Geraint Paul Rees and Isabel Gibert

A TEXTBOOK OR CHATGPT
Which Helps Novice Programmers Most with Unknown Terms?

Abstract Existing research on ChatGPT in lexicography is undoubtedly valuable. However, it has tended to focus on metalexicographic concerns rather than effectiveness in resolving user queries directly. Moreover, it has mostly dealt with general-purpose English lexicography, often ignoring other languages and specific purposes. Focussing on 33 L1 Spanish users completing an introductory training course on the use of the Python programming language for linguistic research at a Spanish university, this study attempts to fill these gaps. Participants responded to ten multiple-choice questions designed to test understanding of basic programming terms. Approximately half received explanations from a respected introductory Python textbook written in Spanish. The remainder received ChatGPT 3.5-produced explanations written in Spanish. GPT-generated explanations offer performance advantages while textbook definitions offer advantages in processing time. In follow-up interviews, several participants reported feeling overwhelmed by the quantity of explanation provided by ChatGPT.

Keywords chatbots; ChatGPT; specific-purpose lexicography; Spanish; Python; user study

1. Introduction

Large language models paired with chatbots, such as ChatGPT, offer the potential of producing lexicographic resources, cheaply, quickly, and with minimal human involvement. Consequently, it is hardly surprising that since the launch of OpenAI ChatGPT in November 2022, there has been a substantial amount of research on its use for lexicography (c.f. de Schryver, 2023; Lew, 2024). Although undeniably valuable, there are several limitations to existing research. Firstly, much of this research has focused on lexicographers' and lexicography researchers' opinions of GPT's output rather than more direct measures of its effectiveness in resolving users' doubts about word meaning and use (e.g., Jakubíček & Rundell, 2023; Lew, 2023). Secondly, most studies have focused on the production of *general-purpose resources* for English learners (e.g., Rees & Lew, 2024). Ironically, users of these resources are perhaps amongst the best served by existing learners dictionaries and writing assistants. In comparison, the potential advantages offered by GPT in the creation of *specific-purpose lexicographic resources* for *less well-resourced languages* are greater. Thirdly, although much existing research has envisaged a place for GPT in the production of dictionaries, the possibility that users skip the dictionary altogether and go directly to the chatbot to resolve their language doubts has, for the most part, been only grudgingly acknowledged. Drawing a parallel with the shift from paper to online dictionaries, we believe that this is a distinct possibility, especially when one considers

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

the tendency observed in previous research for users to arrive at entry pages directly from search engines (Lorentzen & Theilgaard, 2012).

Concentrating on Spanish lexicography for specific purposes, in an attempt to address these gaps, this paper reports on a study involving post-graduate students at a Spanish university. These participants are L1 users of Spanish who were taking part in an online introductory training course on the use of the Python programming language for linguistic research. For L2 users of English like these novice programmers, the technical challenge of mastering new technology and the intellectual challenge of understanding new concepts are compounded by linguistic challenges. For example, although instructional material is available in Spanish, the meaning of key programming terms which were first coined in English is not always clear. Moreover, L2 English users are also at a disadvantage in understanding documentation and error messages which are often written in English (Becker, 2019; Guo, 2018). Students like these, then, often need to search for explanations of unknown programming terms in their L1. This study examines whether, in doing so, they would be best served by the kind of explanations found in a traditional programming textbook (Marzal & Gracia, 2020) or those generated by ChatGPT. We acknowledge, at the outset, that dictionaries may not be the most likely port of call for novice programmers wishing to clarify the meaning of an unknown term. However, we feel that the parallels between the comparison of information about an unknown term provided by a chatbot with that provided by a more traditional resource on the one hand, and the comparison of information provided in response to a specific language query by a chatbot with that provided by a dictionary, on the other, make this investigation worthwhile. With this in mind, we aim to answer the following research questions:

1. Do ChatGPT-generated explanations contribute to higher success than textbook explanations in a reception task dealing with key programming terms?

2. Do ChatGPT-generated explanations contribute to faster consultations than textbook explanations in a reception task dealing with key programming terms?

2. Methods

2.1 Participants and Setting

The study involved 33 post-graduate students and staff who were studying or teaching on the MA in the teaching of Spanish as a Foreign language or PhD programmes at the Hispanic Studies or Catalan departments at a university in Catalonia. These participants are L1 users of Spanish who were taking part in a voluntary introductory training course on the use of the Python programming language for linguistic research. The course took place online and comprised a two-hour video lecture. Before the class began, students were asked to complete a multiple-choice questionnaire aimed at testing their receptive knowledge of key programming terms. While completing the questionnaire, approximately half these students (n = 16) were

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

provided with definitions from a respected introductory Python textbook written in Spanish (Marzal & Gracia, 2020), while the remainder (n = 17) were provided with explanations produced by ChatGPT 3.5.

2.2 Instruments and Procedure

2.2.1 Multiple-Choice Questions and Explanations

In an attempt to ensure that the terms tested were likely to pose problems for novice Python programmers, we first created a pool of candidate questions inspired by existing online quizzes on basic programming terms. We focused on lexically oriented questions specifically intended to test knowledge of term meaning, leaving aside questions focussing on problem solving or analysis of code. In consultation with the course instructor and two other experienced programmers, we narrowed the pool of potential questions to ten items (see supplementary material). Happily, all the terms dealt with by these questions were present in the textbook from which the traditional explanations were drawn.

The questions were presented to participants using the 1KA online survey platform (*1KA*, 2023). After giving informed consent, participants were shown the first question along with four possible answers and a *no sé* ('I don't know') option. No explanation of the term of interest was provided at this stage. On clicking *no sé*, participants were given another opportunity to answer the question, this time with an explanation from the traditional textbook (TRAD) or one generated by GPT. For certain terms (e.g., *flotante* in supplementary material), the textbook contains something akin to a lexicographic definition. These were reproduced without their surrounding context. For other terms, the textbook provides more extensive explanation occasionally accompanied by code snippets (e.g., *def* in supplementary material). The GPT explanations were generated with ChatGPT 3.5. The prompt used was: "¿Qué significa X en el contexto de Python/programación?" ('What does X mean in the context of Python/Programming?'). The default settings for the free version of ChatGPT were used. The assumption being that this type of simple natural language prompt is likely to be used by novice programmers and that, like many software users and indeed online dictionary users (Trap-Jensen, 2010), these users are unlikely to alter the default settings. The code snippets from both sources were reproduced on the survey platform as image files.

2.2.2 Follow-up Interviews

In the weeks after the procedure, follow up interviews were carried out with six of the novice programmer participants to gain further insight about the relative merits of using the traditional resource or the more direct ChatGPT method to resolve language doubts. In these interviews, students were presented with both ChatGPT explanations and textbook definitions for an unknown term and asked to reflect on which, if any, they found more useful and why.

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

3. Results and Discussion

Two participants originally assigned to the GPT group are excluded from analysis as they abandoned the quiz without answering any items. The relatively large mean number (6.3 out of 10) of terms participants marked as unknown, suggests that the multiple-choice items were sufficiently challenging to probe the research questions and, more generally, lends weight to the assertion that understanding such terms poses a significant barrier to learning to code.

The following analysis focuses on items reported as being unknown. The first research question was answered by fitting a binary logistic mixed model which predicts success or failure for an item answered with the support of an explanation in the two conditions. The second research question was answered by fitting a linear regression model which predicts the time needed to answer an item in each condition. In both cases, model selection was carried out using the lme4 R package (Bates et al., 2015) with the assistance of the buildmer R package (Voeten, 2021). The R code for the fitting of both models can be found in the supplementary material.

3.1 Success

The final selected model for success included parameter estimates for random intercepts for items and participants. A summary of the best model is given in Figure 1.

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation)
(p-values based on Wald z-scores) ['glmerMod']
 Family: binomial  ( logit )
Formula: Success ~ 1 + Exp.Source + (1 | Item) + (1 | Participant)
   Data: PythonData

     AIC      BIC   logLik deviance df.resid
   204.2    217.0    -98.1    196.2      178

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.8063 -0.5367  0.3095  0.4768  2.2519

Random effects:
 Groups      Name        Variance Std.Dev.
 Participant (Intercept) 2.220    1.490
 Item        (Intercept) 1.435    1.198
Number of obs: 182, groups:  Participant, 29; Item, 10

Fixed effects:
             Estimate Std. Error z value Pr(>|z|) Pr(>|t|)
(Intercept)     0.5693     0.6042  0.9423    0.346   0.3461
Exp.SourceGPT   1.6695     0.8121  2.0558    0.040   0.0398 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr)
Exp.SorcGPT -0.441
```

Fig. 1: A summary of the best logistic regression model

The parameter estimates for the GPT level are significantly higher (z = 2.0558, p = 0.0398 < 0.05) than those for the traditional textbook resource (TRAD) level. In Figure 2, these parameter estimates are displayed in terms of probabilities.

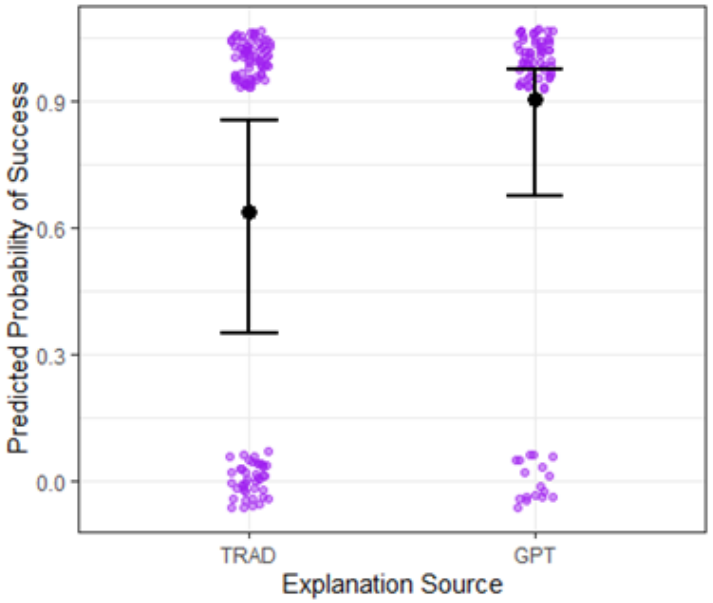

Fig. 2: Predicted probability of success for the two levels of explanation source, with 95% confidence intervals. Raw data points are shown jittered.

3.2 Time

Time measurements for every item were collected automatically using the paradata function of 1KA. Two clear cases of erroneous time measurements (416 & 2263 seconds)—possibly, due to participants leaving a question open and then returning to it later—are excluded from the analysis. The mean time per unknown item is markedly higher for GPT than TRAD (62.77 seconds vs. 35.22 seconds). Though the difference is exaggerated due to the skewness of the data. For further analysis, in line with accepted practice, we have transformed the data using logTime effectively compressing the scale of values, reducing the influence of extreme values thus making the distribution more symmetrical. A mixed regression model was fitted on logTime. The model parameters are given in Figure 3.

```
Linear mixed model fit by REML
(p-values based on Wald z-scores) ['lmerMod']
Formula: log_time ~ 1 + Exp.Source + (1 | Participant) + (1 + Exp.Source | Item)
   Data: PythonTime
REML criterion at convergence: 291.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.4062 -0.6472  0.0683  0.5910  2.3121

Random effects:
 Groups      Name          Variance Std.Dev. Corr
 Participant (Intercept)   0.23393  0.4837
 Item        (Intercept)   0.06360  0.2522
             Exp.SourceGPT 0.07428  0.2725   -0.44
 Residual                  0.18108  0.4255
Number of obs: 181, groups:  Participant, 29; Item, 10

Fixed effects:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.3153     0.1522   21.78  < 2e-16 ***
Exp.SourceGPT   0.6840     0.2151    3.18  0.00147 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr)
Exp.SorcGPT -0.604
```

Fig. 3: A summary of the best linear regression model

Figure 4 gives the predicted time per item in seconds for the two explanation sources.

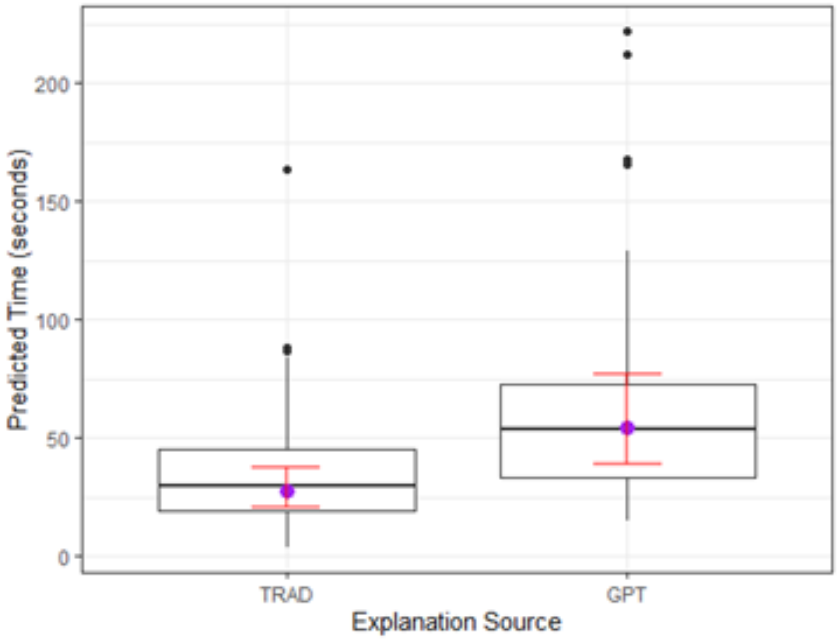

Fig. 4: Predicted time per item (in seconds) for the two levels of explanation source, with purple dots indicating mean and 95% confidence intervals are shown in red

3.3 Interviews

All six interviewees noted that the GPT definitions were more detailed. For certain items, they appreciated the extra information and examples. However, all but one interviewee reported feeling overwhelmed by the quantity of information provided at times (e.g., "ChatGPT gives a much larger and more complex explanation, which means that you forget a lot of information along the way"). Interviewees also expressed concerns about the appropriateness of the GPT explanations for novices (e.g., "ChatGPT offers a 250-word explanation, to put a number on it, of which 100 are incomprehensible. The examples don't help either 'the inverse of the square of 3', no idea!"). More generally there were concerns about "irrelevant", "repetitive", and "redundant" information. These were echoed by the experienced programmers consulted when writing the multiple-choice questions. This chimes with debates in pedagogical lexicography about 'good' dictionary definitions and restricted defining vocabularies.

In line with previous research (Becker, 2019; Guo, 2018), four of the five interviewees who addressed the issue agreed that they, as users whose L1 is not English, face an additional linguistic challenge when learning to program. Not only do they have to comprehend key programming terms, they must also face the challenge of understanding new concepts (e.g., "Knowing English helps you to remind yourself about concepts and understand the commands"). See supplementary material for a (Spanish) summary of the interviews.

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

4. Conclusions

This study set out to address a gap in the research on the use of chatbots in non-English, specific-purpose, lexicographic tasks by examining a case similar to one in which the user could feasibly skip a lexicographic resource and resolve their queries directly with the chatbot. The success scores reported above suggest that ChatGPT can be more useful than a traditional textbook for this group of L1 Spanish novice programmers in supporting comprehension of unknown programming terms. Taken in isolation, this suggests that users might skip the traditional reference resource altogether and resolve their linguistic doubt directly using ChatGPT. However, the significantly greater mean time needed to resolve queries using ChatGPT and participants' reservations about the relevance and quantity of information provided by ChatGPT suggest that there may be a role for more carefully curated reference resources in conjunction with chatbot queries.

Care should be taken not to overgeneralise these conclusions. They were based on the responses of a relatively small number of participants to questions designed to test receptive knowledge of only ten basic programming terms. It is also notable that while this Spanish study counters English dominance in AI research, there are a great number of languages which are even less researched. Furthermore, the data collection instrument underplays the real-world affordances of ChatGPT. For example, on stating that they did not know a term, users were immediately presented with explanations from one of the sources. This meant that neither the time needed to find an item in a traditional resource and other accessibility advantages (Rees, 2023), nor the possibility of asking follow-up questions in ChatGPT were accounted for. Finally, chatbot technology is constantly evolving, not only in terms of underlying models, but also in terms of closer integration with development environments and productivity software (e.g., BethanyJep, n.d.) in a similar way to previous integration of dictionaries with word processors via writing assistants (e.g., Frankenberg-Garcia et al., 2019). These limitations present interesting avenues for future research.

References

1KA. (2023). [Computer software]. Faculty of Social Sciences. https://www.1ka.si

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Becker, B. A. (2019). Parlez-vous Java? Bonjour La Monde != Hello World: Barriers to Programming Language Acquisition for Non-Native English Speakers. In M. Marasoiu, L. Church, & L. Marshall (Eds.), *Proceedings of the 30th Annual Workshop of the Psychology of Programming Interest Group, PPIG 2019, Newcastle University, UK, August 28—30, 2019*. Psychology of Programming Interest Group. https://ppig.org/papers/2019-ppig-30th-becker/

BethanyJep. (n.d.). *Using GitHub Copilot with Python—Training*. Retrieved 29 May, 2024, from https://learn.microsoft.com/en-us/training/modules/introduction-copilot-python/

de Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, *36*(4), 355–387. https://doi.org/10.1093/ijl/ecad021

Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P., & Sharma, N. (2019). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, *31*(1), 23–39. https://doi.org/10.1017/S0958344018000150

Guo, P. J. (2018). Non-Native English Speakers Learning Computer Programming: Barriers, Desires, and Design Opportunities. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3173970

Jakubíček, M., & Rundell, M. (2023). The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-editing Lexicography? In M. Medveď, M. Měchura, I. Kosem, J. Kallas, C. Tiberius, & M. Jakubíček (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference* (pp. 518–533). Lexical Computing CZ s.r.o.

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

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, *11*(1), 1–8. https://doi.org/10.1057/s41599-024-02889-7

Lorentzen, H., & Theilgaard, L. (2012). Online dictionaries – how do users find them and what do they do once they have? In R. V. Fjeld, & J. M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress* (pp. 654–660). Department of Linguistics and Scandinavian Studies, University of Oslo.

Marzal, A., & Gracia, I. (2020). *Introducción a la programación con Python*. Publicacions de la Universitat Jaume I.

Rees, G. P. (2023). Online Dictionaries and Accessibility for People with Visual Impairments. *International Journal of Lexicography*, *36*(2), 107–132. https://doi.org/10.1093/ijl/ecac021

Rees, G. P., & Lew, R. (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. https://doi.org/10.1093/ijl/ecad030

Trap-Jensen, L. (2010). One, Two, Many: Customization and User Profiles in Internet Dictionaries. In A. Dykstra, & T. Schoonheim (Eds.), *Proceedings of the 14th EURALEX International Congress* (pp. 1133–1143). Fryske Akademy.

Voeten, C. C. (2021). *Using 'buildmer' to automatically find & compare maximal (mixed) models*. https://cran.r-project.org/web/packages/buildmer/vignettes/buildmer.html

Supplementary Material

Supplementary material including the R code for the model selection and figures, the multiple-choice questions and explanations, the participant response data, and a Spanish summary of the interviews can be found here: https://osf.io/9rbdw/

Acknowledgements

The authors are grateful for the support of the Research Group in Language and Technology (ReLaTe) (2021 SGR 00151).

Contact information

Geraint Paul Rees
Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona, Spain
geraintpaul.rees@upf.edu

Isabel Gibert
Department of Romance Studies, Universitat Rovira i Virgili, Tarragona, Spain
isabel.gibert@urv.cat