctness
Pronunciation (max. 6)	Availability, accentuation available	Correctness, accentuation correct
Grammar (max. 14)	Indication of part of speech available, genus specification available, inflection available, paradigm fully covered	Part of speech correct, genus indication correct, inflection indication correct, no redundancy, paradigm correct
Orthography (max. 6)	Availability, hyphenation available	Correctness, hyphenation correct
Frequency (max. 3)	Availability	Correctness
Comment on semantics (max. 86)		
Polysemy (max. 4)	Availability, completeness	Usualness
Pragmatics (max. 6)	Availability, completeness	Linguistic correctness, differentiation
Meanings (max. 16)	Availability, completeness	Use of general language, information density, linguistic correctness, accuracy of content, genus proximum, differentia specifica, no redundancy

Examples (max. 14)	Availability, completeness
	Clarity, linguistic correctness, accuracy of content, authenticity, appropriateness of content, sources provided
Collocations (max. 14)	Availability, amount
	Clarity, linguistic correctness, accuracy of content, authenticity, empirically verifiable, aligned with meanings
Proverbs & idioms (max. 5)	Availability
	Linguistic correctness, authenticity
Word formation (max. 10)	Availability, amount
	Linguistic correctness, accuracy of content, multiple types of word formation covered, usualness
Synonymy (max. 10)	Availability, completeness
	Linguistic correctness, content accurate, usualness, empirically verifiable
Etymology (max. 7)	Availability
	Linguistic correctness, accuracy of content, information density

The evaluation grid in Table 2 comprises a total of 136 points to be awarded, with 30 points for quantitative criteria and 106 points for qualitative criteria. A maximum of 18 points are awarded for general categories, 32 points for the comment on form, and 86 points for comment on semantics. For the overall evaluation of an entry, the total number of points is converted to a percentage and graded based on the ranges shown in Table 3, with 94–100% representing an excellent result (grade I) and anything below 50% considered inadequate for a dictionary (grade V).

Table 3: Grading scheme for the overall evaluation

98–100%	133–136 points	Grade I: *meets quantitative and qualitative requirements for the dictionary type excellently*
94–97%	128–132 points	
90–93%	122–127 points	Grade II: *meets quantitative and qualitative requirements for the dictionary type well (above par)*
85–89%	116–121 points	
80–84%	109–115 points	
76–79%	103–108 points	Grade III: *meets quantitative and qualitative requirements for the dictionary type satisfactorily*
71–75%	97–104 points	
66–70%	90–96 points	
61–65%	83–89 points	Grade IV: *meets quantitative and qualitative requirements for the dictionary type fairly (below par)*
50–60%	68–82 points	
0–49%	00-67 points	Grade V: *does not meet quantitative and qualitative requirements for the dictionary type*

4.3 Exemplary Assessment of the Lemma Maus (duden.de)

As an example for the application of our evaluation grid, we briefly describe the assessment of the lemma *Maus* [mouse] on duden.de; the result also serves as a point of reference for the dictionary comparison in the second level of our evaluation framework. This dictionary entry achieves an overall rating of 70% with only 95 out of 136 points and therefore has to be located in the lower range of Grade III. Of course, this rating cannot be generalised to the dictionary as a whole. For example, the lemma *Katze* [cat] as a control entry is at least in the upper range of Grade III (with 78%).

Due to space constraints, we can only describe the assessment of the category Meanings in detail. Meaning indications are provided in the duden.de entry on *Maus*, but are incomplete for sense 6 "weibliche Scham; Vulva" [female pubis, vulva] (lacking even a link to the lemma *Vulva*, as is provided for sense 2 "Kosewort" [term of endearment]). The meaning indications or links provided are correct in terms of language and content. They use general language, being neither technical nor colloquial. A classification of the lemma into superordinate semantic categories such as biological order is also provided: "kleines [graues] Nagetier" [small [grey] rodent], although differentiation within the category is too unspecific. The two sub-senses "4. a) Geld [money]" and "b) Euro, Mark o.Ä. [euro, mark, etc.]" appear to be redundant and over-differentiated. Information density varies: while meaning indications for senses 1 (rodent) and 5 (computer mouse) are relatively extensive, others such as sense 6 (female pubis, vulva) are only paraphrased with a synonymous word group. Usualness is evaluated in the polysemy category because it may differ between (sub-)senses: it is at least questionable whether sense 6 is indeed frequently used in general language.[2]

The poor overall score of the dictionary entry is partly due to the absence of nominal compounds (in the word formation category) and collocations, losing a total of 24 possible points. It is thus a crucial advantage of our evaluation grid that dictionary entries can be judged at a more fine-grained level (as will be shown in Table 4).

5. Application of the Multi-level Analysis Framework

In this section, we evaluate the dictionary entries from our writing experiment according to the multi-level analysis framework presented in Section 3.

5.1 Assessment of the Dictionary Entries

The dictionary entries from the writing experiment were independently evaluated by two experts on the basis of the evaluation grid shown in Table 2. As an added benefit, the grid allows us to validate the reliability of ratings much better than by comparing overall scores.[3] For this purpose, we calculate Cohen's kappa as a measure of the inter-annotator agreement between the two experts (Artstein & Poesio, 2008) and thus the reliability of the assessment. Kappa values usually range from 0 to 1, with scores above .8 considered as good agreement and above .67 as acceptable (Artstein

[2] According to the *Early New High German Dictionary* it was used in this sense by just a few authors during the 15th and 16th century. Other dictionaries – except for *Duden Universalwörterbuch* [General dictionary] (from 1996 onwards) and *Brockhaus-Wahrig Deutsches Wörterbuch* [German dictionary] (1982) – do not show this sense of *Maus*.

[3] For illustration, assume that both experts are lazy and award their points completely at random. Both overall scores would very likely fall somewhere between 43% and 57%, at first glance suggesting fairly good agreement. Only a closer look at the individual criteria in the grid would reveal that they only agree on approximately half of all points – exactly the value that is to be expected due to chance.

& Poesio, 2008, p. 576). To apply kappa to our data, we interpret the evaluation of qualitative criteria as a combination of two binary decisions: one point is awarded for minimum requirements, another point if the criterion is fully met. In this way, we obtain a satisfactory overall kappa of .73, with individual scores of .67 to .80 for the entries from the writing experiment, and as high as .86 and .87 for the two duden.de entries. As expected, the overall scores obtained by the two experts are much more similar than the kappa values might suggest.

For the sake of clarity, Table 4 only shows the assessment of one of the two experts for the various entries from the writing experiments (1.1–4.2), two expert entries (5.1, 5.2) and a completely AI-generated entry with expert prompting (6.1). In addition to the overall score, the table also shows how many points have been awarded in each general area and for individual categories. The colour coding is based on the colours in Table 3 and rates each category with one of the Grades I–V.

Table 4: Assessment of all dictionary entries according to the evaluation grid.

Source		writing experiments										Duden		AI
Entry		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
Using AI		y	Y	y	y	y	y	n	n	n	n	–	–	y
Lex. introduction		y	Y	y	n	n	n	y	y	n	n	–	–	–
Made by experts		n	N	n	n	n	n	n	n	n	n	y	y	y
%	100	46	58	57	40	50	40	59	46	32	43	70	78	90
Points	136	63	79	77	55	68	55	80	62	43	59	95	106	122
Whole microstructure	18	5	17	15	13	16	10	16	13	12	15	18	18	18
Textual structure / design	8	2	8	8	8	8	5	8	6	6	8	8	8	8
Overall rating	10	3	9	7	5	8	5	8	7	6	7	10	10	10
Comment on form	32	15	16	17	9	15	14	14	11	14	8	32	32	32
Lemma	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Pronunciation	6	1	0	3	3	2	3	0	2	0	0	6	6	6
Grammar	14	11	10	11	3	10	8	11	6	11	5	14	14	14
Orthography	6	0	0	0	0	0	0	0	0	0	0	6	6	6
Frequency	3	0	3	0	0	0	0	0	0	0	0	3	3	3
Comment on semantics	86	43	46	45	33	37	31	50	38	17	36	45	56	72
Polysemy	4	0	2	3	2	2	2	4	4	4	4	3	3	4
Pragmatics	6	0	0	5	0	0	0	0	0	0	5	6	5	6
Meaning	16	14	11	11	9	11	12	12	12	13	13	11	14	13
Examples	14	11	10	11	9	9	11	14	11	0	14	10	13	12
Collocations	14	0	9	0	0	0	0	0	0	0	0	0	0	12
Word formation	10	8	0	0	0	0	0	10	6	0	0	0	0	8
Proverbs & idioms	5	0	0	0	0	0	0	0	5	0	0	5	5	4
Synonyms	10	10	9	9	7	9	0	10	0	0	0	4	10	7
Etymology	7	0	5	6	6	6	6	0	0	0	0	6	6	6

A first, striking observation is that only four of the ten dictionary entries from the writing experiment meet the minimum quality requirements according to Table 3. With one exception, they are from subjects that have received a lexicographical introduction. Nevertheless, even these entries remain in the lower range of Grade IV.

Due to space constraints, we can only discuss the best AI-generated dictionary entry 1.2 from the writing experiment in detail. Comment on form: Neither spelling nor pronunciation are provided. While all expected grammatical information is included, the inflectional paradigm is incomplete. The duplicated specification of the plural form is redundant. Content, orthography, and grammar are correct. Comment on semantics: Categories word formation, pragmatics and proverbs & idioms are missing. The meaning indication only covers three of the four required (sub)-senses, whereas sense 3 in the entry is questionable: "scheuer Mensch" [shy person]. Meaning indications for the main senses *rodent* and *computer input device* are very short or completely absent. Semantic categorisation is present at least for sense 1, but uses technical terms: "Familie Muridae" [family Muridae]. Each sense includes both an example and several collocations. Those of senses 1 and 2 appear authentic and appropriate; one of the collocations for sense 3 ("Angst vor Mäusen haben" [being afraid of mice]) obviously refers to another sense. In general, some collocations have doubtful empirical support due to lack of salience, e.g., "Mäuse loswerden" [getting rid of mice]. Sources for the examples are never provided in AI-generated entries. Synonym information is available for senses 1 and 2. Etymology information is also provided but has low information density. Except for the aspects mentioned above, all information is correct in terms of content and language. Regarding general criteria, the entry contains microstructure indicators and other structural elements that give the impression of a dictionary entry. The entry is written clearly, fulfils the dictionary functions and takes target users into account.

5.2 Comparison With Other Dictionary Entries

All entries from the writing experiments achieve much lower scores than the expert dictionary entry from duden.de (5.1 in Table 4). This was to be expected, as a short lexicographical introduction cannot make up for the decades of experience of professional lexicographers. Table 4 highlights some interesting details, e.g., for the comparison with dictionary entry 1.2: Whereas the expert entry achieves full marks in the areas of whole microstructure and comment on form, entry 1.2 falls far short of the requirements, as various categories are missing or incomplete. Both entries have almost the same score for the comment on semantics (46 vs. 45 points), mainly due to the missing information mentioned above. Information they do provide (meaning, examples) is similar in quality. A closer look at the qualitative differences between expert and AI-assisted dictionary entries is beyond the scope of the present paper.

5.3 Analysis of the Writing Experiment

In our writing experiment, the test subjects also completed a questionnaire with questions about their prior experience with lexicography/dictionaries, AI, and the writing task. With two exceptions, all subjects had only moderate to little experience with both dictionaries and AI. Interestingly, entry 1.2 was generated by a subject who reported extensive experience with lexicography, while entry 2.1 from a subject with a high level of AI experience lags far behind at 40%. This suggests that AI experience may have little positive effect on the quality of an entry. There is no room here for a more detailed analysis of the questionnaires; establishing possible correlations between the quality of the generated texts and the level of knowledge of the subjects will require an experiment at a much larger scale. A comparison of the prompts used by the different groups is also beyond the scope of the paper.

5.4 AI-specific Considerations

In this Section, we discuss questions concerning the limitations of AI-assisted dictionary writing and the specific errors made by ChatGPT (1). We also address the question of prompting and, closely related, compare our findings to a purely AI-generated dictionary entry prompted by an expert (2).

(1) In general, GPT-3.5 is unable to generate audio snippets, as well as images or word clouds – features that are found in modern online dictionaries. Within the scope of our experiment, the limitations of AI are apparent mainly in the lack of comprehensiveness and the unclear structuring of the generated entries. Both issues can be solved by adequate prompting (see below). Comprehensiveness refers not only to the number of information categories, but also to the data within a category, e.g., missing forms of an inflectional paradigm, or missing word senses in the polysemy indication. Concerning the latter, ChatGPT often generates unclear, non-existent, rare or highly specialised senses such as “Scheue Person” [shy person] (entry 1.1) or “Biologisches Modell” [biological model] (entry 2.2). This can also happen for entries written by experts, though, as shown in Section 4.3. Similarly, meaning indications often use specialised technical terms, which is inappropriate for a general dictionary. This did not happen in the expert-generated entry, although it was not specifically forbidden. Further problems include giving not enough or too much information. The group without lexicographical introduction in particular created entries with redundant and unnecessary information, which contributed to disqualifying them as dictionary entries.

(2) The complex prompt below was developed by an expert to generate the optimized entry 6.1 in Table 4, which achieved an impressive overall score of 90% in the assessment:[4]

> Generate a monolingual dictionary entry in German for the word Maus [mouse]; including the following categories: Lemma [lemma], Aussprache [pronunciation], Betonung [accentuation], Wortart [part of speech], Genus [genus], Deklination [declination], Rechtschreibung [spelling], Worttrennung [hyphenation], Häufigkeit [frequency], Bedeutung [meaning] (4) and just in brackets (pragmatische Einordnung) [(pragmatic classification)], Beispiele

[4] One has to keep in mind, though, that the training data of GPT-3.5 very likely include several entries on the lemma *Maus* [mouse] from German online di8ctionaries.

> [examples], Redewendungen und Sprichwörter [idioms and proverbs], Kollokationen [collocations] (sorted by meanings), Synonyme [synonyms], Wortbildung und Etymologie [word formation and etymology]. Please highlight the category names using bold type.

A closer look at Table 4 shows that the entry achieves full marks in the first two areas. It loses 14 points in the comment on semantics, which is due *inter alia* to the missing sources for examples, inaccuracies in the collocational information ("eine Mausfalle aufstellen" [set up a mousetrap] is a collocation of the compound *Mausefalle* [mousetrap]) and atypical synonyms. Nevertheless, it is evident that AI can in fact generate a highly-rated dictionary entry, which even surpasses an expert entry in term of completeness and breadth of the lexicographic information, if it is provided with a suitable optimised prompt. Consider the prompt for entry 1.2 by comparison, which obtained a much lower score:

> 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 grammatik 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[5]

The biggest differences between the two prompts are (a) language of the prompt, (b) structure of the prompt, (c) linguistic style and self-corrections. Whereas the expert combines English (instructions) and German (content) in his prompt, the test subject only uses German. His style of dialogue anthropomorphises the AI, addressing it directly ("kannst du" [can you]) rather than giving commands ("Generate"). His prompt is written in conceptually spoken form and can be described as linear-associative. In terms of content, the prototypical structure of a dictionary entry is not provided to the AI. The second prompt continues the dialogue rather than reformulating the original prompt. Experience from optimising the expert AI entry shows that the extension of a functioning prompt gives much better results than asking the AI to rewrite and improve its first draft, contrary to popular recommendations.

[5] I have to write dictionary entries on the noun Maus [mouse] and the verb köpfen [behead]. can you do it for me like the dictionary entries in duden online. so first information about grammar, e.g., genus, pluralisation, declension etc. and then meaning with 5 examples each and also collocations. It would be great if you could also include related words, including synonyms, antonyms and other related words. The dictionary entries must also include information on frequency and origin. Write the entries for Maus and köpfen for me. | can you improve the information about the grammar. name the headings so that all people can understand them. also give information about how to form genitives. can you try to give information about the origin for specific meanings and not for the lemma [our translation]

5.5 Conclusions

For a final synthesis of the analysis, three main conclusions can be drawn (exemplified here for entry 1.2): (1) Expertise in relevant types of specialised texts often leads to better results, as the comparison with entries 2.1–2.3/4.1/4.2 shows. (2) Concise and targeted prompts, which are expanded in the correction phase, also produce better results, as the comparison with expert-prompted AI (6.1) shows. (3) Compared to expert dictionary entries (5.1), structural and qualitative differences are particularly noticeable. Based on these conclusions, we can give some specific recommendations for the use of AI chatbots in generating dictionary entries: (a) Prompting: We suggest to use complex prompts in the style of a command, containing all needed information in a structured manner, accompanied by specific instructions pertaining to the desired linguistic style and design. In the event that the results are deemed unsatisfactory, the entire prompt should be adjusted, reformulated, and used to generate a new version of the entry (rather than asking ChatGPT to revise the entry within the same thread); otherwise, the AI tends to retain the format of the last output and just adds or changes single pieces of information. (b) Output: Regardless of how prompts are crafted, we strongly recommend not to trust the generated data uncritically. A verification of the output against existing printed or online dictionaries is essential to assess the correctness of the generated content. In an educational context, it has the additional benefit of promoting dictionary literacy while demonstrating to users how actual dictionaries produced by experts function. (c) Our experiments indicate that using AI chatbots to create a dictionary entry can only be successful if the human user has a high level of lexicographic knowledge and text competence. Only against this background, a well-founded evaluation of the output can take place, which enables the user to make necessary revisions and improve prompts.

6. Outlook

Generative AI has reached a point where it takes more than just a brief, superficial glance at texts to recognise that they may not have been written by a human. A detailed (but time-consuming) qualitative evaluation that concentrates on the content and semantic aspects of the text, in terms of categories specific to the text type, provides much better insights, and is to be preferred to the superficial quantitative evaluation common in AI research. The elaborate evaluation grid proposed in this paper is a first step towards putting such thorough evaluation into practice.

Methodologically, our multi-level analysis framework still poses many challenges and unanswered questions. It is already clear, though, that it enables a relatively comprehensive evaluation of dictionary microstructure if the intended functions of the dictionary and its target users are taken into account. One of its key advantages is that the different evaluation categories and individual criteria are listed and assessed separately. This makes for a much more detailed and informative comparison of dictionary entries than mere overall evaluation scores. Of course, the evaluation grid needs to be dynamically adapted to the respective dictionary type and function. We have already done this for bilingual Chinese-German dictionary entries (Rink et al., 2024), which enabled us to identify different types of errors made by the AI in a highly systematic manner. This paves the way to a better use of AI-generated dictionary entries in language teaching and learning.

We believe that our multi-level analysis framework and evaluation grid can serve as a general model for assessing and improving AI-generated texts of all types.

References

Abel, A., Glaznieks, A., Linthe, M., & Wolfer. S. (Eds.) (2020). Textqualität im digitalen Zeitalter. *Themenheft Deutsche Sprache*, *48*, 97–100. https://doi.org/10.37307/j.1868-775X.2020.02

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. https://doi.org/10.5788/34-1-1879

Artstein, R., & Poesio, M. (2008). Survey entry: 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 (Eds.), *Deutschunterricht in Theorie und Praxis. (DTP). Handbuch zur Didaktik der deutschen Sprache und Literatur in elf Bänden. 4. Schriftlicher Sprachgebrauch – Texte verfassen* (pp. 501–513). Baltmannsweiler.

Celikyilmaz, A., Clark, E., & Gao, J. (2020). Evaluation of text generation: A survey. arXiv:2006.14799. https://doi.org/10.48550/arXiv.2006.14799

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. https://doi.org/10.1093/ijl/ecad021

duden.de: *Duden – mehr als ein Wörterbuch*. Cornelsen Verlag GmbH. Retrieved May 31st, 2024, from https://www.duden.de/

Engelberg, S., & Lemnitzer, L. (2009). *Lexikographie und Wörterbuchbenutzung*. Stauffenberg.

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

Fritz, G. (2017). *Dynamische Texttheorie*. Gießener Elektronische Bibliothek. https://core.ac.uk/download/pdf/84117983.pdf

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

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

Neumann, A. (2017). Zugänge zur Bestimmung von Textqualität. In M. Becker-Mrotzek, J.

Grabowski, & T. Steinhoff (Eds.), *Forschungshandbuch empirische Schreibdidaktik* (pp. 203–219). Waxmann.

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

Nielsen, S. (2009). Reviewing printed and electronic dictionaries. A theoretical and practical framework. In S. Nielse, & S. Tarp (Eds.), *Lexicography in the 21st Century. In honour of Henning Bergenholtz.* (pp. 23–41). 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), pp. 201–210.

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

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. (ASIALEX. 2024. Proceedings), in print.

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

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 (Eds.), *Wörterbuchkritik – Dictionary Criticism.* (pp. 113–132). Walter de Gruyter.

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

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

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

Contact information

Stephanie Evert
Friedrich-Alexander-Universität Erlangen-Nürnberg
stephanie.evert@fau.de

Christine Ganslmayer
Friedrich-Alexander-Universität Erlangen-Nürnberg
christine.ganslmayer@fau.de

Christian Rink
Friedrich-Alexander-Universität Erlangen-Nürnberg
christian.rink@fau.de