http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

A Comparative Study on the Effectiveness of AI Chatbots and Dictionary Apps for Lexical Tasks and Retention

Yuzhen Chen, *College of Foreign Languages, Putian University, Fujian, P.R.C. (287323222@qq.com) (https://orcid.org/0009-0004-6431-7768)*

Abstract: This study compared an AI chatbot (Kimi) and a bilingual dictionary app (NCD) in supporting vocabulary tasks among Chinese junior English majors. Sixty-six participants used either Kimi or NCD to complete both receptive and productive lexical tasks. Questionnaires gathered user feedback on tool use, and a surprise retention test assessed long-term vocabulary retention one week later. Results showed that Kimi significantly outperformed NCD in vocabulary comprehension, collocation production, and productive knowledge retention. Additionally, Kimi demonstrated more consistent performance than NCD across all test items, highlighting its reliability. The study underscores the potential of AI chatbots to address language-related queries and enhance vocabulary acquisition. It also advocates for aligning technological advancements with pedagogical goals to optimize language learning tools and create a sustainable learning environment.

Keywords: AI CHATBOT, DICTIONARY APP, VOCABULARY RECEPTION, COLLOCATION PRODUCTION, RETENTION

Opsomming: 'n Vergelykende studie van die effektiwiteit van KI-kletsbotte en woordeboektoepasssings in leksikale take en retensie. In hierdie studie word 'n KI-kletsbot (Kimi) en 'n tweetalige woordeboektoepassing (NCD) vergelyk ten opsigte van die steun wat hulle aan Chinese junior studente met Engels as hoofvak in woordeskatopdragte bied. Ses-en-sestig deelnemers het óf Kimi óf NCD gebruik om beide reseptiewe en produktiewe leksikale opdragte te voltooi. Gebruikersterugvoer oor hulpmiddelgebruik is met behulp van vraelyste ingesamel, en 'n week later is die langtermynwoordeskatretensie in 'n onverwagse retensietoets bepaal. Die resultate het getoon dat Kimi beduidend beter as NCD in woordeskatbegrip, kollokasieproduksie en produktiewe kennisretensie gevaar het. Daarbenewens het Kimi in al die toetsitems konsekwenter vertoon as NCD, wat die betroubaarheid van Kimi beklemtoon. Hierdie studie benadruk die potensiaal van KI-kletsbotte om taalverwante vrae te hanteer en om woordeskatverwerwing te verbeter. Dit bepleit ook die belyning van tegnologiese vooruitgang met pedagogiese doelwitte om hulpmiddels vir die aanleer van taal te optimaliseer en 'n volhoubare leeromgewing te skep.

Sleutelwoorde: KI-KLETSBOT, WOORDEBOEKTOEPASSING, WOORDESKATRESEPSIE, KOLLOKASIEPRODUKSIE, RETENSIE

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

1. Introduction

The field of Artificial Intelligence (AI), especially Natural Language Processing (NLP), has achieved significant breakthroughs since the emergence of transformer-based architectures, exemplified by generative Large Language Models (LLMs) such as OpenAI's GPT series, Google's PaLM, Meta's LLaMA and Baidu's Ernie. These models possess unprecedented application potential and have brought disruptive impacts to scientific research, technological innovation, and daily life (Yuan 2023: 8). In the field of lexicography, a new age of the successful application of generative AI has dawned (De Schryver 2023: 1). An increasing number of scholars are investigating how LLMs can reshape lexicographical practices and the roles of dictionaries and lexicographers in the AI era. However, a critical area remains under-researched: the effectiveness of AI chatbots in resolving language-related uncertainties and enhancing vocabulary acquisition.

Recent studies have presented mixed results regarding AI's impact on vocabulary comprehension and production. For instance, Rees and Lew (2024) found no significant difference between AI-generated and human-authored definitions in reading comprehension tasks. In contrast, Ptasznik et al. (2024) and Lew et al. (2024) revealed the advantages of AI chatbots over monolingual online dictionaries in lexical reception and production. Despite these findings, the long-term impact of AI chatbots on vocabulary retention remains unexplored. This gap points to the need for further research to evaluate the role of intelligent tools in meeting learners' lexical needs through comparative studies from a user perspective.

2. Literature review

2.1 Generative AI and lexicography

Current research on generative AI in lexicography primarily explores three dimensions. The most prominent strand examines the capabilities and limitations of AI models in producing lexicographical content, e.g. De Schryver and Joffe (2023) who successfully integrated ChatGPT into the TLex dictionary system, Phoodai and Rikk (2023) who revealed ChatGPT's superiority in structuring entries for high-frequency words, capturing nuanced elements like phonetic transcription and morphological details, and McKean and Fitzgerald (2024) who found that AI models do not meet human editorial standards, necessitating significant human oversight (see also De Schryver 2023, Jakubíček and Rundell 2023, Lew 2023 and Rundell 2023).

The second dimension explores the opportunities, challenges, and future trajectory of AI in lexicography, investigating the dynamics of human-AI collaboration and AI's capabilities to streamline workflows and increase productivity. Tarp and Nomdedeu-Rull (2024) and Huete-García and Tarp (2024) researched

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

the development of a Spanish AI writing assistant project, underscoring the complexity of human-machine interaction and the crucial role of human lexicographers in maintaining quality, while Zhao (2023) indicated that LLMs can cluster concordance lines and streamline semantic categorization, thereby significantly reducing lexicographers' workload (see also Fuertes-Olivera 2024 and Lew 2024).

The last and relatively understudied but arguably most relevant dimension for this article, gauges the efficacy of AI chatbots in addressing real-world vocabulary inquiries from a user perspective. Rees and Lew (2024) divided 43 students into three groups for a reading comprehension task. One group used definitions from the *Macmillan English Dictionary* (MED), another used AI-generated definitions, and a third had no dictionary access. The results indicated that the MED group surpassed the non-dictionary group in performance. However, there were no significant differences between the MED and AI groups, or between the AI and non-dictionary groups. Furthermore, no notable variations were found in the time taken to complete the task between the two groups exposed to definitions. Nonetheless, the authors maintained that AI's benefits for language learners, educators, translators, and lexicographers remain attainable (Rees and Lew 2024: 65).

Ptasznik et al. (2024) investigated the utility of AI in facilitating vocabulary reception and production compared with the online *Longman Dictionary of Contemporary English* (LDOCE). The research involved 223 university students divided into two groups, with one group using ChatGPT and the other resorting to LDOCE. These participants performed two tasks: Translating 20 low-frequency English words from contextualized sentences into Polish, and converting twenty Polish sentences, with the English verbs highlighted, into English. The results indicated ChatGPT's superior performance in both receptive and productive tasks. Additionally, ChatGPT was faster in the production task than LDOCE. The study positions AI chatbot as a formidable competitor to traditional dictionaries and advocates for hybrid strategies that preserve learner autonomy.

Lew et al. (2024) assessed the effectiveness of ChatGPT against a monolingual dictionary (LDOCE) and a popular bilingual dictionary (Diki.pl) in production and reception tasks mirroring Ptasznik et al.'s (2024) experimental design. The findings showcased ChatGPT's enhanced effectiveness over both dictionaries in the production task, while in the reception task, it surpassed the monolingual but not the bilingual dictionary. The study confirmed that a general-purpose chatbot such as ChatGPT can be a viable alternative to traditional dictionaries in both production and reception tasks (Lew et al. 2024: 8).

The research conducted by Ptasznik et al. (2024) and Lew et al. (2024) not only emphasizes the prospects of AI chatbots in language learning but also raises important questions about the future role of traditional dictionaries. As intelligent tools become more prominent, further research is needed to probe how these new technologies can be integrated into language learning and how they compare with traditional dictionaries in addressing users' lexical reference needs.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

In particular, efforts should be made to explore how AI chatbots affect learners' long-term vocabulary retention, an area that remains uncharted so far.

2.2 Dictionary use in lexical tasks

While AI tools offer new possibilities, traditional dictionaries remain a cornerstone in lexical learning, as evidenced by numerous studies on their effectiveness in vocabulary tasks.

Researchers have long examined the utility of dictionaries in vocabulary acquisition by comparing different conditions under which lexical tasks are fulfilled. In a study conducted by Chen (2012), participants were divided into three groups to complete a reading comprehension task that focused on ten target lexical items. Each group used either a paper dictionary, an electronic version of the same dictionary, or no dictionary at all. The results indicated that both groups using diction significantly outperformed the group that did not use a dictionary. This suggests that dictionaries are more effective for vocabulary comprehension than relying solely on contextual cues.

Several other studies have confirmed the beneficial contribution of dictionaries in lexical learning. Li and Xu (2015) observed substantial improvements in task performance after consulting an online dictionary for meaning determination. Alzi'abi (2017) noted increased appropriate responses when electronic dictionaries were used in a verb-adverb collocation task. Chen (2017, 2020) reported that the use of CALL (Computer-Assisted Language Learning) and online dictionaries significantly improved learners' accuracy in producing target collocations.

The impact of dictionary medium on lexical learning has also been a focal area. Chen (2010) compared pocket electronic dictionaries and paper dictionaries in receptive and productive tasks. Despite the faster processing speed of electronic dictionaries, there was no significant difference in vocabulary comprehension, production, or retention between the two types. Dziemianko (2010, 2017) evaluated the efficacy of electronic and paper dictionaries in decoding, encoding, and retaining target vocabulary. Results were mixed, with one study reporting advantages for electronic dictionaries, while the other found no substantial differences. These discrepancies emphasize the complexity of understanding the effects of different dictionary forms on vocabulary learning.

Beyond the medium, researchers have also investigated how specific design features in electronic dictionaries affect vocabulary learning outcomes. Lew and Doroszewska (2009) revealed that the availability of L1 equivalents, either alone or with L2 definitions, strongly predicted word retention. Dziemianko (2015) highlighted the efficacy of color-coded functional labels in accelerating search efficiency and enhancing information recall. Further advancing this line of inquiry, Dziemianko's (2022) comparative analysis identified line drawings as the most pedagogically impactful visual format, demonstrating their dual capacity to streamline cognitive processing and reinforce long-term memory consolidation.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

From the literature review, it is apparent that the investigation of dictionaries' contribution to vocabulary acquisition predominantly focuses on electronic dictionaries. However, with the rise of generative AI, there is growing ambiguity regarding user preferences, the efficacy of AI models in meeting learners' reference needs, and the competitiveness of electronic dictionaries compared with AI chatbots. As we transition into the AI era, it becomes crucial to undertake comparative studies that examine the relative advantages and synergies between electronic dictionaries and intelligent tools within language learning.

3. Study design

Inspired by Ptasznik et al. (2024), this study aims to compare an AI chatbot and a dictionary app in vocabulary tasks. It employs a Chinese AI model and a bilingual dictionary app, differing from ChatGPT and LDOCE, the tools used in the original study. Notably, the study involves an investigation into vocabulary retention.

3.1 Research questions

The research seeks to address the following questions.

RQ 1: Which tool more effectively enhances the comprehension of target words in a receptive lexical task, the Kimi intelligent assistant (Kimi henceforth) or the mobile app of *New Century English–Chinese Chinese–English Dictionary* (NCD henceforth)?
RQ 2: Which tool provides more effective support for the production of target items in a productive lexical task, Kimi or NCD?
RQ 3: Which tool leads to superior retention of target items, Kimi or NCD?

3.2 Participants

The study involved 75 junior English majors from a Chinese university, with 67 females and 8 males aged 21–22. All participants had eight to nine years of EFL learning experience and had passed the Test for English Majors (TEM, Band Four), a national test to evaluate learners' English proficiency. None had used the NCD app before, and only a few had used Kimi, though they were familiar with other AI chatbots or dictionary apps.

3.3 Instruments

The study adopted a pre-test, main test and post-test design. Each phase incorporated two paper-based lexical tasks — one receptive and the other productive — which remained uniform throughout all three stages (see Appendix A).

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

The receptive task

In this task, participants were asked to provide meanings of the target words highlighted within sentences, either in Chinese or English. For example, "*She was led away in a state of* ***stupor***". These target words were chosen from the 10,000-word level of the Vocabulary Levels Test (Schmitt et al. 2001: 87-88) based on several criteria. Firstly, they were deemed unfamiliar to the participants, necessitating the use of tools for task completion. Secondly, they spanned various parts of speech, encompassing content words of significant pedagogical relevance. Lastly, they were not morphologically complex, not obsolete, technical or specialized terms, thereby reducing retention difficulties. After consulting with three teachers from the participants' classes, ten words were chosen: *dabble*, *scrawl*, *vie*, *stint*, *wily*, *squirm*, *torrid*, *banter*, *swagger*, and *translucent*.

Example sentences for the task were sourced directly from either online *Oxford Advanced Learner's Dictionary* (OALD, https://www.oxfordlearnersdictionaries.com/) or LDOCE (https://www.ldoceonline.com), both renowned globally for their exceptional quality. It was ensured that the sentences in the task did not overlap with those provided by NCD, confirmed through trial consultations, or those produced by Kimi in response to prompts like "Please give an example sentence for (the target word)" or "how to use (the target word) in a sentence".

The productive task

In the productive task, participants had to complete missing verbs in verb–noun collocations within sentences, with meanings provided in Chinese, e.g., "*I could see he was trying to ___ a fight with me (寻衅)*". The selection of these collocations was guided by several criteria: (1) mutual information score >3 from the British National Corpus, indicating strong collocation strength; (2) verbs and nouns are among the top 2000 most common words but rarely combined; (3) covered in NCD and retrievable via Kimi; (4) semantically transparent, with meanings inferable from components; (5) some collocations and their translation are congruent while others are not; (6) excluding specialized, formal, or informal collocations to focus on general use. These criteria ensured that the collocations were representative, unfamiliar to most participants, and held significant experimental relevance. Ultimately, ten target collocations were chosen: *write a cheque, lift a ban, blow a kiss, cut a tooth, jump the queue, wear perfume, say a prayer, pick a fight, break a habit, take a joke*. Sentences for the task were similarly sourced from the online versions of OALD and LDOCE, avoiding overlap with NCD or Kimi-generated examples.

It is worth noting that the current study featured only 10 target items per task — half the amount in Ptasznik et al.'s (2024) research. This reduction was necessary, as participants were required to retain the target items, and memory capacity is inherently limited.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

The questionnaire survey

Two concise semi-structured questionnaires, each with eight questions, were designed to gather participants' feedback on Kimi or NCD (see Appendix B). They focused on: (1) perceived quality and quantity of retrieved information, (2) tool usability, (3) challenges encountered during tool usage, and (4) comparative advantages and disadvantages relative to alternative resources.

Despite their brevity, the surveys demonstrated adequate reliability and validity. Both employed structured, closed-ended questions with clearly defined response options, such as Likert scales and multiple-choice formats, which enhanced their internal consistency. Moreover, the inclusion of open-ended questions enriched the construct validity by allowing for the capture of more nuanced user perspectives.

Tools for Task Assistance

For the study, two tools were selected: Kimi and NCD, based on accessibility and participants' dictionary usage habits.

Given the unavailability of ChatGPT at the author's institution, Kimi, a domestic AI chatbot in China, was chosen as an alternative out of several considerations. To start with, it is a free AI chatbot with high accessibility, making it a practical choice for academic and research purposes. Secondly, it has earned wide recognition for its reliability and robust capabilities. Furthermore, Kimi excels in text interpretation and processing, particularly in Chinese, which aligns well with the needs of Chinese users and the context of this study.

Surveys point to the fact that dictionary apps are very popular among Chinese EFL learners (Liu et al. 2019, Ma 2019, Liu et al. 2021). Using these apps in the study reflects participants' real-life dictionary habits. The chosen app, NCD, is the first of its kind in China that integrates a prestigious L2–L1 dictionary (the *New Century English–Chinese Dictionary,* 2016) with an L1–L2 one (the *New Century Chinese–English Dictionary,* 2nd edition, 2016). It offers an innovative "two-in-one" functionality with a convenient "jump" feature for bidirectional searches.

The choice of Kimi and NCD was also informed by participants' dictionary habits. Extensive research has confirmed that most EFL learners in China prefer bilingual dictionaries over monolingual ones. NCD and Kimi both offer bilingual functionality, fitting well with participants' habits and preferences.

3.4 Procedure

The study lasted three weeks. Participants were randomly assigned to the Kimi or NCD group. In Week One, they attended 15-minute preparatory sessions: One group was granted free access to download NCD on smartphones and received training on its features, functions, and consultation methods, while the other group learned to use Kimi, focusing on composing prompts for lexical inquiries.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

In Week Two, participants took a pretest to assess their prior knowledge of target lexical items, performing receptive and productive tasks without reference tools. After the pretest, they completed a main test with the same tasks but in a randomized order to reduce carry-over effects. One group used NCD, while the other interacted with Kimi. Participants had a 20-minute time limit and were not allowed to use other reference sources. Afterward, they filled out a 5-minute questionnaire. In Week Three, a surprise delayed retention test was conducted without reference tools.

3.5 Data processing

Each participant received six scores: three for receptive tasks and three for productive tasks across the pre-test, main test, and retention test. The maximum score for each task was ten points, with one point for each correct response. Minor errors like spelling or verb inflections were ignored. SPSS 27 was employed for data analysis.

The pre-test results affirmed that most participants were unfamiliar with the test items. Only four participants knew the meaning of two or more target words, and five had prior productive knowledge of target collocations. To ensure uniform familiarity, these nine students were excluded from further analysis, leaving a final sample of 66 participants (33 in each class). The extremely low pre-test scores were not included in the analysis. Questionnaire responses were analyzed both quantitatively and qualitatively.

4. Results

4.1 The receptive task

The main test results, presented in Table 1, indicate that students using Kimi achieved perfect scores on the receptive task (M = 10.000, SD = 0.000), while the performance of the NCD group exhibited slight variability (M = 9.697, SD = 0.529). An independent-samples *t*-test revealed a statistically significant difference between the two groups (t = 3.288, p = 0.002, df = 64, two-tailed; Cohen's d = 0.81) (Table 2), suggesting that Kimi surpassed NCD in the receptive main test. As manifested in Figure 1, Kimi achieved a 100% success rate, whereas NCD experienced a few noticeable dips.

Table 1: Group Statistics (Max = 10) (receptive task, main test)

	Class	N	Mean	Std. Deviation	Std. Error Mean
Score	Kimi	33	10.000	0.000	0.000
	NCD	33	9.697	0.529	0.092

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

Table 2: Independent Samples Test (receptive task, main test)

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	Upper
Score	Equal variances assumed	80.188	0.000	3.288	64	0.002	0.303	0.092	0.119	0.487
	Equal variances not assumed			3.288	32.000	0.002	0.303	0.092	0.115	0.491

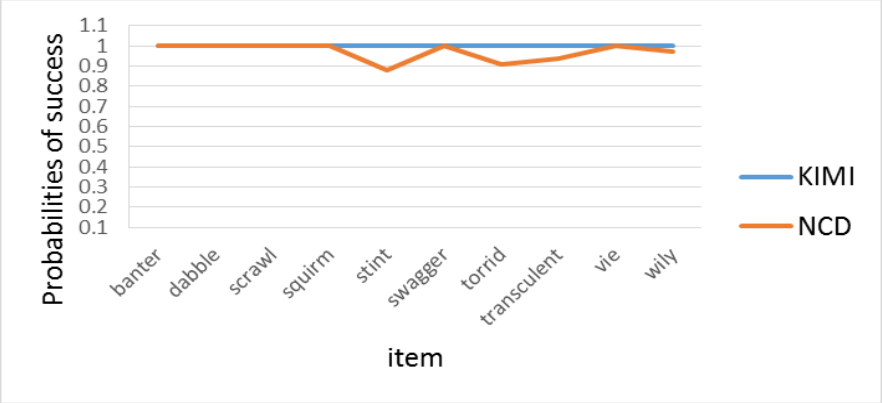

Figure 1: Probabilities of success for each test item in the receptive main test

In the retention test (Table 3), both groups obtained a comparable retention rate of target words. The Kimi group scored 7.667 out of 10 (76.7%), while the NCD group garnered 7.576 out of 9.697 (78.1%). An independent-samples *t*-test found no significant difference between the groups (t = 0.174, p = 0.863, df = 62.197, two-tailed) (Table 4). As depicted in Figure 2, while overall performance was similar, Kimi demonstrated more consistent retention patterns, whereas NCD showed greater variation across test items.

Table 3: Group Statistics (Max = 10) (receptive task, retention test)

	Class	N	Mean	Std. Deviation	Std. Error Mean
Score	Kimi	33	7.667	2.300	0.400
	NCD	33	7.576	1.937	0.372

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

Table 4: Independent Samples Test (receptive task, retention test)

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Score	Equal variances assumed	0.850	0.360	0.174	64	0.863	0.091	0.524	-0.954	1.136
	Equal variances not assumed			0.174	62.197	0.863	0.091	0.524	-0.955	1.137

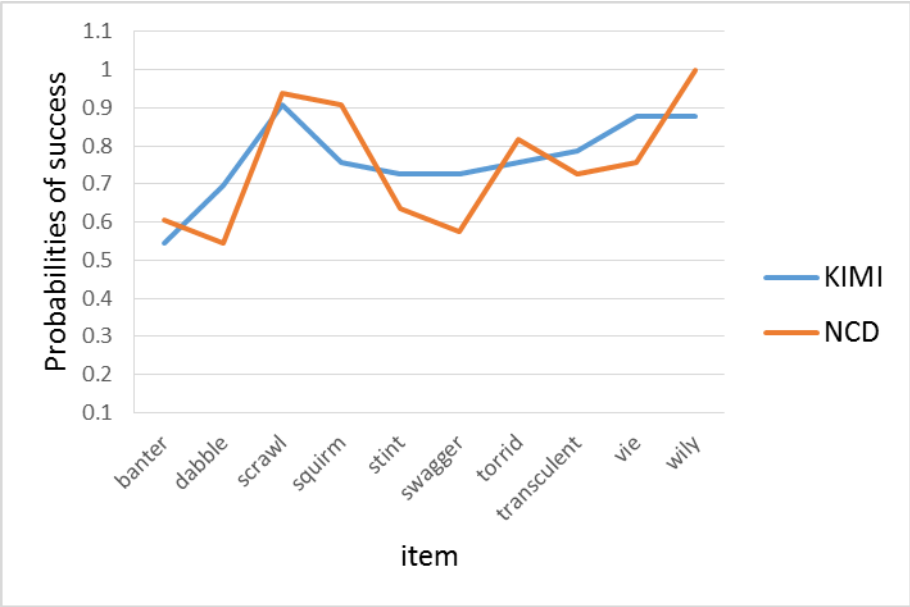

Figure 2: Probabilities of success for each item in the receptive retention test

Table 5 displays the mean score differences for each test item in the main and retention tests as well as detailed performance differences between the groups.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

Table 5: Mean difference in each test item in main and retention tests (receptive task)

Test item	Tool	Mean score (main test)	Mean difference (main test)	Mean score (retention test)	Mean difference (retention test)
dabble	Kimi	1.00	0.00	0.70	0.15
	NCD	1.00		0.55	
scrawl	Kimi	1.00	0.00	0.91	0.03
	NCD	1.00		0.94	
vie	Kimi	1.00	0.00	0.88	0.12
	NCD	1.00		0.76	
stint	Kimi	1.00	0.12	0.73	0.09
	NCD	0.88		0.64	
wily	Kimi	1.00	0.03	0.88	0.07
	NCD	0.97		0.81	
squirm	Kimi	1.00	0.00	0.76	-0.15
	NCD	1.00		0.91	
torrid	Kimi	1.00	0.09	0.76	-0.06
	NCD	0.91		0.82	
banter	Kimi	1.00	0.00	0.55	-0.06
	NCD	1.00		0.61	
swagger	Kimi	1.00	0.00	0.73	0.15
	NCD	1.00		0.58	
translucent	Kimi	1.00	0.06	0.79	0.06
	NCD	0.94		0.73	

In the main test, Kimi users achieved full scores across all target words, while NCD users obtained marginally lower on four items, with differences ranging from 0.03 to 0.12 points. Analysis identified two main reasons for these errors: incorrect sense selection and partial comprehension of definitions. Some students chose the wrong sense for polysemous words. For instance, for "torrid," three students chose the wrong sense (sense 3, 热情似火的) in the sentence "They face a torrid time in tonight's game," despite the correct sense (sense 4, 难熬的，艰难的) being available, probably due to a mismatch with the contextual clues. Another example is "stint," where four students mistranslated it as "工作 (work)","分配 (allotment)", or "disturbed task" instead of understanding it as a "period of work" as a result of incomplete comprehension of the definition. These findings imply that the effectiveness of vocabulary tools depends on users' proficiency in utilizing them.

In the retention test, both groups had a balanced performance, with each outperforming the other on certain items. The highest retention scores were observed for "scrawl" (Kimi: 0.91, NCD: 0.94), while the lowest were for "banter" (Kimi: 0.55, NCD: 0.61). Despite being morphologically simpler, "banter" was less remembered than "scrawl". This may be due to the sentence contexts: "He scrawled his name at the bottom" provides clearer context than "He enjoyed exchanging banter with the customers", which allows for varied interpretations due to ambiguous context. Incorrect responses included terms like "交易 (transaction)", "货物" (goods)", "讨价还价 (bargain)", "友善" (kindness)", "零钱" (changes)", "柜台 (counter)", and "小道消息 (gossip)". It seems the context of a word can significantly influence its retention, irrespective of the tools employed.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

4.2 The productive task

In the main test (Table 6), the Kimi group scored higher (M = 8.697, SD = 0.951) than the NCD group (M = 7.939, SD = 1.144) in the productive task. An independent-samples *t*-test indicated a significant difference between the groups (t = 2.925, p = 0.005, df = 64, two-tailed; Cohen's d = 0.72) (Table 7). As evident from Figure 3, Kimi maintained consistent success rates across items, while NCD displayed more fluctuation in performance.

Table 6: Group Statistics (Max = 10) (productive task, main test)

	Class	N	Mean	Std. Deviation	Std. Error Mean
Score	Kimi	33	8.697	0.951	0.166
	NCD	33	7.939	1.144	0.199

Table 7: Independent Samples Test (productive task, main test)

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Score	Equal variances assumed	0.015	0.902	2.925	64	0.005	0.758	0.259	0.240	1.275
	Equal variances not assumed			2.925	61.944	0.005	0.758	0.259	0.239	1.275

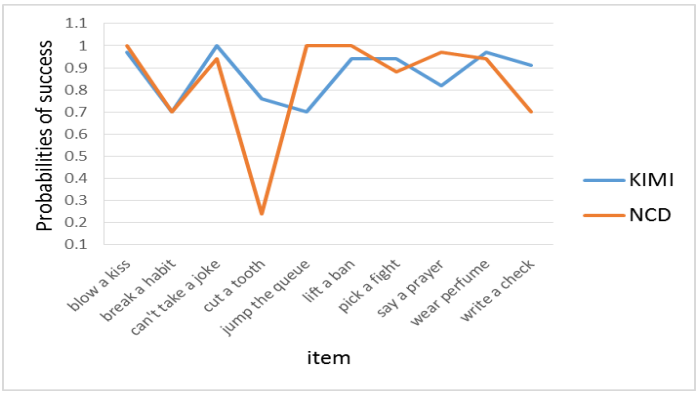

Figure 3: Probabilities of success for each item in the productive main test

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

In the retention test (Table 8), the Kimi group retained significantly more productive collocational knowledge (6.152/8.697, 70.7%) compared with the NCD group (3.606/7.939, 45.4%), with a 25.3% difference in retention rates. An independent-samples *t*-test confirmed a large, statistically significant advantage for Kimi (t = 4.085, df = 64, p < 0.001, two-tailed; Cohen's d = 1.006) (Table 9). Figure 4 further depicts Kimi's consistent retention versus NCD's item-specific fluctuations.

Table 8: Group Statistics (Max = 10) (productive task, retention test)

	Class	N	Mean	Std. Deviation	Std. Error Mean
Score	Kimi	33	6.152	2.949	0.513
	NCD	33	3.606	2.030	0.353

Table 9: Independent Samples Test (productive task, retention test)

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	Upper
Score	Equal variances assumed	7.042	0.010	4.085	64	0.000	2.545	0.623	1.300	3.790
	Equal variances not assumed			4.085	56.770	0.000	2.545	0.623	1.297	3.793

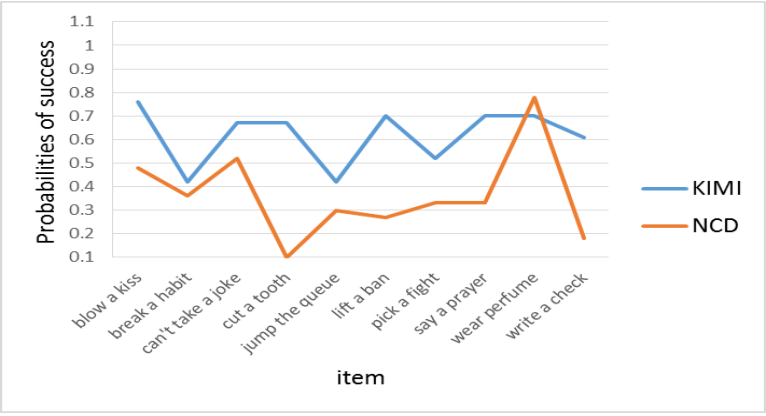

Figure 4: Probabilities of success for each item in the productive retention test

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

Compared with the receptive task, the productive task was more challenging for participants due to the higher cognitive demand required for production. Table 10 illustrates the differential performance of the Kimi and NCD groups on each target collocation in both the main and retention tests. In the main test, Kimi users scored higher on some collocations, whereas NCD users excelled in others. However, in the retention test, the Kimi group consistently outperformed the NCD group across all target collocations, with score differences between 0.04 and 0.43 points.

Table 10: Mean difference in each test item in main and retention tests (productive task)

Test item	Tool	Mean score (main test)	Mean difference (main test)	Mean score (retention test)	Mean difference (retention test)
插队 jump the queue	Kimi	0.70	- 0.30	0.42	0.12
	NCD	1.00		0.30	
开支票 write a cheque	Kimi	0.91	0.21	0.61	0.43
	NCD	0.70		0.18	
寻衅 pick a fight	Kimi	0.94	0.06	0.52	0.19
	NCD	0.88		0.33	
祷告 say a prayer	Kimi	0.82	- 0.15	0.70	0.37
	NCD	0.97		0.33	
戒掉习惯 break a habit	Kimi	0.67	-0.03	0.42	0.06
	NCD	0.70		0.36	
解除禁令 lift a ban	Kimi	0.94	0.06	0.70	0.42
	NCD	1.00		0.28	
长牙齿 cut a tooth	Kimi	0.88	0.43	0.48	0.09
	NCD	0.45		0.39	
送飞吻 blow a kiss	Kimi	0.97	- 0.03	0.76	0.18
	NCD	1.00		0.48	
抹香水 wear perfume	Kimi	0.97	0.03	0.70	0.04
	NCD	0.94		0.66	
开不起玩笑 can't take a joke	Kimi	1.00	0.04	0.67	0.15
	NCD	0.94		0.52	

In the main test, the Kimi group scored lowest on "戒掉习惯" (0.67 points) and "插队" (0.70 points). For "戒掉习惯," 11 students wrote "quit a habit" instead of the correct "break/kick a habit" generated by Kimi. Since "quit" is a direct translation of "戒掉," these students might have instinctively paired "quit" with "habit," considering it a natural collocation. This suggests that they did not seek Kimi's assistance.

For "插队", nine students in the Kimi group chose "cut" instead of "jump," resulting in the incorrect collocation "cut the queue." Actually, Kimi provided two translations: "cut in line" and "jump the queue", along with detailed explanations and sentence examples. However, these students were not sufficiently attentive when reading the information.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

The NCD group scored very low on "长牙齿" (0.45 points), with only 19 out of 33 students providing correct responses. Eight students used "teethe," creating the erroneous phrase "teethe another tooth." According to the questionnaire, some students reported difficulty locating the collocation in NCD, as it was placed in a less accessible Sentence Bank Module which requires additional navigation. This user-unfriendly arrangement clearly affected the students' performance.

Another low-scoring collocation was "开支票" (0.7 points), with eight students using "invoice a cheque" instead of "write a cheque." These students probably knew that "invoice" means "开支票" and assumed "invoice a cheque" is an acceptable combination. Had they consulted the noun "cheque" in NCD, they would have found the correct collocation.

In the retention test, the NCD group had the lowest scores on "开支票" (0.18 points) and "解除禁令" (0.28 points). For "开支票," ten students left it unanswered, and others wrote incorrect phrases like "open a cheque" or "invoice a cheque." For "解除禁令", four left it blank, and others used wrong verbs like "abandon the ban", "stop the ban", "forbid the ban," or "loose the ban". These low retention scores may be attributed to the incongruent nature of these two collocations between English and Chinese.

4.3 Responses to the questionnaires

Questionnaire responses regarding Kimi

Among the 33 respondents, 63.6% (n = 21) had no prior experience with Kimi, though some were aware of it. Only a few used it regularly (15.2%, n = 5) or occasionally (18.9%, n = 6) for EFL learning, with one participant using it solely for non-linguistic purposes. In contrast, 60.6% (n = 20) regularly used other AI chatbots for EFL learning, primarily for tasks including vocabulary acquisition, grammar clarification, translation, writing support, speech drafting, text analysis, and error correction. A smaller subset (33.3%, n = 11) used AI tools sporadically for language learning, while two respondents reported non-linguistic applications.

Regarding lexical information provision for tasks in the main test, 69.7% (n = 23) affirmed that Kimi generated all required data, while 30.3% (n = 10) noted near-complete coverage. Satisfaction levels with the quality of information were notably high: 33.3% (n = 11) expressed strong satisfaction, 63.6% (n = 21) reported satisfaction, and only one respondent remained neutral.

All participants found Kimi easy to access, with 63.6% (n = 21) describing it as "convenient" and 36.4% (n = 12) as "highly convenient." Nevertheless, while 63.6% (n = 21) experienced smooth interactions, 33.3% (n = 11) reported uncertainty in formulating effective prompts, and one noted insufficient content depth. Compared with traditional dictionary apps, 75.8% (n = 25) of respondents rated Kimi as "far superior", 18.2% (n = 6) as "slightly better", and two considered it "comparable".

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

When it comes to the strengths and weaknesses of AI chatbots versus traditional dictionary apps, a clear consensus emerged among the respondents. They generally agreed that AI chatbots have three main advantages. To begin with, AI tools can provide prompt and instantaneous responses to queries, which is highly convenient. In addition, they can engage users in interactive learning through conversations, making the learning process more engaging and tailored to individual needs. Lastly, AI tools can offer a broader range of information than any single dictionary app.

Conversely, AI tools also have three main disadvantages compared with dictionary apps. First, the quality of the generated content depends on the prompts provided by users. Inadequate prompts may result in simplistic or superficial information. Secondly, while dictionaries present all relevant lexicographical information in a single entry, chatbot users must formulate prompts to elicit specific information. Thirdly, AI chatbots typically provide only textual information for lexical reference, whereas dictionary apps often offer visual aids such as images and audio pronunciations, which can enhance the learning experience.

Questionnaire responses regarding NCD

According to the questionnaire on NCD, all the 33 respondents reported no prior experience with this tool. In terms of dictionary information accuracy, the majority rated it positively: 33.3% (n = 11) described it as "very good," and 57% (n = 19) deemed it "good," while only three participants (9.1%) considered it "neutral." Similarly, assessments of the richness of dictionary content confirmed broad satisfaction, with 21.2% (n = 7) expressing "high satisfaction" and 57.6% (n = 19) indicating "satisfaction." The remaining 21.2% (n = 7) found it "average".

Regarding usability, 30.3% (n = 10) of respondents were "very satisfied" with NCD's interface and functionality, while 48.5% (n = 16) found it "satisfactory." A minority (21.2%, n = 7) felt the usability was merely "neutral". When asked about the effectiveness of NCD for vocabulary tasks, 27.3% (n = 9) reported being "very satisfied," and 66.7% (n = 22) expressed general satisfaction. Only two participants (6%) remained neutral.

Despite these positive trends, 66.7% of respondents (n = 22) faced challenges while using NCD dictionary use. Common issues included the absence of full-sentence translation capabilities, difficulties locating English equivalents for specific Chinese collocations, and lack of English definitions alongside Chinese ones. However, a strong majority (78.8%, n = 26) still considered NCD superior to mainstream alternatives like Youdao, Eduict, and Baidu.

Suggestions for improvement were provided by 81.8% (n = 27) of participants. Key recommendations included expanding vocabulary coverage, integrating sentence-level translation features, enhancing access to collocations and phrases, redesigning the interface for greater intuitiveness, adding English explanations for Chinese entries, and incorporating tools to support word retention.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

A comparative analysis

From the questionnaire responses, several similarities were identified between Kimi and NCD users. Despite having limited or no prior experience with the respective tools, both groups generally provided positive evaluations regarding the quality and quantity of retrieved information, as well as the usability and usefulness of the tools for task fulfilment. They found the tools more useful for task fulfilment than other dictionary apps they usually use. Moreover, most users in both groups had balanced perceptions of the strengths and limitations of the tools. These findings may partially explain the better results of the current study relative to previous research, as will be discussed in the next section.

Significant differences also emerged between Kimi and NCD users. Kimi users reported higher levels of satisfaction with the quality of output and the usability of the tool compared with NCD users. For example, 100% of Kimi users rated the tool as convenient or highly convenient, while seven NCD users rated it as neutral. Each group encountered different obstacles in using the tools. In addition, NCD users found dictionary use more challenging than Kimi users. These disparities likely contributed to the better performance of Kimi users in vocabulary comprehension, production and productive retention.

5. Discussion

5.1 Tool effectiveness for vocabulary reception

To address the first research question, the study offers convincing proof that Kimi significantly outperforms NCD in facilitating vocabulary comprehension. On one hand, the advantages of Kimi may chiefly stem from its powerful capabilities in generating interactive and contextually relevant content in response to users' inquiries. These features enable users to understand word meanings more accurately and efficiently than traditional dictionaries, leading to better comprehension outcomes. On the other hand, NCD users exhibited inadequate dictionary usage skills, which hindered their ability to leverage the dictionary for lexical reception. As detailed in Section 4.1, some students were unable to discern the appropriate meaning of a polysemous word based on contextual cues and some failed to grasp word definitions accurately. These inadequacies obviously impacted on the dictionary's effectiveness for vocabulary reception.

The current study achieved higher success rates for both tools in the receptive task (Kimi: 100%, NCD: 97.0%) compared with Ptasznik et al.'s (2024) research, where the dictionary group had a 73% mean success rate and ChatGPT attained 87%. This discrepancy primarily stems from task difficulty differences. Ptasznik et al.'s task was more challenging, featuring more complex target words (e.g., "sycophant," "circumlocution", "equanimity," "boondoggle") and twice as many items (twenty vs. ten in the current study). Additionally, the sentences employed exhibited greater syntactic complexity and length.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

5.2 Tool effectiveness for vocabulary production

In answering the second research question, the study illustrates that Kimi notably surpasses NCD in supporting vocabulary production, maintaining a more consistent performance across ten target collocations. These results may be attributed to Kimi's advanced capabilities to generate context-specific suggestions through interactive dialogue. Users can receive immediately relevant feedback, free from error-prone and time-consuming dictionary searches (Ptasznik et al. 2024: 334). In contrast, NCD offers a broader range of options without the same contextual guidance, resulting in more variability in student performance.

The study identifies several reasons for the relatively low scores on certain test items in the productive task. Participants often viewed collocations as arbitrary word pairings rather than conventionalized patterns. This misconception led some students to overestimate their collocation knowledge and underestimate the need for reference tools. In addition, a lack of attentiveness when interpreting accessed information resulted in avoidable mistakes. Moreover, the unintuitive layout and presentation of some entry information in NCD made it difficult for students to quickly and accurately locate what they needed.

The current study demonstrated higher success rates in the productive task relative to Ptasznik et al. (2024), with Kimi achieving 87% and NCD 79% efficacy, surpassing their reported results of 53% for the dictionary group and 81% for ChatGPT. Similar to the receptive task, the improved outcomes in production may be attributed to differences in task design. The current study focused on ten collocations comprising high-frequency verbs and nouns, whereas Ptasznik et al.'s research involved twenty highly polysemous verbs in complex verb complementation patterns, such as "see something out," "carry something off," "play up to somebody," and "fix somebody up with somebody". Evidently, the simpler nature of the current productive task reduced cognitive demand, thereby enhancing performance.

5.3 Tool effectiveness for vocabulary retention

Regarding the third research question, the study offers interesting findings on the tools' effectiveness for lexical retention. For receptive retention, no significant difference was found between the Kimi and NCD groups, indicating that both tools are equally effective in helping users remember word meanings. Specifically, NCD users achieved a slightly higher retention rate (78.13%) than Kimi users (76.67%), probably as a result of inter-learner strategy variation. Nevertheless, the latter exhibited more consistent performance across ten target collocations.

In the productive task, Kimi achieved a retention rate nearly 25% higher than NCD (70.7% vs. 45.4%), affirming its superior ability to help users recall collocations. The Kimi group also exhibited more consistent performance across different items. These differences may be ascribed to Kimi's strong interactivity, which

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

can reduce users' cognitive load and enhance their engagement with the tool, thereby improving memory retention.

Two factors were identified as influencing word retention. One is the context in which a word appears. It seems more specific context cues improve retention by helping students associate words with relevant contexts. The other is L1–L2 congruency. Incongruent collocations between English and Chinese, which lack direct equivalents, tend to have lower retention rates due to the difficulty in establishing meaningful connections.

Compared with the author's previous study (Chen 2017), which used an electronic dictionary and reported a retention rate of 36.1%, the current study reaped much better results (Kimi: 70.7%; NCD: 45.4%). This improvement is due to the advanced technological tools employed in the current study, which offer more sophisticated features than the traditional electronic dictionary used previously.

6. Conclusion

The study provides compelling evidence for the superior effectiveness of Kimi over NCD in facilitating vocabulary comprehension and production, echoing the findings of Ptasznik et al. (2024) and Lew (2024) on the advantages of AI chatbots over traditional dictionaries. Notably, Kimi excels in retaining productive collocation knowledge. These results highlight the transformative potential of generative AI tools as valuable complements to conventional lexicographical resources in language-learning contexts. Beyond vocabulary acquisition, AI chatbots like Kimi can play a pivotal role in diverse linguistic activities, including translation, writing, and reading, by offering real-time, context-aware assistance that adapts to learners' needs.

However, the effectiveness of these tools hinges on users' ability to craft precise prompts, a skill that requires explicit training. To maximize the utility of AI chatbots, it is crucial to integrate AI literacy modules into language curricula. These modules should cover prompt engineering, critical evaluation of AI-generated content, and ethical considerations. Meanwhile, developers should refine AI algorithms, enhance computing power, and improve inference abilities to better understand user inputs and meet their needs. Furthermore, incorporating multimodal features like audio-visual aids could significantly enhance the usability of AI models.

While AI chatbots hold great potential for language learning, learners should caution against overreliance. These tools can sometimes generate inaccurate, biased or even hallucinated content. Traditional dictionaries, on the other hand, still remain indispensable because they can guarantee linguistic accuracy and preserve nuanced semantic distinctions. Future research should focus on combining AI's interactive features with the reliable editing standards of dictionaries. It is also important to gauge the long-term effects of using AI tools on learners' ability to think independently and critically. Such efforts will ensure that tech-

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

nological advancements align with pedagogical goals, creating a balanced and sustainable environment for language learning.

Meanwhile, in the era of AI, to boost the competitiveness of dictionary apps against AI-powered resources, interdisciplinary collaboration is crucial. Experts from lexicography, education, information technology, and particularly the AI industry should work together to develop AI-enhanced dictionaries. The form, content, access structure, and functionalities of dictionaries need substantial upgrades to withstand the competitive onslaught of intelligent tools and ensure their continued relevance and utility. The integration of AI technologies can facilitate the provision of more contextually relevant examples, interactive learning features, and personalized recommendations, thereby significantly improving the overall learning experience for users.

7. Limitations and suggestions for future research

The study is not without its limitations. It focused on outcomes rather than the processes of tool use. Recording and analysing students' prompts could provide deeper insights into how they construct queries and use AI-generated responses. Future research could explore in depth the interaction process between AI models and users. This could employ qualitative research methodologies, such as case studies, interviews, observational studies, and think-aloud protocols to capture the varied experiences of individuals as they interact with these tools.

Another limitation of the study is that, each task included only ten target items, which may limit the generalizability of the findings. Considering the limited memory capacity and attention span of learners within a relatively short period of time, it is reasonable to involve twenty lexical items in a twenty-minute task session. However, the limited number of target items might influence the representativeness of results. Future research could explore the effects of varying the number of target items within certain time frames to determine the optimal balance between quantity and retention efficiency.

Furthermore, the study employed a between-group design rather than a within-group design. While this approach allowed for a comparison of different tools across distinct groups, it may have overlooked the nuanced differences in how the same learners interact with and evaluate multiple tools. Future research could adopt a within-group design to compare the same group's usage patterns and evaluations of different tools within the same set of tasks. This would provide a deeper grasp of each tool's strengths and weaknesses, as well as how learners adapt their strategies when switching between tools.

To extend the research scope of the present study, further investigations could involve a diverse array of language learning activities, such as reading, writing, and translation. By employing a variety of chatbots and dictionaries with differing features and capabilities, researchers could gain more insights into the efficacy of these tools across various linguistic tasks and learner profiles.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

Given the rapid advancement of AI technologies, it is also imperative to explore the evolving role of dictionaries in language learning. As AI continues to permeate and transform the landscape of language resources, the traditional functions and usage patterns of dictionaries are likely to undergo significant shifts. Research should therefore focus on tracking and analysing the changes in users' dictionary use habits and preferences over time, as well as the potential impact of these shifts on language learning outcomes.

Acknowledgements

The author extends deep gratitude to the anonymous peer reviewers and André du Plessis, editor of *Lexikos*, for their invaluable comments and suggestions on the previous drafts of this article. Particular appreciation is directed to the participants of the study for their indispensable contributions. Special thanks are also due to my colleagues, Dr. Lin Lin and Ms. Wang Rui for their assistance in collecting the experimental data.

Funding disclosure statement

This work was supported by the Fujian Social Science Foundation under Grant FJ2023B031.

GenAI use disclosure statement

During the manuscript preparation, the author utilized Kimi (a free LLM accessible at https://kimi.moonshot.cn/) solely for enhancing readability and language polishing. The author carefully reviewed and edited all AI-generated content and assumes full responsibility for the final text.

References

Alzi'abi, S.E. 2017. Guessing Verb–Adverb Collocations: Arab EFL Learners' Use of Electronic Dictionaries. *Lexikos* 27: 50-77.

Chen, Yuzhen. 2010. Dictionary Use and EFL Learning: A Contrastive Study of Pocket Electronic Dictionaries and Paper Dictionaries. *International Journal of Lexicography* 23(3): 275-306.

Chen, Yuzhen. 2012. Dictionary Use and Vocabulary Learning in the Context of Reading. *International Journal of Lexicography* 25(2): 216-247.

Chen, Yuzhen. 2017. Dictionary Use for Collocation Production and Retention: A CALL-based Study. *International Journal of Lexicography* 30(2): 225-251.

Chen, Yuzhen. 2020. The Effectiveness of Dictionary Use for Collocation Production and Retention: A Case Study of Smartphone Online Dictionaries. *Lexicographical Studies* 2: 20-31.

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.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

De Schryver, G.-M. and D. Joffe. 2023. *The End of Lexicography, Welcome to the Machine: On How ChatGPT Can Already Take over All of the Dictionary Maker's Tasks.* Conference paper. 20th CODH Seminar, Center for Open Data in the Humanities, Research Organization of Information and Systems, National Institute of Informatics, Tokyo, Japan, 27 February 2023. http://codh.rois.ac.jp/seminar/lexicography-chatgpt-20230227/.

Dziemianko, A. 2010. Paper or Electronic? The Role of Dictionary Form in Language Reception, Production and the Retention of Meaning and Collocations. *International Journal of Lexicography* 23(3): 257-273.

Dziemianko, A. 2015. Colors in Online Dictionaries: A Case of Functional Labels. *International Journal of Lexicography* 28(1): 27-61.

Dziemianko, A. 2017. Dictionary Form in Decoding, Encoding and Retention: Further Insights. *ReCALL* 29(3): 335-356.

Dziemianko, A. 2022. The Usefulness of Graphic Illustrations in Online Dictionaries. *ReCall* 34(2): 218-234.

Fuertes-Olivera, P.A. 2024. Making Lexicography Sustainable: Using ChatGPT and Reusing Data for Lexicographic Purposes. *Lexikos* 34(1):123-140.

Huete-García, Á. and S. Tarp. 2024. Training an AI-based Writing Assistant for Spanish Learners: The Usefulness of Chatbots and the Indispensability of Human-assisted Intelligence. *Lexikos* 34(1): 21-40.

Jakubíček, M. and M. Rundell. 2023. The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-editing Lexicography? Medveď, M. et.al. (Eds.). 2023. *Electronic Lexicography in the 21st century (eLex2023): Invisible Lexicography. Proceedings of the eLex 2023 Conference, Brno, 27–29 June 2023:* 518-533. Brno: Lexical Computing CZ s.r.o.

Lew, R. 2023. ChatGPT as a COBUILD Lexicographer. *Humanities and Social Sciences Communications* 10(704):1-10.

Lew, R. 2024. Dictionaries and Lexicography in the AI Era. *Humanities and Social Sciences Communications* 11(1): 1-8.

Lew, R. and J. Doroszewska. 2009. Electronic Dictionary Entries with Animated Pictures: Lookup Preferences and Word Retention. *International Journal of Lexicography* 22(3): 239-257.

Lew, R., B. Ptasznik and S. Wolfer. 2024. The Effectiveness of ChatGPT as a Lexical Tool for English, Compared with a Bilingual Dictionary and a Monolingual Learner's Dictionary. *Humanities and Social Sciences Communications* 11(1): 1-10.

Li, Lingling and Hai Xu. 2015. Using an Online Dictionary for Identifying the Meanings of Verb Phrases by Chinese EFL Learners. *Lexikos* 25:191-209.

Liu, Dilin, Yaochen Deng and Shiyan Yang. 2021. Evaluating Popular Online English–Chinese Dictionaries in China by Applying Lew and Szarowska's (2017) Evaluation Framework. *International Journal of Lexicography* 34(2): 157-182.

Liu, Xiqin, Dongping Zheng and Yushuai Chen. 2019. Latent Classes of Smartphone Dictionary Users among Chinese EFL Learners: A Mixed-method Inquiry into Motivation for Mobile Assisted Language Learning. *International Journal of Lexicography* 32(1): 68-91.

Longman Dictionary of Contemporary English: https://www.ldoceonline.com

Ma, Qing. 2019. University L2 Learners' Voices and Experience in Making Use of Dictionary Apps in Mobile Assisted Language Learning (MALL). *International Journal of Computer-Assisted Language Learning and Teaching* 9(4): 18-36.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

McKean, E. and W. Fitzgerald. 2024. The ROI of AI in Lexicography. *Lexicography* 11(1): 7-27.

Oxford Advanced Learner's Dictionary: https://www.oxfordlearnersdictionaries.com

Ptasznik, B., S. Wolfer and R. Lew. 2024. A Learners' Dictionary Versus ChatGPT in Receptive and Productive Lexical Tasks. *International Journal of Lexicography* 37(3): 322-336.

Phoodai, C. and R. Rikk. 2023. Exploring the Capabilities of ChatGPT for Lexicographical Purposes: A Comparison with *Oxford Advanced Learners' Dictionary* within the Microstructural Framework. Medveď, M. et al. (Eds.). 2023. *Electronic Lexicography in the 21st Century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 Conference, Brno, 27–29 June 2023:* 345-375. Brno: Lexical Computing CZ s.r.o.

Rees, G.P. and R. Lew. 2024. The Effectiveness of OpenAI GPT-generated Definitions Versus Definitions from an English Learners' Dictionary in a Lexically Orientated Reading Task. *International Journal of Lexicography* 37(1): 50-74.

Rundell, M. 2023. Automating the Creation of Dictionaries: Are We Nearly There? *Proceedings of the 16th International Conference of the Asian Association for Lexicography (Asialex 2023 Proceedings), 22–24 June 2023, Seoul, Korea: Lexicography, Artificial Intelligence, and Dictionary Users:* 9-17. Seoul: Yonsei University.

Schmitt, N., D. Schmitt and C. Clapham. 2001. Developing and Exploring the Behaviour of Two New Versions of the Vocabulary Levels Test. *Language Testing* 18(1): 55-88.

Tarp, S. and A. Nomdedeu-Rull. 2024. Who has the Last Word? Lessons from Using ChatGPT to Develop an AI-based Spanish Writing Assistant. *Círculo de lingüística aplicada a la comunicación* 97: 309-321.

Yuan, Yulin. 2023. Theoretical Reflections on Linguistic Studies against the Background of AI Great Leap Forward. *Chinese Journal of Language Policy and Planning* 46(4): 7-18.

Zhao, Chong. 2023. *Research on Contextual Embedding for Sense Identification in Learner's Dictionaries.* Beijing: Beijing Foreign Studies University.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

Appendix A

The receptive lexical task

Directions: Please provide the meaning of the target word that is highlighted in the sentence. You may explain it in either Chinese or English.

1. He **dabbles** in local politics.
   dabble : _____
2. He **scrawled** his name at the bottom.
   scrawl: _____
3. There are at least twenty restaurants **vying** with each other for custom.
   vie: _____
4. He hated his two-year **stint** in the Navy.
   stint: _____
5. The boss is a **wily** old fox.
   wily: _____
6. Christine **squirmed** uncomfortably in her chair.
   squirm: _____
7. They face a **torrid** time in tonight's game.
   torrid:_____
8. He enjoyed exchanging **banter** with the customers.
   banter: _____
9. He **swaggered** over towards me.
   swagger: _____
10. His skin was **translucent** with age.
    translucent: _____

The productive lexical task

Directions: Please complete each sentence by filling in the blank with a suitable verb in its correct form to complete an accurate verb–noun collocation. The intended meaning of the collocation is provided at the end of each sentence.

1. An argument developed when she tried to ___ **the queue**. (插队)
2. I had to ____ **a cheque** for £360 yesterday. （开发票）
3. I could see he was trying to___ **a fight** with me. (寻衅)
4. The children ____ their **prayers** and got into bed. (祷告)
5. I've smoked for years, but I really want to ___ **the habit**. (戒掉习惯)
6. There are no plans to ___ **the ban** on the sale of fireworks to children. (解除禁令)
7. Poor little Patrick was _____ **another tooth** and we had hardly had any sleep. （长牙齿）
8. As the train drew away he ____ her **a kiss**. （飞吻）
9. She was ____ too much **perfume**. （抹香水）
10. The trouble with her is she **can't** _____ **a joke**. （开不起玩笑）

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

Appendix B

KIMI AI Assistant Usage Survey

Directions: This survey is designed to gather your feedback on your interaction with Kimi for the lexical tasks you completed just now. Please select your response (✔) or provide comments based on your recent experience. Thank you for your participation.

1. How often do you use KIMI?
   A. Never heard of it B. Heard of it but never used before
   C. Occasionally for non-language learning purposes D. Occasionally for English learning
   E. Frequently for English learning F. Other situations _____
2. How often do you use other AI models (such as ChatGPT, Baidu's Ernie, etc.)?
   A. Never
   B. Occasionally for non-language learning purposes
   C. Occasionally for English learning
   D. Frequently for English learning, such as _____
   E. Other situations _____
3. Did Kimi generate sufficient content needed for the lexical tasks just now?
   A. Yes, it provided all the content needed for the tasks.
   B. Yes, it generated most, but not all the needed information.
   C. No, it barely provided anything useful.
   D. No, it generated some irrelevant content.
4. How satisfied are you with the quality of the content generated by KIMI?
   A. Highly satisfied B. Satisfied C. Neutral D. Dissatisfied E. Extremely dissatisfied
5. How would you evaluate Kimi in terms of ease of use?
   A. Very convenient B. Convenient C. Neutral D. Difficult to access E. Slow in response
6. What obstacles did you encounter during your interaction with Kimi just now? (Multiple choices allowed)
   A. No obstacles, highly smooth interaction B. uncertainty about prompt formulation
   C. Slow response D. Inaccurate content
   E. Incomplete content F. Overloaded content
   G. Irrelevant content H. Complex interface
   I. Other issues _____
7. Compared with traditional dictionary apps, how would you rate KIMI's overall performance?
   A. Far Superior B. Slightly Better C. Comparable D. Slightly Worse E. Far Inferior
8. In your opinion, what are the key advantages and limitations of AI assistants (like Kimi, ChatGPT) compared with traditional dictionaries?
   Advantages: _____
   Limitations: _____

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2027 (Article)

Dictionary Application Usability Survey

Directions: This survey is designed to gather your feedback on the dictionary application you used to complete your tasks just now. Please select your response (✔) or provide comments below based on your recent experience. Thank you for your participation.

1. Have you used this dictionary application before?
   A. Yes B. No
2. How satisfied are you with the accuracy of the dictionary content? (e.g., definitions, examples, pronunciations)
   A. Extremely Satisfied B. Satisfied C. Neutral D. Dissatisfied E. Extremely Dissatisfied
3. How satisfied are you with the comprehensiveness of the information? (e.g., coverage of words, synonyms, usage notes)
   A. Extremely Satisfied B. Satisfied C. Neutral D. Dissatisfied E. Extremely Dissatisfied
4. How satisfied are you with the ease of searching/retrieving information? (e.g., search bar functionality, filters)
   A. Extremely Satisfied B. Satisfied C. Neutral D. Dissatisfied E. Extremely Dissatisfied
5. How satisfied are you with the overall effectiveness in assisting vocabulary tasks? (e.g., interface design, time efficiency, relevance of results)
   A. Extremely Satisfied B. Satisfied C. Neutral D. Dissatisfied E. Extremely Dissatisfied
6. Compared with other dictionary applications you regularly use, how would you rate this app's overall performance?
   A. Far Superior B. Slightly Better C. Comparable D. Slightly Worse E. Far Inferior
7. Did you encounter any difficulties while using this application?
   A. Yes (Please describe: _____)
   B. No
8. Do you have any suggestions for improving this application?
   A. Yes (Please specify: _____)
   B. No