

# 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

Christian Rink<sup>1</sup>, Christine Ganslmayer<sup>2</sup> and Stephanie Evert<sup>3</sup>

<sup>1</sup> Friedrich-Alexander-Universität Erlangen-Nürnberg, christian.rink@fau.de

<sup>2</sup> Friedrich-Alexander-Universität Erlangen-Nürnberg, christine.ganslmayer@fau.de

<sup>3</sup> Friedrich-Alexander-Universität Erlangen-Nürnberg, stephanie.evert@fau.de

## Abstract

In this paper, we take an analytical and a practical approach to the topic of AI-generated lexicographical data. In the context of a language-learning environment, we want to take a closer look at the quality of bilingual lexicographical data (Chinese-German) produced by generative large language models (LLMs). Our aim is to create a scientifically supported and comprehensive method that includes effective prompting strategies, an objective evaluation of the generated lexicographical data, and learner strategies for the use of conversational AI. The focus of our paper will be on the evaluation method, for which we use a multi-level analysis based on lexicographical standards, in comparison with existing lexicographical resources made by experts, the consideration of the findings of a didactical experiment, as well as a discussion about AI specific questions, and on the proposal of a didactical concept, which introduces a reasonable use of conversational AI chatbots in a language-learning environment. The results of this study are expected to give further insight into the functionality of generative LLMs in lexicographical settings, to provide a critical framework for the evaluation of AI-generated lexicographical data, and offer perspectives on how it can be used as a valuable lexicographical tool in language-learning.

**Keywords:** lexicographical data generated by conversational AI, evaluation methods, bilingual lexicography, Chinese as a foreign language, didactical concepts

## 1. Introduction

In this paper, we will concentrate on the evaluation of AI-generated lexicographical data, and a didactical concept, which will introduce a certain way of using AI for language learning. In order to achieve that, we will utilize the Multi-level Evaluation Framework for AI-generated dictionary entries introduced in Evert et al. (2024), which includes a report about a didactical experiment carried out in a Chinese class, the evaluation of AI-generated dictionary entries using our evaluation grid adjusted for bilingual dictionary entries, a comparison with existing online dictionaries, and a discussion about certain AI-specific aspects. Moreover, we will introduce a didactical concept, which considers the findings of the Multi-level Evaluation Framework. This concept is to fulfill three premises: a) using conversational AI for language learning effectively, b) using the errors made by the AI as an asset, and 3) providing students with dictionary literacy.

In Section 2, we will report about the didactical experiment, Section 3 will outline the findings of the Multi-level Evaluation Framework for AI-generated dictionary entries. The proposal of the didactical concept will be shown in Section 4.

## 2. Didactical experiment

In this Section, we will display and discuss a non-representative, exploratory didactical experiment, which was carried out in April 2024 in a Chinese course for adults held by the German adult education center (Volkshochschule) in Erlangen, Germany. The 10 participants were all adults, 7/10 of them aged 60 to 80 years and three of them 25 to 40. The language level of the course content is HSK 4<sup>1</sup>, but the actual language level of the participants might range just around HSK 3 (A2-B1) level. The course participants were divided into two groups; both groups had to work on the same task, which consisted of the translation of two sentences, the first of which was a translation into German (L1) and the second into Chinese (L2). The tasks were as follows:

- 1) Please translate the following sentence into German using the added dictionary entries pp. 2-6: „在业余时间, 叔叔很喜欢收藏各种古董玩意儿, 比如老画和漂亮的玉器。” [In his spare time, my uncle likes to collect various antiques, such as old paintings and beautiful pieces of jade.]
- 2) Please translate the following sentence into Chinese using the added dictionary entries pp. 7-9: I love watching Japanese anime at the weekend, especially stories about friendship and adventure.

For both tasks, dictionary entries were provided for almost every word used in the sentences; just very simple words of lower HSK levels such as 很 [very], 周末 [weekend], 故事 [story], 是 [to be] etc. were omitted. The dictionary entries of group 1 were entries taken from the HanDeDict online dictionary. Group 2 was provided with the entries generated by the Chinese AI Tongyi qianwen discussed in Section 3. For organizational reasons all dictionary entries were printed on paper; the participants were not supposed to use other tools (such as printed dictionaries or the internet) and

thus could not work with the AI directly. Furthermore, the group using the entries of the online dictionary was also provided with all entries the online dictionary showed after the given search to simulate a situation in which participants would use a device with internet access. After 30 minutes, the participants were asked to complete a questionnaire with questions about the task and its completion. The analysis of the experiment will be displayed in Section 3.2.4 as part of the Multi-level Evaluation Framework.

### 3. Evaluation of AI-generated dictionary entries

We consider the quality of the (generated) lexicographical data to be crucial for the learning effect that language learners get from the information provided. For the purpose of this study, dictionary entries generated by the Chinese AI Tongyi Qianwen (通义千问) developed by Alibaba were used on the basis of our assumption that a Chinese conversational AI is trained with more Chinese data than other conversational AI tools. The theoretical foundation of the Multi-level Evaluation Framework for AI-generated dictionary entries can be found in Evert et al. (2024) where the framework was firstly utilized for the microstructures of AI-generated monolingual dictionary entries of German. In the following, the framework will be shortly introduced and adjusted for the microstructure of a bilingual dictionary.

#### 3.1 Multi-level evaluation framework for AI-generated dictionary entries

The Multi-level Evaluation Framework for AI-generated dictionary entries consist of four main parts<sup>2</sup>: 1) the assessment of individual dictionary entries (by one or more reviewers) on the basis of a detailed evaluation grid; 2) the discussion of AI-specific questions; 3) the analytical comparison with other dictionary entries made by experts, and 4) the analysis of the experiment (results of the writing tasks, questionnaires filled in by subjects).

Under 1) the original evaluation grid was used containing three main sections (i.e. overall microstructure, comment on the form, comment on the semantics) and several information categories per section, which were evaluated concerning quantitative and qualitative criteria such as the existence of a definition phrase (quantity) and its linguistic correctness (quality). The main sections will also be used in the evaluation grid for this study, but the information categories as well as the criteria will have to be adjusted for a specific bilingual dictionary. In general, different types of dictionaries with their respective addressees need different criteria for an evaluation; therefore, the information categories and the specific criteria in our evaluation grid can be exchanged dynamically according to the dictionary type. Whereas the focus in the experiment in Evert et al. (2024) was on the definition phrase of a monolingual dictionary of German, we will focus here on the categories of the equivalent, the example, and collocations in a bilingual dictionary for German learners of Chinese. The information categories for the overall microstructure section (a) will be the text structure / text design, and a general assessment. The categories for the comment on the form section (b) are the lemma, and the indication of the pronunciation. The comment on the semantics section (c) contains the equivalent, the definition phrase, examples, and collocations. The choice of the information categories of (b) and (c) are based on a study conducted by Zhang (2022, pp. 170-178), in which 379 German learners of Chinese were asked about the features a learner's dictionary of Chinese-German should have. Features that cannot be provided by AI are not subject of the evaluation such as the stroke order of a Chinese character or the indication of the radical.

In particular, the criteria for every information category will be an evaluation of the quality of the given feature, because the prompting for all entries was identical. Therefore, the quantity assessment is of little importance since all entries (should) consist of the same items. Table 1 shows the criteria developed for the main sections of all information categories, which have been slightly adjusted for the entries in German-Chinese, which is not shown in Table 1:

Table 1. Evaluation grid for the microstructure of a dictionary entry

Table 1: Evaluation grid for the microstructure of a thematic entry		
Main sections	Criteria (max. 60)	Quantity (0–1 each, total max. 4)
Information category		Quality (0–2 each, total max. 56)
Overall microstructure (max. 12)		
Textual structure / design (max. 6)	Non/typographical microstructural indicator, paragraphs, entry structure recognisable	
	Correctness microstructural indicator	
General assessment (max. 6)	Overall coherence, overall use of a learner-friendly language, no overall redundancy	
Comment on the form (max. 6)		
Lemma (max. 2)	Linguistic correctness	
Pronunciation (max. 4)	Correctness of pinyin, tones	
Comment on the semantics (max. 42)		
Equivalents (max. 4)	Linguistic correctness, accuracy of content	
Meanings (max. 6)	Linguistic correctness German, accuracy of content, clarity	

Examples (max. 16)	Linguistic correctness of German / Chinese characters / Chinese pronunciation, accuracy of content in Chinese, authenticity in Chinese, appropriateness of the content in Chinese, accuracy of the German translation, clarity of the German translation
Collocations (max. 16)	Linguistic correctness of German / Chinese characters / Chinese pronunciation, accuracy of the content in Chinese, authenticity in Chinese, appropriateness of the content in Chinese, accuracy of the German translation, clarity of the German translation

The evaluation grid in Table 1 comprises a total of 60 points to be awarded, with 4 points for quantitative criteria and 56 points for qualitative criteria. A maximum of 12 points can be awarded for main section (a), 6 points for main section (b), and 42 points for main section (c). The intended dictionary type is a bilingual learner's dictionary of Chinese-German for German users, for which all criteria shown in Table 1 have been optimized.<sup>3</sup> Quantitative criteria were awarded 0 or 1 points, qualitative criteria 0-2 points. For the overall evaluation of an entry, the total number of points is converted to a percentage and a grade based on the ranges shown in Table 2, with 94–100% representing an excellent result (Grade I) and anything below 50% being considered inadequate for a dictionary (Grade V).

Table 2. Grading scheme for the overall evaluation

98–100%	59–60 points	Grade I: <i>meets quantitative and qualitative requirements for the intended dictionary type excellently</i>
94–97%	56–58 points	
90–93%	54–55 points	Grade II: <i>meets the quantitative and qualitative requirements for the intended dictionary type well (above par)</i>
85–89%	51–53 points	
80–84%	48–50 points	
76–79%	46–47 points	Grade III: <i>meets the quantitative and qualitative requirements for the intended dictionary type satisfactorily</i>
71–75%	43–45 points	
66–70%	40–42 points	
61–65%	37–39 points	Grade IV: <i>meets the quantitative and qualitative requirements for the intended dictionary type fairly (below par)</i>
50–60%	30–36 points	
0–49%	00–29 points	Grade V: <i>does not meet the quantitative and qualitative requirements of the intended dictionary type</i>

AI-specific questions will be discussed under 2) (cf. 3.2.2). Under 3), a general comparison with the entries of existing dictionary entries will follow (cf. 3.2.3), whereas these entries will not be evaluated on the basis of the grid, because firstly, no online learner's dictionaries for Chinese / German are available in general, and secondly, the existing online dictionaries for this pair of languages very much lack information categories in main section (3), what makes a systematic comparison using the evaluation grid nearly impossible. Under 4), the experiments and test persons will be considered. (cf. 3.2.4).

### 3.2 Application of the multi-level evaluation framework

In this chapter, we will evaluate the dictionary entries according to the Multi-level Evaluation Framework presented in Section 3.1. The errors produced by the AI are particularly interesting, since they will be the basis of our assessments and suggestions for the proposal of the didactical concept (cf. Section 4).

#### 3.2.1 Assessment of the dictionary entries

The dictionary entries used in the experiment were evaluated together by two experts on the basis of the evaluation grid shown in Table 1. Both evaluators have expertise in lexicography; one of the experts' mother tongues is Chinese, the other's is German. This was arranged to ensure that the two languages could be evaluated adequately. Table 3 shows the results of the evaluation:

Table 3. Assessment of all dictionary entries according to the evaluation grid

Chinese-German German-Chinese	Entry no.	default	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	2.1	2.2	2.3	2.4	2.5
%		100	77	78	92	92	90	77	70	82	88	85	72	68	90
Points		60	46	47	55	55	54	46	42	49	53	51	43	41	54
Overall microstructure (max. 12)		12	8	8	9	9	9	8	9	6	8	10	6	7	8
Text structure / design (max. 6)		6	6	6	6	6	6	6	6	6	6	6	6	6	6
Typographical microstructural indicator (MSI)		1	1	1	1	1	1	1	1	1	1	1	1	1	1
Non-typographical microstructural indicator		1	1	1	1	1	1	1	1	1	1	1	1	1	1
Paragraphs		1	1	1	1	1	1	1	1	1	1	1	1	1	1

Entry structure recognisable	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Correctness MSI	2	2	2	2	2	2	2	2	2	2	2	2	2	2
<b>Overall assessment (max. 6)</b>	6	2	2	3	3	3	2	3	0	2	4	0	1	2
Overall coherence	2	1	1	2	2	2	1	1	0	1	1	0	0	1
Overall use of learner-friendly language	2	1	1	1	1	1	1	2	0	1	2	0	1	1
No overall redundancy	2	0	0	0	0	0	0	0	0	0	1	0	0	0
Comment on the form (max. 6)	6	4	6	6	6	6	6	6	6	6	6	6	6	6
<b>Lemma (max. 2)</b>	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Linguistic correctness	2	2	2	2	2	2	2	2	2	2	2	2	2	2
<b>Pronunciation (max. 4)</b>	4	2	4	4	4	4	4	4	4	4	4	4	4	4
Correctness of pinyin letters	2	1	2	2	2	2	2	2	2	2	2	2	2	2
Correctness of pinyin tones	2	1	2	2	2	2	2	2	2	2	2	2	2	2
Comment on the semantics (max. 42)	42	34	33	40	40	39	32	27	37	39	35	31	28	40
<b>Equivalents (max. 4)</b>	4	2	4	4	4	4	3	4	3	4	4	4	4	4
Linguistic correctness	2	1	2	2	2	2	2	2	2	2	2	2	2	2
Accuracy of content	2	1	2	2	2	2	1	2	1	2	2	2	2	2
<b>Meanings (max. 6)</b>	6	4	3	4	5	6	2	4	4	6	5	6	6	6
Linguistic correctness of German	2	2	2	2	2	2	2	2	1	2	2	2	2	2
Accuracy of the content	2	1	0	1	2	2	0	1	1	2	2	2	2	2
Clarity	2	1	1	1	1	2	0	1	2	2	1	2	2	2
<b>Examples (max. 16)</b>	16	16	15	16	15	14	11	16	15	15	13	12	14	15
Linguistic correctness of German	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Linguistic correctness of Chinese characters	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Linguistic correctness of Chinese pronunciation	2	2	1	2	1	1	1	2	1	2	1	0	1	2
Accuracy of the content in Chinese	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Authenticity of Chinese	2	2	2	2	2	2	1	2	2	2	2	2	2	2
Appropriateness of the content in Chinese	2	2	2	2	2	2	1	2	2	2	2	0	2	2
Accuracy of the translation in German	2	2	2	2	2	2	1	2	2	2	1	2	2	2
Clarity of the translation in German	2	2	2	2	2	1	1	2	2	1	1	2	1	1
<b>Collocations (max. 16)</b>	16	12	11	16	16	15	16	3	15	14	13	9	4	15
Linguistic correctness of German	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Linguistic correctness of Chinese characters	2	2	2	2	2	2	2	0	2	2	2	2	0	2
Linguistic correctness of Chinese pronunciation	2	2	2	2	2	2	2	1	2	2	2	0	2	1
Accuracy of the content in Chinese	2	1	2	2	2	2	2	0	2	2	1	1	0	2
Authenticity of Chinese	2	2	1	2	2	2	2	0	2	2	2	2	0	2
Appropriateness of the content in Chinese	2	2	2	2	2	2	2	0	2	2	1	0	0	2
Accuracy of the translation in German	2	1	0	2	2	2	2	0	2	1	2	2	0	2
Clarity of the translation in German	2	0	0	2	2	1	2	0	1	1	1	0	0	2

The following remarks are important to understand the rating within the evaluation grid: a) The overall assessment category does not necessarily reflect the ratings of the other main categories, because it shows overall ratings, which have not been applied to every individual information category. b) Within the area of the German-Chinese entries, the category “Correctness of pinyin letters and tones” applies to the equivalent and not the lemma category. c) Quality criteria can be awarded with 0, 1 or 2 points, e.g. if no mistakes were found in the pronunciation indication of the example category, 2 points were awarded for 0 mistakes, 1 point for 1 or 2 mistakes, and 0 points for 3 mistakes or more.

A first observation is that the overall quality of all entries is relatively high (between 68 and 92%). They are all located between “satisfactory” and “well”: 6/13 reached Grade III, and seven entries were awarded Grade II. Three entries even

reached the upper range of Grade II with 90 resp. 92%. The first experiment of Evert et al. (2024) already shows that entries generated through prompts of lexicographical experts range comparably much higher than those made by lay persons, which also relates to the high grades of the entries in this study. Nevertheless, the AI shows very specific flaws, which will be discussed under Section 3.2.2. In general, the entries show remarkable flaws concerning the pronunciation indication in the example category, in the overall coherence, and also regarding the redundancy category.

Gruppe 2

**Deutsches Wort:** Freundschaft (feminine, kein Plural) (Nomen, bedeutet: persönliche Verbundenheit und gegenseitige Zuneigung zwischen zwei oder mehr Menschen aufgrund gemeinsamer Interessen, Einstellungen oder Erfahrungen)

**Chinesische Übersetzung:** 友情 (yǒuqíng) (Nomen)

**Pinyin-Aussprache:** yǒu qíng

**Bedeutung:** Eine enge, emotionale Verbindung zwischen Personen, die durch gegenseitiges Vertrauen, Wertschätzung und Unterstützung gekennzeichnet ist.

**Chinesische Beispiele:**

1. Unsere Freundschaft reicht zurück bis in unsere Kindheit. (Wir sind seit unserer Kindheit befreundet.) 我们的友情可以追溯到孩提时代。(Wömen de yǒuqíng kě yǐ zhuīsù dào hái tí shí dài.)
2. Eine echte Freundschaft besteht auch durch schwierige Zeiten hindurch. (Eine wahre Freundschaft hält auch bei Problemen an.) 真正的友情即使在困难时期也能维持不变。(zhēnzhèng de yǒuqíng jǐshǐ zài kùnnán shíqī yě néng wéichí bùbiàn.)
3. Sie schenkte mir ihre Freundschaft, was ich sehr schätze. (Sie bot mir ihre Freundschaft an, die ich sehr wertschätze.) 她赋予了我她的友情, 这是我非常珍视的。(Tā fù yǔ le wǒ tā de yǒuqíng, zhè shì wǒ fēicháng zhēnshì de.)

**Chinesische Kollokationen:**

1. Freundschaftsbekundung (友情的表现): yǒuqíng de biǎoxiàn Beispiel: Schenkung eines Geschenks als Freundschaftsbekundung (送礼物作为友情的表现 Sòng lǐwù zuówéi yǒuqíng de biǎoxiàn)
2. Freundschaftsdienst (友情的帮助): yǒuqíng de bāngzhù Beispiel: In einer Notlage einen Freundschaftsdienst erweisen (在危难时伸出援手做友情的帮助 Zài wēinán shí shēnchū yuánshǒu zuò yǒuqíng de bāngzhù)
3. Freundschaftsbeziehung (友情关系): yǒuqíng guānxì Beispiel: Eine lange und stabile Freundschaftsbeziehung pflegen (维护长久稳定的友情关系 Wéihù chángjiǔ wěndìng de yǒuqíng guānxì)

Figure 2. Display of entry 2.4

玩意 wán yì

收起 ^

1. Deutsche Übersetzung: Spielzeug, Ding, Gadget
2. Aussprache Pinyin (mit Tonmarkierung): wán yì (1.3.)
3. Bedeutung: Das Wort "玩意" bezieht sich auf Spielzeuge, Gadgets oder interessante Dinge im Allgemeinen, die unterhaltsam oder faszinierend sind.
4. Beispielsätze (inkl. Pinyin):
  - 我刚买了一个新玩意, 是个智能手表。 Wǒ gāng mǎi le yī gè xīn wán yì, shì gè zhì néng shǒu biǎo. (Ich habe gerade ein neues Gadget gekauft, einen Smartwatch.)
  - 这个小玩意儿真不错, 能帮我记住很多事情。 Zhè gè xiǎo wán yì ér zhēn bù cuò, néng bāng wǒ jì zhù hěn duō shì qíng. (Dieses kleine Ding ist wirklich gut, es hilft mir, vieles zu merken.)
  - 他在旧货市场淘到了不少古董玩意。 Tā zài jiù huò shì chǎng táo dào le bù shǎo gǔ dòng wán yì. (Er hat auf dem Flohmarkt einige antike Spielzeuge oder Kuriositäten gefunden.)
5. Kollokationen:
  - 新鲜玩意儿 (xīn xiān wán yì ér) - neue Gadgets, aktuelle Spielzeuge
  - 搞些玩意 (gǎo xiē wán yì) - einige Dinge basteln, Hobbys betreiben
  - 玩意收藏家 (wán yì shōu cáng jiā) - Sammler von Spielzeug oder Kuriositäten

Figure 1. Display of entry 1.6 with an optimized prompt

An exemplified display and a description of entry 2.4 follow (cf. Figure 1), displaying the lowest result of all entries. It also shows significant errors that the AI is producing: The colour coding in the displayed entry shown in Figure 1 should be read as follows: orange = redundancy, yellow = pronunciation mistake, pink = unrequested information, and green = mistake in the overall usage of the displayed information. The entry for the lemma *Freundschaft* [friendship], which is spelled correctly, is followed by two brackets that show unrequested (and false) grammatical information “(feminine, kein Plural)” [female gender, no plural], and an extra definition phrase. These kinds of mistakes did not occur when AI produced entries for the direction Chinese-German. The correctly written equivalent shows a redundant second indication of pronunciation, and again unrequested grammatical information. The indication of the pronunciation and the definition phrase are linguistically correct, with a clear and accurate definition phrase for the given lemma. The example category shows two different flaws, which can be found in most of the generated entries: The first common mistake the AI is making concerns the pronunciation; in this case the tone is missing for the vowel “nan” (难) [difficult]. More of AI’s issues with the Pinyin romanization system will be discussed in 3.2.2. The second mistake concerns again a general redundancy which the AI tends to produce – in the case of entry 2.4 it gives a second German version of the already written sentence after the German example sentence, which seems to be a more literal translation of the original Chinese sentence. But this might confuse learners more than help due to information overload. The Chinese examples are correctly spelled, appropriate and authentic. But the Chinese sentence should be given before the German example, since the entry is made for German learners of Chinese. The collocation section appears to be rather chaotic: firstly, the collocations sound very rarely and unnatural according to the Chinese expert, and secondly, the example sentences given for every single collocation were not requested. This article was awarded the worst amongst the evaluated data reaching only 68%, which is just two percentage points above GRADE IV.

### 3.2.2 Discussion of AI-specific questions

In this section, we will discuss three AI-specific questions, which arose from the results of the evaluation of the AI-generated dictionary entries shown in Section 3.2.1: 1) Which specific flaws does the conversational AI Tongyi Qianwen

produce in the given task? 2) What role does the prompting play, and how can it be changed to produce better results? 3) Do other conversational AI chatbots as ChatGPT 3.5 or Gemini produce different kind of problems?

To 1): Here we would like to categorize the errors made in linguistic, content, and structural mistakes, which can be marked to be occurring more or less frequently. The most common linguistic errors for Chinese are mistakes in the pronunciation (~25) of the example category, and in a few cases in the collocation, and just one time in the pronunciation categories. The problems that appear in generating the Hanyu pinyin romanization system are three folded: In some cases, it either does not show the correct tone (as in entry 1.1: “yě” instead of yè in 业余 [spare time]), or it entirely lacks the tones (as in entry 1.5: “paimai” instead of pāimài in 拍卖 [auction]), or it displays the wrong vowels as in entry 1.6 (“shōujiě” instead of sōují in 搜集 [to collect]). In less common cases, it also lacks a whole vowel, although the character was given in the example sentence before as in entry 1.5, where the yǒu of 有价值 [valuable] is missing in the first sentence. In one case in entry 1.7, the AI even generates characters instead of Pinyin: “lèisi bǐ rú zhè 样的 qíngkuàng”. All of this might indicate that the Ai, when creating content, does not relate to the sentences it just wrote.

Regarding the content, the AI sometimes tends to use non-existing collocations as in entry 1.7 (“正如比如所说” for ‘as mentioned before’), which should be the “比如前面提到的” which the AI generated at first. After that, it apparently just made up two collocations. To be fair, a lemma as ‘for example’ shows considerable difficulties in terms of finding collocations in all languages. Moreover, it uses collocations that are very rarely used or sound unnatural, such as the three collocations in entry 2.4: “友情的帮助” [\*service of friendship] for example is a phrase which simply cannot be used. Another recurring problem, which is also a structural one, is the change of the actual lemma, as is the case in the collocation category in entry 2.2: here the AI uses “突出强调” [to highlight, emphasize] instead of the entry lemma “特别” [especially]; in the phrase “业余时间” [leisure time] in entry 1.1 it changes parts of the lemma in the collocation indication: “丰富的业余生活” [rich leisure life] or “增加业余爱好” [increase hobbies] instead of the given lemma. Another problem in this content category is the inaccuracy in the indication of meaning in entry 1.2, where the lemma “叔叔” [uncle, younger brother of the father] falsely and imprecisely is defined as “brother of the father or mother”. The second meaning of the lemma – the addressing of older adult male strangers by younger people and children – is entirely omitted.

Structural problems mainly concern the overall redundancy and coherence of the entries. In all entries, the double display of Pinyin after the lemma and in the pronunciation section is redundant, and there is a not-requested summary of the whole entry at its end, which is the case in 12/15 entries; in the other three entries the AI gave a second definition phrase right after the lemma as described in 3.2.1. Redundancy is also occurring in the two similar example sentences (cf. 3.2.1), and in the unrequested examples sentences in the collocation indication in entries 2.2-2.5. In 2.3, the AI even gives another unrequested English translation of the examples, which addresses the general problem of a partly missing overall coherence. The mixture of English and German category headwords is also a common problem, occurring in more than half of the generated entries. The reason might be that the prompt was phrased bilingually. Another structural problem occurring in entry 1.6 “玩意儿” [toy, thing, plaything] is the pollution of the entry with lemmas of earlier generated entries: in this case, each of the three example sentences uses the lemma of 1.5 “古董” [antiquities] such as in “1. 我在旧货市场上发现了一个有趣的古董玩意儿。” [I found an interesting antique gadget at a yard sale.] The reason for this is most likely that all entries were generated in the same thread. However, the clear entry structure seems to be preserved within the same thread. If the same prompt is used in a new thread, the structure of the entry might change, and additional prompting must be used to restore the intended structure.

To 2): The prompt for generating the entries in any case had the same structure:

“Generate a bilingual dictionary entry Chinese-German for German users for the word 比如; including: 1. deutsche. Übersetzung [German translation], 2. Aussprache [pronunciation] Pinyin, 3. Bedeutung [meaning], 4. 1-3 Beispiele [examples] and 5. Kollokationen [collocations].”

An optimized prompt used in a new thread created an entry that was awarded 93 % by the two experts. The prompt shows additionally specifying elements that concern the use of English in the entry, the entry’s structure, and the different aspects of pronunciation:

“Generate a bilingual dictionary entry Chinese-German for German users for the word 玩意; including: 1. deutsche. Übersetzung, 2. Aussprache Pinyin (including tones), 3. Bedeutung (in German), 4. 1-3 Beispielsätze (including Pinyin) and 5. Kollokationen. Highlight only the category headwords in bold type. Don’t use any English in the entry.”

The result is shown in Figure 2. Most of the found redundancies have been eliminated (except the double display of the Pinyin after the lemma), the pollution of this entry has entirely disappeared, and no mistakes can be found in the pronunciation anymore, which might be just a coincidence. The only two mistakes found are a German grammar issue with the English loanword “Smartwatch”: the gender of the article should be female and not neutral, and the German verb in example 2 is incorrect as it demands the reflexive form of the verb using another pronoun (in this case “mir” [me,



myself]). Moreover, in the new thread the AI changed the microstructural indicators while numbering the different categories, which is untypical for dictionary entries, but was to be expected as mentioned above. Two main conclusions for achieving better results can be drawn here: a) give more specific instruction with additional remarks, and b) use new threads to avoid pollution, but give extra instructions concerning the structure.

To 3): Looking at a few samples for generated entries in other AI chatbots shows interesting results: The optimized prompt mentioned above was also given to Google's Gemini and OpenAI's ChatGPT 3.5. While Gemini fails in almost every category, the overall results of ChatGPT 3.5 are in the upper GRADE II and the whole GRADE I range of the evaluation grid. Gemini's generated entry achieved 52% in the rating and thus almost failed to be eligible as a dictionary entry due to several remarkable errors: the first issue is a direct indication of the tone of a vowel in brackets ("wán (3. Ton)"), which is not only false (it should be "2. tone") but also randomly repeated by the AI in the example indication: "Zhè zhī wán yì hěn nán wán. (Zhè zhī (zhī) wán (3. Ton) yì hěn (hěn) nán (nán) wán (3. Ton)." The second and much bigger problem is also shown in this quotation: the examples are given in Pinyin and not in Chinese characters; furthermore, there is a duplication of the vowels with or without tones in brackets after a given tone, which means it is being treated as character. It is most likely due to the given prompt that the AI did not give examples in Chinese and Pinyin in brackets afterwards. The last problem is the collocations: the AI changes the lemma in Tongyi qianwen and is generating collocations for 玩具 [toy] instead of 玩意 [toy, thing, plaything]: "儿童玩具 (ér tóng wán jù – Kinderpielzeug [sic!])". ChatGPT on the other hand produces consistently good results ranging above 90%. After the first sample of 玩意 that reached a rating of 98%, five more samples with the lemmas used in the experiment were given to ChatGPT to confirm or falsify this result. Except one, all entries ranged between 90 and 92%. The outlier was the lemma 比如 [for example] discussed above, for which collocations are difficult to find at all. All other entries had linguistically correct, accurate, authentic, and appropriate examples and collocations, with only two exceptions lacking authenticity or appropriateness in the collocation segment. Structurally speaking, also ChatGPT did number the different categories, which indicates that the prompting must be adjusted in this regard.

It is a striking acknowledgement that ChatGPT has considerably fewer mistakes in terms of the official Chinese pronunciation system than the Chinese AI Tongyi Qianwen. Moreover, the American AI can produce rather authentic examples, even if they tend to be very short and simple: "孩子们喜欢玩这些玩意。" [Children like to play with those toys.].

### 3.2.3 Comparison with existing online dictionaries of Chinese-German

The comparison with existing Chinese-German online dictionaries shows significant advantages of the AI in terms of language learning. This is due to a lack of information in all four considered dictionaries (HanDeDict, Leo, Pons, Langenscheidt) which in most cases do not show other than these equivalents: "玩意儿•玩意兒 wán yìr (1) Ding, Sache (S) (2) Spielzeug (S)" (HanDeDict). In the case of this lemma, there is not even an entry in Leo and Pons. For the lemma "收藏 [to collect]", the dictionaries show slightly better results: Leo displays four possible equivalents ("aufbewahren, lagern sammeln, speichern [to store, to stock, to collect, to save]") and three compounds with the lemma such as: "画品收藏 [畫品收藏] huàpǐn shōucáng – die Gemäldesammlung [collection of paintings]". HanDeDict shows the lemma with the equivalent for "to collect", and 10 other lemmas, of which 收藏 is a part. Pons shows three possible equivalents and two collocations that are not marked as such. Langenscheidt just gives two possible equivalents. All four dictionaries greatly lack information and/or a well-organized microstructure; extra paragraphs with examples and collocations are apparently not considered content. Thus, we dare to assume that AI can provide much more informative and structured data for language learners. The existing dictionaries seem to be merely equivalence dictionaries, as the few samples considered here have shown. Since this test is not exhaustive, it cannot be considered as representative.

### 3.2.4 Analysis of the experiment

The results and evaluation of the tasks given in the experiment are described in Section 2: For the evaluation of the two sentences, the following criteria were considered: understandability of the translation, the test person's understanding of the sentence, grammar/orthography, and the expression and readability (Chinese into German); understandability of the sentence, the correctness of the words, the correctness of the characters, and the correctness of the grammar (German to Chinese). Due to the errors produced by the AI and the information overload in the displayed entries of the online dictionary, extra points could be awarded, which happened 4 times.

Whereas 4/5 of group 2 reached 100% of the first task, just 1/5 of group 1 reached the same result, 3/5 were above 50% (50, 75, 88%). In every group, 1/5 failed the task (13%/38%). The second task shows somewhat different results, since the translation into L2 is considered more difficult than the other way around. In group 1, 3/5 subjects reached more than 50% of the evaluation (50, 68, 70%), and in group 2, just 2/5 reached more than 50 % (78, 83%). In terms of AI use, group 2 achieved better results in the translation L2 to L1, and group 1 slightly better results in the translation task L1 to L2. However, at this point it is not clear if the results are a consequence of the use of different data base (AI vs. online dictionary entries) or happened due to the language level of the test persons.

Analysis of the questionnaire: In the questionnaire, questions about a general rating of the dictionary entries, a rating of the entries L2-L1 / L1-L2 particularly useful, missing, and redundant information – and detected errors had to be answered.

In general, the questionnaire was either not completed or not completed sufficiently 3 times. In group 2 (AI group), only 3 people sufficiently completed it. The following remarks will concentrate on the negative evaluation through the test persons, since the errors are crucial for the didactical concept we will present in Section 4. The main criticism regards the length of the entries, especially the (unrequested) summary at their end, too many examples, respectively too many examples and collocations considered as unnecessary. One participant even recognized that the equivalent, pronunciation and meaning were sufficient for the translation, whereas examples and collocations are more valuable for self-study. Additionally, one test person criticized the lack of audio (which could not be realized mainly due to the print format, but also neither Tongyi, nor ChatGPT 3.5 provide this feature). Other criticisms regarded the mixture of English and German, and the random bold type of some characters shown in the printed entries (the latter was due to the transition to a word document and not shown in the chat threads of the AI).

#### 4. Proposal of a didactical concept using AI in Chinese language teaching

As long as conversational AI chatbots get caught producing errors, it seems obvious to take advantage of those mistakes, and make the AI's weaknesses a valuable asset for language learning. In order to develop a systematic approach in this regard, we will firstly show a compilation of the errors of AI described above, and secondly, we will derive tasks from those flaws – also considering the conclusions drawn in Evert et al. 2024 – for a Chinese class with participants of beginner to intermediate knowledge of Chinese. Table 4 shows an overview of the categorization of the detected errors in this study (the color-coding shows possible AI inherent reasons for errors = **training**, **generating**, **prompting**, **thread**, **feature**):

Table 4. Categorization of the detected AI's errors

source: error types:	Multi-level evaluation framework				Questionnaire	
linguistic errors	<b>Pronunciation</b>					
content errors	inaccuracies in the collocation indication	change of the lemma within example & collocation indication	inaccuracies in the meaning indication			
structural errors	redundancy	unrequested information	pollution of the entries	mixing of English and German	missing of audio	information overload / redundancy

For this didactical concept, we will not take every mistake displayed in Table 4 into account, but concentrate on errors in particular, which we consider important for learners to recognize and review. Moreover, we also aim at introducing conversational AI chatbots to learners as a valuable tool in language-learning as well as convey dictionary literacy (*sensu* Chi 2020). More simply speaking, learners should benefit from those tasks in terms of learning a foreign language and learn how to use AI sufficiently as well. Additionally, they may (indirectly) learn more about the value of lexicographical data. Therefore, we will take AI's issues with the pronunciation and the inaccuracies in the meaning indication for the part of language learning, and the teaching of a sufficient prompting to avoid information overload and redundancy as a part of using AI. Regarding dictionary literacy, the test persons will be encouraged to use relevant and appropriate lexicographical sources for the review of found mistakes and concerning the disadvantages of using AI. For the latter, teachers must possess “[p]edagogical lexicographical knowledge (both print and digital dictionaries) related to language teaching, [and possibly knowledge and skills] in dictionary criticism [...]” (Chi 2020, p. 92).

In the following, we will introduce two didactical concepts in a class setting, which will concentrate 1) on prompting, and 2) on reviewing data given by AI on the basis of certain dictionaries.

As a prerequisite, every student in class needs access to a conversational AI chatbot, whereas – as shown above – different AI chatbots produce different errors, and even might increasingly get better in producing certain outputs as the results produced by ChatGPT 3.5 show. It is thus reasonable that every participant uses the same AI for the task.

To 1): In terms of prompting, students should be enabled to understand how a certain prompting produces different results, while learning that certain formulations, and also using very specific prompting can create better results. To achieve a step-by-step approach, participants may be encouraged to ask the AI about information for a certain word, e.g. about the verb 吃饭 [to eat] in a beginners' class, using a simple formulation such as “please give me information about (word)”. After the AI has produced an output, participants can evaluate, and especially criticize the output without using other sources. In a second step, it can be suggested to ask the AI about generating a dictionary entry for the word and evaluate the output again. In a possible intermediate third step, participants or groups of them could be asked to formulate prompts by themselves that could create better results; this is also a question of the time available. In a fourth step, the teacher might suggest to give the kind of prompting we introduced in 3.2.2 to get information about the word; for that the teacher can work with the students on the content that the dictionary article should comprehend according to the needs of the participants. The results of all three (or more) outputs can be compared by the class within groups. We generally recommend using dehumanized language in the prompting, concentrating on the information needed by the user. This could be the beginning of a teaching method, using AI for a whole semester and/or for self-learning, in which students will generally obtain information about words in this way, and start to compile personal dictionary entries made by AI. To guarantee appropriate information for language learning, participants will need to be enabled to verify or falsify the



results shown by the AI, which we will introduce in the following.

To 2): The information given by the AI should obviously not entirely be trusted without checking its accuracy and correctness. Taking the example of the word 叔叔 [uncle, younger brother of the father], which in the case of Tongyi Qianwen showed a partly false and incomplete definition phrase, students/test persons should be suggested to check the meaning using e.g. paper dictionaries such as *Das Neue Chinesisch-Deutsche Wörterbuch* [The New Chinese-German Dictionary], which shows the following entry: “叔叔 shūshu 1) <口> (叔父) jüngerer Bruder des Vaters; Onkel m 2) (称父辈男子) Onkel m” [Uncle shūshu 1) <colloquial> (uncle) younger brother of the father; uncle m 2) (Calling men of the father's generation) uncle m]. As can be seen, all information in this dictionary is sufficient; the only caveat found is the second definition phrase, where the (pragmatic) meaning lies within the Chinese shown in the brackets; the teacher must help in this case and explain the meaning of the pragmatic indication. The word can also be checked in the above-mentioned online dictionaries, almost all of which show the two meanings of the word; only Pons shows only the second meaning “叔叔 shūshu Anrede für jüngere männliche Personen der Generation des Vaters [Addressing of younger males of the father's generation]”. Teachers should compile a collection of lexicographical recourses in advance for such cases and provide it to their students/test persons. If they are speakers of English, we recommend two other dictionaries for more information about a given lemma: The first dictionary is of Dutch origin and can be found on the internet, its name is *MDBG* and it shows exhaustive information about Chinese characters such as pronunciation, meanings, measure words, HSK levels, stroke count, radical, variants, character decomposition, (animated) stroke order, audio pronunciation, and the possibility to search for combinations with the given character (also systematically with only one character before, after or before and after the lemma). Moreover, it shows scientific information about the pronunciation in Cantonese, Sino-Japanese, Korean, Hangul, and Vietnamese. In addition, it is also displayed where it can be found in the relevant classical and modern Chinese dictionaries, Japanese and Korean dictionaries and relevant classical and modern bilingual dictionaries, Unicode number etc. The latter information is just for scientists and do not give much relevant information for beginners of Chinese. The second dictionary recommended is Pleco, which is a mobile dictionary portal application for Chinese that shows the collected entries of free and bought dictionaries. The free version (Pleco Basic Chinese-English Dictionary) also shows sufficient information about the given lemma: “叔叔 shūshu noun colloquial 1 father's younger brother; uncle 2 [a child's form of address for any young man one generation older] uncle”. Additionally, it shows the meanings of the single character, compounds, and example sentences. Also the pronunciation of given words should be reviewed in the same way and combined with speaking practice in class supervised by the teacher.

If students start to compile AI generated dictionary articles for new vocabulary, they can request the AI to add missing meanings, and straighten out wrong meaning or pronunciation indications by adding the information to the prompt that they found in the lexicographical sources.

## 5. Outlook

In this paper, we tried to achieve two things: First, to further utilize and adjust our Multi-level Evaluation Framework for AI-generated dictionary entries introduced in Evert et al. 2024 for bilingual entries of Chinese/German. These dictionary entries were on the one hand applied in a didactical experiment to find out more about the data used in a practical context, to assess how it helped to solve certain tasks, and to consider the feedback of the participants. On the other hand, the entries were evaluated with our evaluation grid, and compared to entries of existing online dictionaries. Also specific aspects of the AI such typical errors and sufficient prompting were discussed. Second, to propose a didactical concept that reasonably deals with AI-generated lexicographical content by introducing a sufficient way of prompting to generate appropriate information, and by reviewing the information using relevant lexicographical resources. This concept is expected to be applied in further experiments soon and introduced in a follow-up publication.

In summary, we would like to note that we consider AI as a valuable tool for language learning, because it might show more comprehensive information about words than existing online dictionaries. A well-developed microstructure by the AI, containing representative and authentic indications of contextualized examples and collocations, not only promote vocabulary acquisition, but also provide material for production of speech and text. This kind of information can be reviewed and thus verified, falsified, and extended. If users are enabled to detect errors made by the AI, they are able to learn more about the language they are trying to master, about the use of the AI in a language-learning context, and about relevant lexicographical resources created by experts. In a learning environment, our developed Multi-level Evaluation Framework for AI-generated dictionary entries might help experts and teachers to detect errors made by the AI and develop more distinctive didactical concepts to introduce AI as a language-learning tool in class.

## Notes

<sup>1</sup> HSK = Hanyu Shuiping Kaoshi (汉语水平考试) Chinese Proficiency Test, HSK 1-3: Beginner, HSK 4-6: Intermediate, HSK 7-9: Advanced

<sup>2</sup> The order slightly differs from the original Multi-level Evaluation Framework in Evert et al. (2024).

<sup>3</sup> On general characteristics of learner's dictionaries, which were partly utilized for the dictionary entries and the evaluation here, cf. Herbst/Klotz 2003, pp. 242-243; Zöfgen 1991, pp. 2898-2899. Aspects of learner lexicography for Chinese as a foreign language can be found in Zhang 2023, pp. 63-95.

## References

- Chi, M.L.A. 2020. 'Reconstructing the lexicographical triangle through teaching dictionary literacy to teachers of English' *Lexicography* 7, pp. 79-95.
- Evert, S. et al. 2024. 'Multi-level analysis as a systematic approach to evaluating the quality of AI-generated dictionary entries'. *Lexicography and language technologies: the use of computational linguistics, natural language processing and computer science technologies and language resources in lexicography* (EURALEX. 2024. Proceedings), in print.
- HanDeDict 汉德词典. 2015-2024. <https://handedict.zydeo.net/de>. (Last access: 17 June 24)
- Herbst, T, and M. Klotz 2003. *Lexikographie*. Paderborn: Ferdinand Schöningh
- Langenscheidt Chinesisch-Deutsch Wörterbuch. 2024. <https://de.langenscheidt.com/chinesisch-deutsch/>. (Last access: 20 June 24)
- Leo Chinesisch Deutsch. 2006-2024. <https://dict.leo.org/chinesisch-deutsch/>. (Last access: 20 June 24)
- Pleco Basic Chinese-English Dictionary based on A Chinese-English Dictionary, 1995 and 2009. Foreign Language Teaching & Research Press.
- PONS Online-Wörterbuch Chinesisch Deutsch. 2001-2024. <https://de.pons.com/>. (Last access: 20 June 24)
- Das Neue Chinesisch-Deutsche Wörterbuch 新汉德词典. 2004. (XHDCD). Beijing: The Commercial Press.
- Zhang, Y. 2022. *Zeichen- und Wortschatz des Chinesischen effizienter lernen. Konzeption einer Lernwörter-App für deutsche ChaF-Lernende auf der Basis einer empirischen Bedarfsanalyse*. Ph.D. thesis submitted to the Gottfried Wilhelm Leibniz University.
- Zöfgen, E. 1991. Bilingual learner's dictionaries. In F.J. Hausmann et al. (eds.) (1989, 1990, 1991). *Wörterbücher/ Dictionaries/Dictionnaires. An international encyclopedia of lexicography*, 3 vols. Berlin: W. de Gruyter, pp. 2888-2903.