Introduction

Language learners have traditionally used lexical tools, and the canonical tools to use have been dictionaries. As recently as ten years ago (Levy and Steel, 2015), dictionaries were still the most popular type of references consulted in the language-learning context. The final decades of the 20th century have seen some debate as to what dictionary type (bilingual, monolingual, or bilingualized) might serve language learners best, with the focus subsequently shifting to dictionary medium (paper versus digital) and entry navigation devices (menus and signposts).

Lew (2004) provides a fairly comprehensive overview of the research into dictionary type, and reports on an original large-scale study, concluding that, in reception (i.e., understanding additional-language text), bilingual lexical information leveraging the power of the native language is absolutely crucial, and this applies across all proficiency levels (see also Thompson, 1987; Wingate, 2002, p. 23). Such an advantage, however, did not surface from an earlier and widely-cited study by Laufer and Melamed (1994), and that is likely because the bilingual dictionary used in that study was very much limited in scope compared to the monolingual dictionary used in the comparison. By contrast, Lew (2004) controlled for entry coverage and content, with the experimental dictionaries only varying by entry type. Even so, the—modest as it was—bilingual dictionary in Laufer and Melamed (1994) still outperformed the other types in production. This highlights the fundamental problems with monolingual learners' dictionaries when used for production: however rich and helpful the microstructure may be, it is still embedded under a specific word, and it is not going to help the additional-language writer unless they already know which word(s) to use (compare the 'hitch a ride' example in Lew and Adamska-Sałaciak, 2015, p. 53). And, again, this explains the overwhelming preference of language learners for bilingual dictionaries (Atkins and Varantola, 1998; Baxter, 1980; Tomaszczyk, 1979). This tallies well with common classroom experience. Our numerous conversations with practising teachers of English (personal communication) reveal that students often seek the native-language equivalent when confronted by a teacher's paraphrase in English and will not rest until they learn it (sometimes whispered furtively if classroom rules prohibit first-language use). On the other hand, Abecassis (2008, p. 6) contends that, while language learners usually prefer bilingual dictionaries, the latter are "often incompetently used with students only relying on a word-for-word translation or ignoring the whole range of translations a word can have".

As regards the dictionary medium (print or digital), a series of studies (Dziemianko, 2010, 2011, 2012, 2017) returns somewhat conflicting results. It appears that we cannot claim with confidence that either the print or digital medium is more effective for either reception or production. There are also methodological assumptions that impact the results, foremost among these being the fact that digital dictionaries increasingly offer dynamic, adaptive views, perhaps dependent on the context of use, which print dictionaries cannot hope to replicate. It is not obvious if this difference should be controlled to even the field, as it were, or whether it should rather be treated as inherent to the medium. In any case, digital dictionaries have now all but supplanted print dictionaries, thus the question of medium is no longer that relevant. Our study will focus on digital dictionaries.

Just as the transition to the digital format led to a paradigm shift in lexicography (De Schryver, 2003; Lew and De Schryver, 2014), the recent rise of Large Language Models, Generative Transformers, and AI-powered chatbots may well mean an even bigger shift. Interest was spawned in the lexicographic community in applying such technologies to produce dictionary content (De Schryver, 2023; De Schryver and Joffe, 2023; Lew, 2023; Rees and Lew, 2024; Rundell, 2023). However, it has become increasingly obvious that it might make sense to cut the middleman and use a chatbot directly in contexts where, heretofore, one would normally use a dictionary (Lew, 2024). A new question, therefore, arises: how effective would such a generative chatbot be as a lexical tool compared to traditional dictionaries, be it bilingual or monolingual? This is the question that we attempt to answer in the present contribution, addressing both reception and production tasks in English as an additional language, with the help of one of three tools for lexical support: one of two online dictionaries or ChatGPT.

The study

Aim. The aim of the study was to compare the effectiveness in lexically oriented receptive and productive tasks of three tools: ChatGPT-3.5 in its freely available version and two dictionaries: a leading English monolingual learner's dictionary, the *Longman Dictionary of Contemporary English*, or *LDOCE* (LDOCE, 2024), and *Diki.pl* (Diki.pl, 2024), a popular bilingual dictionary between Polish and English. All three tools were accessed via desktop computers.

Participants. Participants were all native speakers of Polish enrolled in a three-year BA program in English at a large state university in Poland, attending study years 1 ($N = 78$), 2 ($N = 46$), and 3 ($N = 42$). Responses were returned from 166 students.

Materials. Approximately a third of our participants ($N = 56$) used an online (mobile web app) version of the Longman Dictionary of Contemporary English on their personal smartphones (LDOCE, 2024). Another third ($N = 55$) used the *Diki.pl* online bilingual dictionary (Diki.pl, 2024). The remaining third ($N = 55$) of our participants used ChatGPT 3.5, normally in its non-subscription (i.e., free) version, as available during the data collection phase between January and April 2024. In a classroom setting, participants were given two challenging lexically oriented tasks on paper: a reception test and a production test (see Supplementary Material). Twenty test items each were used in the production and reception tests. Hence, in total there were forty different test items. Items selected for the two tasks were infrequent phrasal verbs with common verbs. The items were meant to present a challenge and make participants consult the tools assigned to them for the task at hand.

The target sentences in the production test as well as the context (English sentences) provided for the participants in the reception test were adapted from the *Cambridge Advanced Learner's Dictionary* (CALD, 2024; available as part of aggregated content at https://dictionary.cambridge.org/dictionary/english/), *Collins COBUILD Advanced Learner's Dictionary* (COBUILD, 2024; available as part of aggregated content at https://www.collinsdictionary.com/dictionary/english), and the *Oxford Advanced Learners' Dictionary* (OALD, 2024; https://www.oxfordlearnersdictionaries.com/); complete documentation of sources is available as part of Supplementary Material to this article. In a minority of cases, to meet the aims of the study, example sentences taken from the above dictionaries were slightly modified, whereas one example sentence was written by the second author based on corpus-attested uses.

Procedure. The study involved two paper-based tests: a production test and a reception test done with the tools (see Supplementary Material). In the production test, participants translated twenty Polish sentences into English, with the verb part of the English target phrasal verb presented in bold, without

revealing the particle. For the reception test, participants were instructed to make an effort to understand twenty English sentences, focusing on underlined fragments (holding target items), and then translate these fragments into Polish. Each test came in two versions, with different item orders, each separately randomized.

Prior to the tests, participants received 15 minutes of instructions. They were assured of the confidentiality of their data through anonymization techniques. Participants were randomized to one of three online tools: (1) ChatGPT; an online version of the *LDOCE* (a respectable monolingual dictionary for learners of English); or the bilingual dictionary *Diki.pl*. All three tools were consulted via institution-provided desktop computers in the computer lab. A 90-minute time limit was set for completing both the production and reception tests. The participants had already been familiar with using monolingual dictionaries and ChatGPT due to prior discussions and practical sessions at the university.

At the start of each session, the experimenter instructed the participants verbally. The instruction included a practical demonstration involving a sample task relevant to both tests. Participants were told to attempt every item on the test. Those in the ChatGPT group were instructed to interact with the bot using whatever approach they thought might lead to correct answers. Individuals in the two Dictionary groups were reminded to use their designated dictionaries exclusively—either *LDOCE* or *Diki.pl*—and never consult any alternative dictionaries or corpus tools.

Scoring. Scoring was done along more than one parameter.

Target phrasal verb. For production, the use of the target phrasal verb was scored. This was done with the help of regular expressions, and the expressions were tuned in an iterative process whereby Authors 1 and 2 checked for any misclassifications, and Author 1 modified the regular expressions accordingly.

For the reception, regular expressions were also used, similar to the production exercise. Here, however, the goal was to assess that the meaning of the target phrasal verbs was well understood. That understanding could manifest itself through a range of Polish lexical means. Again, a careful iterative process was employed. For reception, the final regular expressions were more elaborate than for production. All regular expression patterns used are available in Supplementary Material.

Meaning conveyed. As part of the production test, we also wanted to see to what extent the availability of the three tools would assist in conveying the given meaning in English, but without requiring the use of the target phrasal verb. Rather, participants were free to express the original meaning by using whatever lexical means as long as the meaning was clear and faithfully rendered. Overall, there were 3639 non-blank responses in the form of English sentences. However, some participants produced identical responses. After de-duplication (also ignoring initial capitalization and final punctuation), we obtained 1639 unique responses. These were then scored for Meaning Conveyed using the Premium ChatGPT service with GPT-4 and custom instructions (2024-04-19). A six-point grading scale from A through F was used; this choice was motivated by the general familiarity of this scale, widely used in the US, to GPT models, but with an intention to conflate the scores into binary values: 'Meaning conveyed satisfactorily' and 'Meaning not conveyed satisfactorily', using the point between grades B and C as the cut-off threshold. The procedure was interactive and used the following initial prompt:

> Let's try something else. You are an English teacher who also understands Polish. For each of the 71 sentences, give a grade, from A to F (like US college grades, no pluses or minuses), reflecting how well each sentence conveys the meaning of this Polish sentence: "Nagle przestała mówić w trakcie swojego przemówienia."
>
> A grade of "A" means: conveys the meaning perfectly
>
> A grade of "F" means: totally fails at conveying the meaning
>
> Ignore the fact that the sentences start with lower case and there's no final punctuation.
>
> Put the grade: A,B,C,D,E, or F after each sentence.
>
> I will paste the 71 sentences next, OK?

The prompts that followed introduced the source Polish sentences in turn, accompanied by a list of all the differing responses to be scored.

The returned scores appeared reasonable at face value, as did the cut-off. To verify their validity and reliability, a random sample of 10% of all the different responses (164 responses) were independently, in a blind procedure, scored by two human judges, two of the authors with native or near-native proficiency in both Polish and English.

The agreement between scores from ChatGPT-4 and both human judges was high, although it was even higher between the two human judges. In more detail (see Table 1), the percent agreement for conflated (binary) scores between the two human judges was 88%, compared to 82% between each of the human judges and ChatGPT-4. The pairwise (bivariate) values of weighted Kappa for the six-way original grades were as follows: 0.89 between the two human judges (henceforth, H1 and H2); 0.73 and 0.72 between ChatGPT-4 and each of the human judges, respectively. As per Landis and Koch (1977), these values represent "almost perfect" or "substantial" agreement. Analogous figures for Spearman correlation, as well as tetrachoric correlation values for conflated (binary) scores, are all given in Table 1. The three-way Light's Kappa was 0.68.

All these measures indicated high agreement between human scores and ChatGPT-4 scores for Meaning conveyed. For

Table 1 Measures of agreement between ratings by two human judges (H1 and H2) and ChatGPT-4.

pair	Measure		
	Pairwise (bivariate) weighted Kappa	Spearman correlation and CI_{95}	Tetrachoric correlation (binary scores) and CI_{95}
H1 v H2	0.89	0.88 [0.84–0.91]	0.94 [0.84–0.91]
H1 v ChatGPT-4	0.73	0.74 [0.66–0.80]	0.74 [0.66–0.80]
H2 v ChatGPT-4	0.72	0.74 [0.66–0.81]	0.74 [0.66–0.81]

completeness, Table 2 reports the three-way confusion matrix for conflated (dichotomized) scores.

In view of the satisfactory human validation of the scores, the total set of 1639 scores was adopted for further analysis, in their conflated binary form.

Data analysis. All data analysis was conducted in the R environment for statistical computing (R Core Team, 2023) (see Supplementary Material). We collected data from 166 participants, and each participant was given 40 items (20 for reception, 20 for production). This would yield 6640 observations, but one participant had blanks in 13 contiguous items, which we thus assumed the participant omitted without attempting to complete. By contrast, we treated any isolated (non-contiguous) blank as a failed attempt at a response. These were relatively rare: 35 blanks over the whole dataset. Thus, there were a total of 6627 non-missing responses.

To estimate success on the reception and production tasks, mixed binary logistic regression models were fitted on the success (versus failure) dichotomous data at the basic item by participant granularity, using the *lme4::glmer* function (Bates et al., 2015). Model selection was guided by theoretical conceptualization, research design, and likelihood-ratio tests, as constrained by model convergence and dispersion (Burnham and Anderson, 2004; Meteyard and Davies, 2020). Model selection was facilitated through the use of the *buildmer* package (Voeten, 2023), using mostly default settings but overriding the default optimizer in favor of the more efficient *BOBYQA* algorithm (Bound Optimization by Quadratic Approximation, Powell, 2009). Model dispersion was computed using the *blmeco::dispersion_glmer* function (Korner-Nievergelt et al., 2015). Further model diagnostics were performed on residuals simulated with the help of the *DHARMa* package (Hartig, 2022). For effect estimation, we also drew on *effects* (Fox and Weisberg, 2019) and *ggeffects* (Lüdecke, 2018). Since we were comparing one chatbot but two dictionaries, ChatGPT was set as the reference level (intercept) against which performance with the dictionaries was compared. The year of study was entered as a factor, with Year 1 set as the reference level. Details of the specific models are given in the "Results" section.

Table 2 Confusion matrix for the two humans (H1, H2) and ChatGPT (C) conflated (dichotomized) scores.

H1	C H2	0	1
0	0	52	16
	1	4	10
1	0	0	5
	1	4	73

Table 3 Observed mean values for target phrasal verb accuracy on production/reception and meaning conveyed.

Tool	Activity	Target item accuracy	Accuracy meaning conveyed
ChatGPT	Production	0.841	0.972
Diki	Production	0.670	0.811
LDOCE	Production	0.400	0.689
ChatGPT	Reception	0.713	
Diki	Reception	0.846	
LDOCE	Reception	0.674	

Results

Overall means categorized by Tool and Activity are given in Table 3. More descriptive detail with visualization will be given in what follows, first for target phrasal verb production, then for reception (understanding), and finally for the accuracy of conveying meaning (with or without the use of the target phrasal verb).

Producing target phrasal verbs. Figure 1 presents observed values as per-item success rates for phrasal verb production, separately for the three tools. Each black dot represents the mean success rate for one Item and a given Tool. The boxplots abide by the usual conventions, with the boxes covering half the data spread, and the inner line drawn at the median. In addition, red diamonds mark the means for each Tool. The distribution of

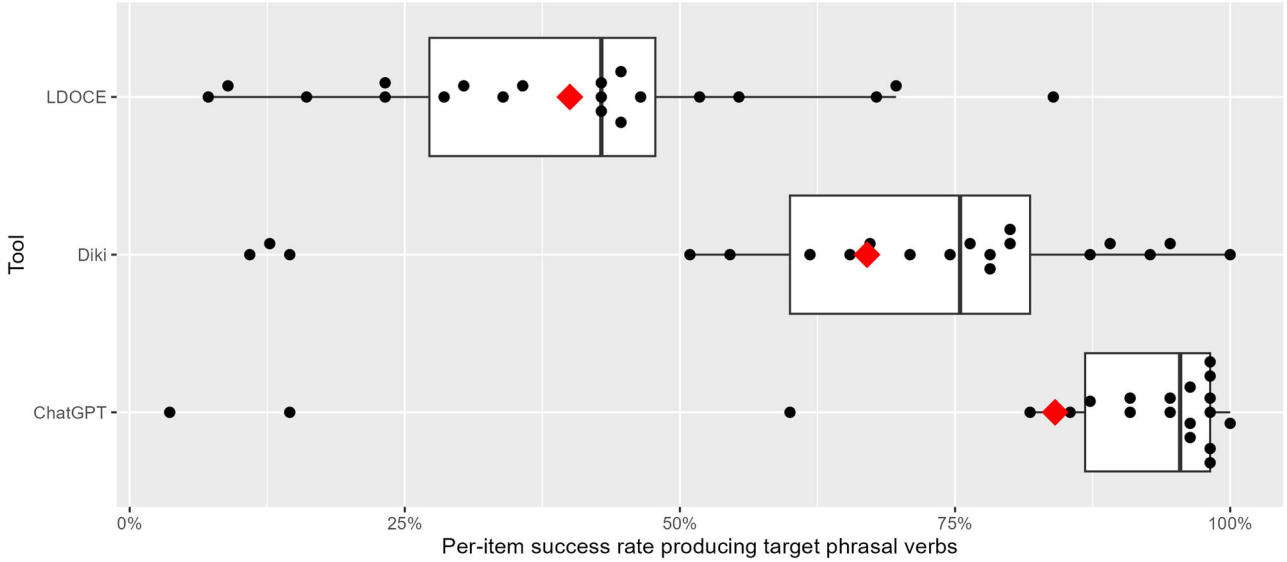

Fig. 1 Boxplots and dot plots for per-item success rates producing target phrasal verbs. Each dot represents the mean success rate for a given Item and Tool combination. Overall means appear as red diamonds.

observed per-item success rates with the three tools suggests a clear progression of success rates, from the very high success rates with ChatGPT (except two or three items), through intermediate success rates with the bilingual Diki dictionary, down to the poorest success using the monolingual *LDOCE* dictionary. To assess these effects while controlling for other intervening variables, let us turn to the modeling.

The model selection procedure concluded with the following glmer formula for Production:

$$\text{Target} \sim 1 + \text{Tool} + \text{Year} + \text{Tool} : \text{Year} + (1 + \text{Tool} \mid \text{Item}) + (1 \mid \text{Participant}),$$

where the binary predicted variable Target indicated whether the Target phrasal verb was successfully used in the response. Estimates of model parameters are given in Table 4. In addition, sum-of-squares partitioning of deviance was done using Wald Chi-squared tests, as implemented in the *car::Anova* (Fox and Weisberg, 2019) function, to confirm the overall significance of the predictors (note that prior confirmation of this status follows from the likelihood-ratio tests conducted as part of model selection). The more conservative Type III testing approach was used, rather than the default Type II Sum of Squares. The overall effect of Tool was highly significant by this test ($\chi^2 = 84.7$, $DF = 2$, $p < 0.001$). The overall effect of Year was not significant, but the Tool:Year interaction was ($\chi^2 = 10.5$, $DF = 4$, $p = 0.03$).

Table 4 Estimated model parameters for success at producing target phrasal verbs.

Terms	Odds ratios	95% CI	*p*
(Intercept)	13.86	4.80–39.97	<0.001
Tool [Diki]	0.14	0.06–0.32	<0.001
Tool [LDOCE]	0.03	0.01–0.08	<0.001
Year [2]	1.28	0.64–2.56	0.485
Year [3]	1.13	0.56–2.27	0.741
Tool [Diki] × Year [2]	1.84	0.75–4.51	0.184
Tool [LDOCE] × Year [2]	0.83	0.35–2.00	0.678
Tool [Diki] × Year [3]	1.35	0.54–3.37	0.517
Tool [LDOCE] × Year [3]	2.29	0.94–5.57	0.068
Varying Effects			
τ00 Participant	0.45		
τ00 Item	4.72		
τ11 Item.ToolDiki	1.77		
τ11 Item.ToolLDOCE	2.03		
ρ01 Item.ToolDiki	−0.64		
ρ01 Item.ToolLDOCE	−0.91		
Marginal R2 / Conditional R2	0.223 / 0.613		

R model formula: Target ~1 + Tool + Year + Tool:Year + (1 + Tool | Item) + (1 | Participant).

Odds Ratios given in Table 4 indicate that, in terms of the odds of success at producing target phrasal verbs, for the reference level of Year-1 students, the bilingual Diki dictionary was about 7 times (= 1/0.14) less effective than ChatGPT, whereas the monolingual *LDOCE* was about 30 times less effective. Year-2 students appeared to fare a bit better with the bilingual dictionary, while Year-3 students managed to reduce the unfavorable odds of using the monolingual learner's dictionary by about twice, and this interaction term was marginally significant.

Results for varying effects (see the lower portion of Table 4) suggest a greater variability of success due to Item (τ00 Item) than due to Participant (τ00 Participant).

Complex interaction effects are easier to conceptualize when visualized using interaction plots of marginal effects. To appreciate this, refer to Fig. 2, which indicates the predicted success at producing target items for each combination of Tool and Year of study, along with 95% confidence intervals. The plot clearly shows the huge general advantage of ChatGPT over *LDOCE*, and a lesser degree of advantage over the Diki bilingual dictionary. At the same time, Year-3 students, being the most advanced, perform relatively better with the monolingual *LDOCE* than the lower years, and this is to be expected. Interestingly, the bilingual dictionary registers the highest average rate of success with Year-2 students, though the differences vis-à-vis other Years of study are not significant.

Next, let us review the results for reception, reflecting how well students were able to understand the meanings of the English phrasal verbs, as indicated by the native-language translation (paraphrase) of the target sentence.

Understanding target phrasal verbs. Figure 3 presents observed values as per-item success rates at understanding target phrasal verbs (reception), separately for the three Tools. Each black dot

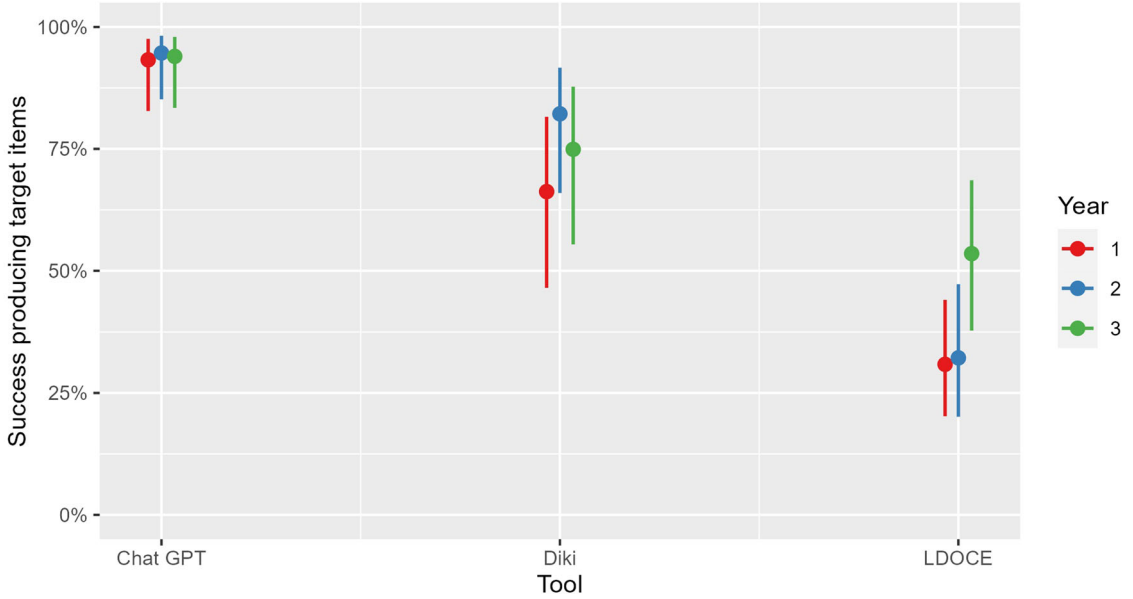

Fig. 2 Success at producing target items by Tool and Year of study. Dots indicate estimated marginal means, and range segments show 95% confidence intervals.

indicates the mean success rate for one Item and a given Tool. As above, the boxplots follow the usual conventions, and we have also added red diamonds to mark the means for each Tool. The pattern suggests a relative advantage of the bilingual dictionary Diki, which registers high scores for all but two items, in which case the dictionary does poorly. The plot for ChatGPT suggests a cluster of eight items with near-perfect success rates, but a wide spread of values for the remaining twelve items, some registering very low success. The median and mean are still somewhat higher than for the monolingual *LDOCE* dictionary, which has relatively fewer extreme (very high and very low) per-item success rates. Next, let us see what the mixed models can tell us about how the Tools performed in terms of helping students understand English phrasal verbs.

The model selection procedure concluded with the following *glmer* formula for Reception:

$$\text{Target} \sim 1 + \text{Tool} + (1 + \text{Tool} \mid \text{Item}) + (1 \mid \text{Participant}),$$

and it is worthy of note that, for reception, the Year of study was not retained in the model: the only variable of interest left in the model was the Tool used in helping students understand the English sentences holding target phrasal verbs. The overall effect of Tool was highly significant by Type III Wald Analysis of Deviance ($\chi^2 = 18.9$, $DF = 2$, $p < 0.001$). The model yielded parameters as set out in Table 5. The Odds Ratios for Tool indicate the odds of understanding a phrasal verb to be close to twice as good with Diki compared to ChatGPT. Conversely, the odds of success with LDOCE are about half those with ChatGPT, and this latter difference was marginally significant.

Just as for production, variability due to Item is quite a bit greater than variability due to Participant (compare τ00 Item and τ00 Participant).

Figure 4 illustrates the marginal means and their 95% confidence intervals for the fixed effect of Tool on the success at understanding the sentences with target phrasal verbs. It will be seen that the use of the bilingual Diki dictionary is associated with the highest marginal mean success rate (91%; $CI_{95\%} = [83\%–95\%]$). The lowest estimated mean was for *LDOCE* (72%; $CI_{95\%} = [61\%–81\%]$). For ChatGPT, the success rate was intermediate but closer to Diki (85%; $CI_{95\%} = [67\%–94\%]$). The lack of overlap between Diki and *LDOCE* indicates that the bilingual dictionary was significantly more effective than the monolingual dictionary on this task.

Table 5 Estimated model parameters for success at understanding target phrasal verbs.

Terms	Odds ratios	95% CI	*p*
(Intercept)	5.5	2.03–14.96	<0.001
Tool [Diki]	1.75	0.73–4.21	0.209
Tool [LDOCE]	0.46	0.21–1.03	0.06
Random Effects			
τ00 Participant	0.22		
τ00 Item	4.74		
τ11 Item.ToolDiki	3.17		
τ11 Item.ToolLDOCE	2.74		
ρ01 Item.ToolDiki	−0.78		
ρ01 Item.ToolLDOCE	−0.89		
Marginal R2 / Conditional R2	0.047 / 0.482		

Formula: Target ~ 1 + Tool + (1 + Tool | Item) + (1 | Participant).

Accuracy at conveying meaning. Our final measure of linguistic success is less related to the form produced and reflects the importance of successful communication, without requiring specific formal elements. Thus, the measure attempts to assess whether the response successfully conveys the assigned meaning in English as an additional language, with or without the use of the target phrasal verb.

Observed per-item success rates at conveying meaning using the three tools are shown in Fig. 5. Clearly, the success is excellent with ChatGPT, and somewhat less so with Diki and (especially) *LDOCE*.

The model selection procedure produced the following final *glmer* formula for Meaning conveyed:

$$\text{Meaning} \sim 1 + \text{Tool} + \text{Year} + (1 \mid \text{Item}) + (1 \mid \text{Participant}),$$

where Meaning stands for the success at conveying the original meaning in English, either with or without using the target

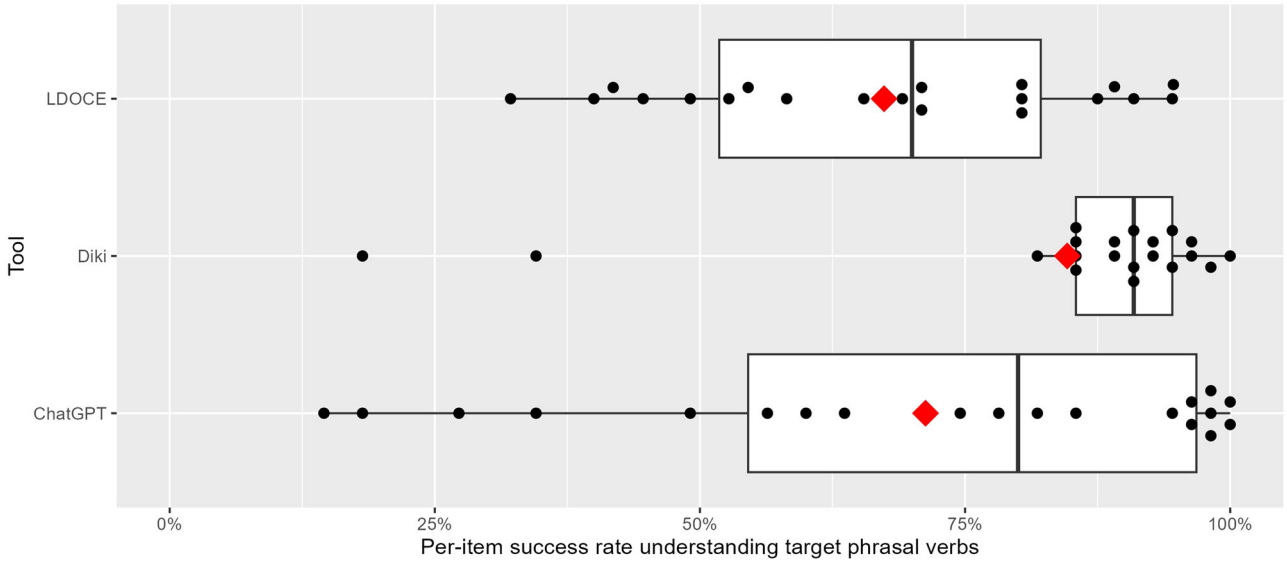

Fig. 3 Boxplots and dot plots for per-item success rates at understanding target phrasal verbs. Each dot represents the mean success rate for a given Item and Tool combination. Overall means appear as red diamonds.

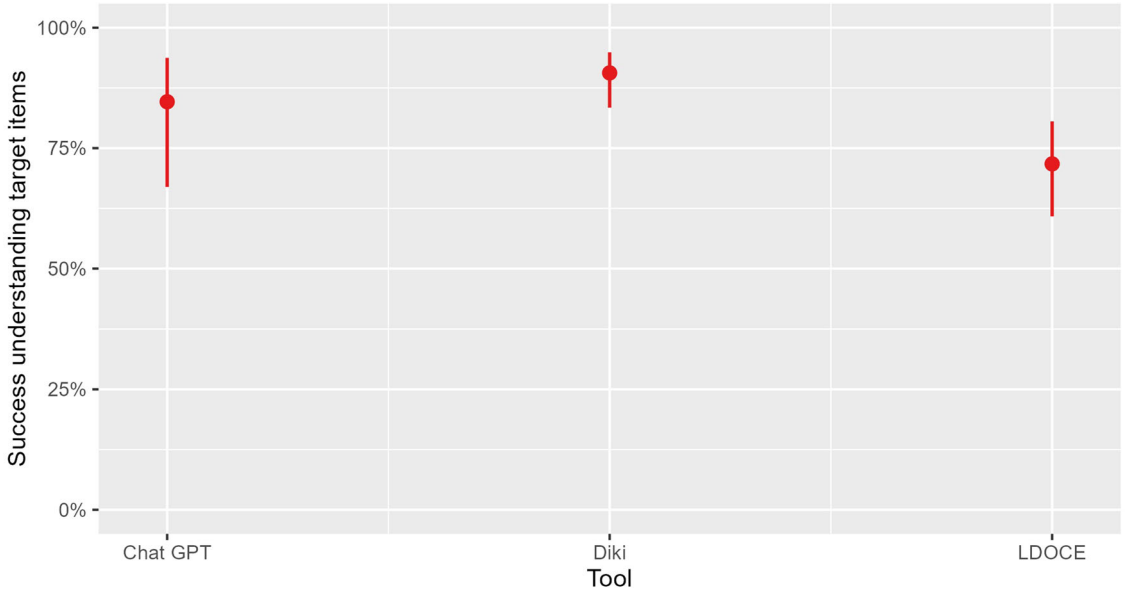

Fig. 4 Success in understanding target items by Tool. Dots indicate estimated marginal means, and range segments show 95% confidence intervals.

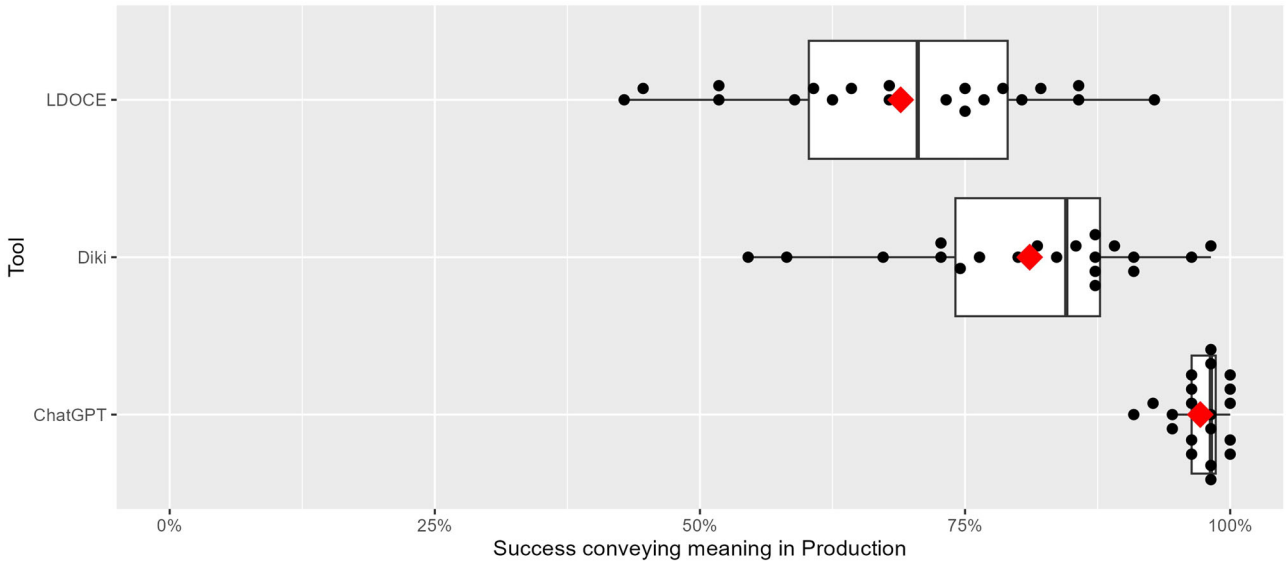

Fig. 5 Per-item success at conveying meaning in production using one of three tools. Each dot represents the mean for a given Item and Tool combination. Overall means appear as red diamonds.

Table 6 Estimated model parameters for accuracy at conveying meaning in English.

Terms	Odds ratios	CI	*p*
(Intercept)	41.16	24.88–68.10	<0.001
Tool [Diki]	0.11	0.07–0.17	<0.001
Tool [LDOCE]	0.05	0.03–0.08	<0.001
Year [2]	1.06	0.78–1.43	0.724
Year [3]	1.81	1.29–2.53	0.001
Varying effects			
τ00 Participant	0.25		
τ00 Item	0.43		
Marginal R2/Conditional R2	0.293 / 0.414		

Formula: Meaning ~ 1 + Tool + Year + (1 | Item) + (1 | Participant).

phrasal verb. The final model did not include an interaction term between Tool and Year, nor a varying slope of Item by Tool.

Model parameters are given in Table 6. The Odds Ratios indicate that the odds of successfully conveying the meaning using ChatGPT were nine times higher than with Diki, and twenty times higher than using *LDOCE*, both these differences being highly significant. In addition, the effect of the Year of study was such that Year-3 students performed nearly twice as well in terms of odds of success. Since the final model does not include an interaction of Tool by Year, the advantage of Year-3 appears to apply irrespective of the Tool used.

Figure 6 represents the additive effects of Tool and Year on the success at conveying meaning in English. The model does not include a Tool-by-Year interaction term. The effect of the Tool is comparatively strong (see Table 6). The respective marginal estimated success rates and their 95% confidence intervals for the

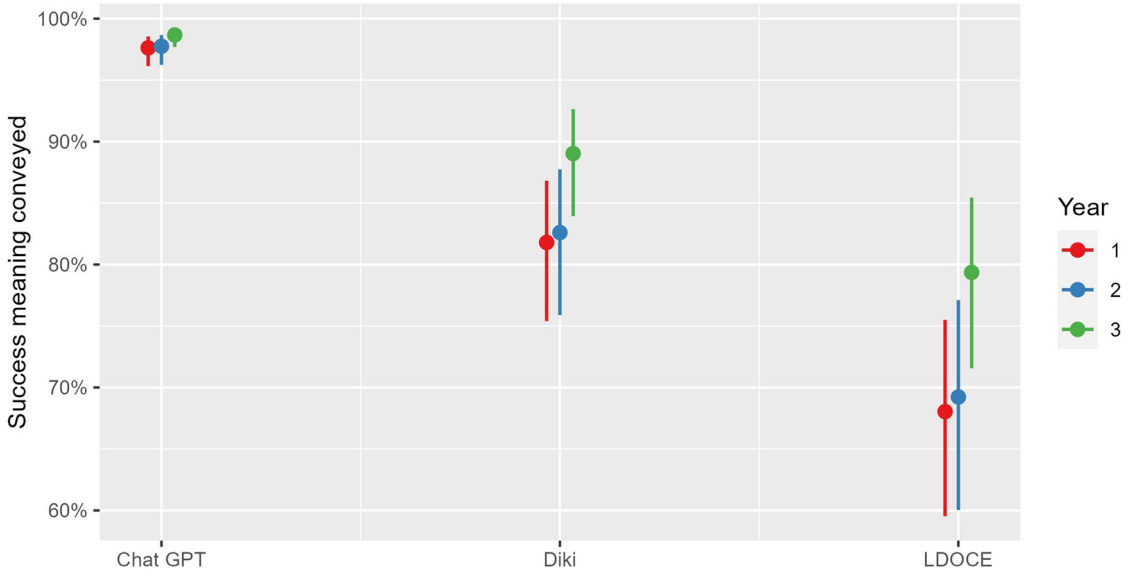

Fig. 6 Success at conveying meaning by Tool and Year. The final model is additive and includes no interaction term. Dots indicate estimated marginal means, and range segments show 95% confidence intervals. Note that, for readability, the vertical scale does not include 0%.

three Tools (averaging over the years) are as follows: ChatGPT 98% [97%–99%]; Diki 84% [79%–88%]; *LDOCE* 72% [64%–78%]. The effect of Year is less pronounced and present for Year-3 students, with the respective estimates for success rates for the three years of study (averaging over tools) being: Year-1 88% [84%–91%]; Year-2 88% [84%–92%]; and Year-3 93% [90%–95%].

Discussion and conclusion

Our results show that, generally speaking, a language-model-based chatbot such as ChatGPT is a serious rival for traditional dictionaries when it comes to helping advanced language learners produce or understand content in a second language, in this case, English. Amongst the three measures used in this study, the largest advantage of ChatGPT vis-à-vis dictionaries was found for assisting students in production: both when a given phrasal verb is expected, and also conveying meaning in English without the requirement that specific lexical items be used. On both production measures, ChatGPT far outperformed both the bilingual (*Diki.pl*) and monolingual (*LDOCE*) dictionaries, but the bilingual dictionary fared better than the monolingual dictionary.

Year-3 students were better at production than the lower study years, which is not at all surprising. However, the advantage held independent of the tool used for conveying meaning, whereas it tended to be restricted to the monolingual dictionary when the use of specific phrasal verbs was required. However, Year-3 students working with *LDOCE* were still at a considerable disadvantage compared to younger students working with ChatGPT, but also those using the bilingual dictionary. This latter result mimics earlier findings by Lew (2004), although, in the present study, it applies to production in English rather than to reception.

ChatGPT was not so clearly advantageous when it came to helping students understand English sentences with difficult phrasal verbs. At this task, it did not perform any better than the bilingual dictionary, though it still outperformed *LDOCE*. In assessing this finding, we must remind ourselves that the explanation of meaning was conducted in the participants' native language, Polish. Now, the GPT models have been trained on predominantly English data, and their performance in other languages, including Polish, is uncontestably inferior (Lai et al., 2023). Looking over the detailed responses of participants in the ChatGPT group, we found quite a few mistranslations, which we were able to replicate later in our own sessions with ChatGPT as problems with chatbot responses, even when using the Premium version running on GPT-4. Non-English language proficiency is clearly a weakness of the currently available versions of ChatGPT. Although work on language models for languages other than English is ongoing, it appears that for such applications, multilingual language models might hold the greatest promise, that is, models that would be highly proficient in (and between) at least two languages.

Going back to traditional lexicographic issues, our study confirms the findings by Lew (2004), re-affirming the role of the native language as a more effective vehicle for meaning indication than the second-language paraphrase which monolingual dictionaries for learners rely on. Our results showed that a bilingual dictionary, even one that is freely available and not coming from a major publisher, helped our advanced learners more than one of the best monolingual learners' dictionaries on every dimension tested.

A further area of interest that should be explored in future research is going beyond immediate success and into measures of learning. It would be interesting to know to what extent working with a chatbot supports language learning, as indicated by delayed retention. On the one hand, asking a bot seems easy, and lexical retention appears to benefit from a certain amount of focused effort, which may not be present as much as for dictionary consultation. On the other hand, interaction with ChatGPT and the like is quite a bit closer in character to interacting with a human, and perhaps this social aspect might promote learning.

Our study suggests that a general-purpose chatbot such as ChatGPT can be a viable alternative to traditional dictionaries in both production and reception tasks in English. The excellent performance of ChatGPT in production is not surprising, given that producing idiomatic, natural English is its first and foremost strength. The relatively poorer success at reception is likely related to the bot's inferior command of Polish. This limitation might be mitigated in the near future as language models proficient in several languages at once become more easily available.

However, for smaller and low-resource languages, dictionaries will probably not be supplanted so readily.

Data availability

The data generated and analyzed during this study are available in the OSF repository at https://osf.io/8egur.

Received: 18 May 2024; Accepted: 12 September 2024;
Published online: 03 October 2024

References

Abecassis M (2008) The ideology of the perfect dictionary: how efficient can a dictionary be? Lexikos 18:1–14. https://doi.org/10.5788/18-0-473
Atkins BTS, Varantola K (1998) Language learners using dictionaries: the final report on the EURALEX/AILA Research Project on Dictionary Use. In: BTS Atkins (ed.), Using dictionaries. Studies of dictionary use by language learners and translators. Niemeyer. pp. 21–81
Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67(1):1–48. https://doi.org/10.18637/jss.v067.i01
Baxter J (1980) The dictionary and vocabulary behavior: a single word or a handful? TESOL Q 14(3):325–336. https://doi.org/10.2307/3586597
Burnham KP, Anderson DR (eds.) (2004) Model selection and multimodel inference. Springer, New York
CALD (2024) Cambridge advanced learners' dictionary. Cambridge University Press. http://dictionary.cambridge.org
COBUILD (2024) Collins COBUILD advanced learners' dictionary. HarperCollins. https://www.collinsdictionary.com/dictionary/english
De Schryver G-M (2003) Lexicographers' dreams in the electronic-dictionary age. Int J Lexicogr 16(2):143–199. https://doi.org/10.1093/ijl/16.2.143
De Schryver G-M (2023) Generative AI and lexicography: the current state of the art using ChatGPT. Int J Lexicogr 36(4):355–387. https://doi.org/10.1093/ijl/ecad021
De Schryver G-M, Joffe D (2023) The end of lexicography, welcome to the machine: On how ChatGPT can already take over all of the dictionary maker's tasks. 20th CODH Seminar, Tokyo. http://codh.rois.ac.jp/seminar/lexicography-chatgpt-20230227/
Diki.pl. (2024) Słownik angielsko-polski, słownik angielski online—Diki. https://www.diki.pl
Dziemianko A (2010) Paper or electronic? The role of dictionary form in language reception, production and the retention of meaning and collocations. Int J Lexicogr 23(3):257–273. https://doi.org/10.1093/ijl/ecp040
Dziemianko A (2011) Does dictionary form really matter? In: Akasu K, Uchida S (eds.). ASIALEX2011 Proceedings. Lexicography: theoretical and practical perspectives. Asian Association for Lexicography. pp. 92–101
Dziemianko A (2012) Why one and two do not make three: dictionary form revisited. Lexikos 22:195–216. https://doi.org/10.5788/22-1-1003
Dziemianko A (2017) Dictionary form in decoding, encoding and retention: further insights. ReCALL 29(3):335–356. https://doi.org/10.1017/S0958344017000131
Fox J, Weisberg S (2019) An R companion to applied regression (third edition). SAGE
Hartig F (2022) DHARMa: residual diagnostics for hierarchical (Multi-level/mixed) regression models. https://CRAN.R-project.org/package=DHARMa
Korner-Nievergelt F, Von Felten S, Roth T, Almasi B, Guélat J, Korner-Nievergelt P (2015) Bayesian data analysis in ecology using linear models with R, BUGS, and Stan. Academic Press
Lai VD, Ngo NT, Veyseh APB, Man H, Dernoncourt F, Bui T, Nguyen TH (2023) ChatGPT beyond English: towards a comprehensive evaluation of large language models in multilingual learning. Find Assoc Comput Linguist 2023:13171–13189
Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33(1):159. https://doi.org/10.2307/2529310
Laufer B, Melamed L (1994) Monolingual, bilingual and 'bilingualized' dictionaries: Which are more effective, for what and for whom? In: Martin W, Meijs W, Moerland M, Ten Pas E, Van Sterkenburg P, Vossen P (eds.). EURALEX '94 Proceedings. Vrije Universiteit. pp. 565–576
LDOCE (2024) Longman dictionary of contemporary English. Pearson Longman. https://www.ldoceonline.com
Levy M, Steel C (2015) Language learner perspectives on the functionality and use of electronic language dictionaries. ReCALL 27(2):177–196. https://doi.org/10.1017/S095834401400038X
Lew R (2004) Which dictionary for whom? Receptive use of bilingual, monolingual and semi-bilingual dictionaries by Polish learners of English. Motivex. https://hdl.handle.net/10593/655
Lew R (2023) ChatGPT as a COBUILD lexicographer. Humanit Soc Sci Commun 10(1):704. https://doi.org/10.1057/s41599-023-02119-6
Lew R (2024) Dictionaries and lexicography in the AI era. Humanit Soc Sci Commun 11(1):426. https://doi.org/10.1057/s41599-024-02889-7
Lew R, Adamska-Sałaciak A (2015) A case for bilingual learners' dictionaries. ELT J 69(1):47–57. https://doi.org/10.1093/elt/ccu038
Lew R, de Schryver G-M (2014) Dictionary users in the digital revolution. Int J Lexicogr 27(4):341–359. https://doi.org/10.1093/ijl/ecu011
Lüdecke D (2018) ggeffects: tidy data frames of marginal effects from regression models. J Open Source Softw 3(26):772. https://doi.org/10.21105/joss.00772
Meteyard L, Davies RAI (2020) Best practice guidance for linear mixed-effects models in psychological science. J Mem Lang 112:104092. https://doi.org/10.1016/j.jml.2020.104092
OALD (2024) Oxford advanced learners' dictionary. Oxford University Press. https://www.oxfordlearnersdictionaries.com
Powell MJD (2009) The BOBYQA algorithm for bound constrained optimization without derivatives. Camb NA Rep. NA2009/06, Univ Camb, Camb 26:26–46
R Core Team (2023) R: A Language and Environment for Statistical Computing [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org
Rees GP, Lew R (2024) The effectiveness of OpenAI GPT-generated definitions versus definitions from an English learners' dictionary in a lexically orientated reading task. Int J Lexicogr 37(1):50–74. https://doi.org/10.1093/ijl/ecad030
Rundell M (2023) Automating the creation of dictionaries: are we nearly there? In: An Y (ed.). Proceedings of ASIALEX 2023. Asialex
Thompson G (1987) Using bilingual dictionaries. Engl Lang Teach J 41(4):282–286. https://doi.org/10.1093/elt/41.4.282
Tomaszczyk J (1979) Dictionaries: users and uses. Glottodidactica 12:103–119
Voeten CC (2023) buildmer: Stepwise elimination and term reordering for mixed-effects regression (Version 2.11) [Computer software]. https://CRAN.R-project.org/package=buildmer
Wingate U (2002) The effectiveness of different learner dictionaries. An investigation into the use of dictionaries for reading comprehension by intermediate learners of German. Niemeyer

Author contributions

Conceptualization: BP RL SW, data curation: RL BP, formal Analysis: RL, funding acquisition: BP SW, investigation: BP, methodology: BP RL SW, project administration: BP, resources: RL BP SW, software: RL, supervision: RL, validation: SW, writing—original draft: RL BP SW, writing—review & editing: RL BP SW.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Ethics approval

This study was performed in accordance with the principles of the Declaration of Helsinki. The Scientific Research Ethics Committee of the University of Warmia and Mazury granted approval (January 22nd, 2024, No. 2/2024).

Informed consent

All participants gave informed consent to participate.

Additional information

Correspondence and requests for materials should be addressed to Sascha Wolfer.