

RECOMMENDED RESOURCE

## A freely-available system for browser-based Q&A practice in English, with speech recognition

Myles O'Brien  
Mie Prefectural College of Nursing, Japan

---

[myles.obrien@mcn.ac.jp](mailto:myles.obrien@mcn.ac.jp)

### Abstract

A browser-based system to facilitate practice in asking and answering simple questions in English was developed. The user may ask or answer by speaking or typing, and the computer's output is in the form of speech and / or text. The types of questions handled and the permitted vocabulary are limited, though the vocabulary items may be edited freely. The system was well received in a small pilot study among Japanese students. It is freely-available for download, and requires no technical expertise to deploy, just the facilities and ability to edit text files and upload to the internet.

**Keywords:** Automatic speech recognition, computer-assisted language learning, Google speech API, pattern practice.

### 1. Introduction

The system is named AARDVARK (Audibly Ask and Respond to Dynamic Vocabulary Addition and Removal Knowledgebase). It aims to provide a stimulating environment for learners of English to practice asking and answering simple questions. It has much in common with a system developed by the author in 1997 (O'Brien, 1997) but takes advantage of the huge developments in internet technology since those days. The old system was written in HyperCard, a very popular CALL tool at that time, so it could run only on Macintosh computers as a standalone program. It employed the very mechanical-sounding speech simulation available then, and had no speech recognition capability. The current system was written with HTML5 and JavaScript, so it will work in any modern browser. The speech recognition function uses Google's automatic speech recognition (ASR) facility, which works only in the Chrome browser, though not on iOS. So, the system can be fully used on Windows, Mac OS, Linux, and Android, and with all features except speech recognition on iOS. A previous paper (O'Brien, 2017) has given a brief history of the use of speech recognition in CALL, and other accounts are given elsewhere ( van Doremalen, Boves, Colpaert, Cucchiari, & Strik, 2016; Ashwell & Elam, 2017; Daniels & Iwago, 2017).

The system includes a sample set of vocabulary items, so it can function out of the box. The items may be edited freely by the user, and the system constructs the questions and answers it generates based purely on its supplied vocabulary and the rules of English grammar and syntax, as written into its algorithms. Its choices are random, without any

element of artificial intelligence. Therefore, a judicious choice of vocabulary is recommended to better simulate meaningful interaction. How the system fits into the overall scheme of CALL is considered in the Discussion section.

The system contains two complementary parts, one which answers the user's questions, and the other which asks questions and checks the user's answers. Since the question forms and vocabulary items which may be used are predefined, they are quite limited, though it was attempted to enable a good selection of basic Q&A types for pattern practice.

## 2. The system interfaces

First, the basic functioning of the system will be outlined through its interfaces. The system consists of two separate, complementary parts:

1. "q-answer" which answers questions put by the user
2. "q-ask" which poses questions to the user and checks the answer given

"q-answer" has 3 interfaces. Figure 1 shows the initial setup screen, which comes up at the start. This allows you to check the available vocabulary, and edit it if you wish. Clicking on "OK" takes you to the main interface.



Figure 1. Initial screen for "q-answer".

Figure 2 shows part of the main screen. You can type a question directly, or click the microphone icon for voice input. Clicking “Enter” calls up an answer. Clicking on the speaker icon allows you to hear the answer again. There is also a tools icon to adjust settings.

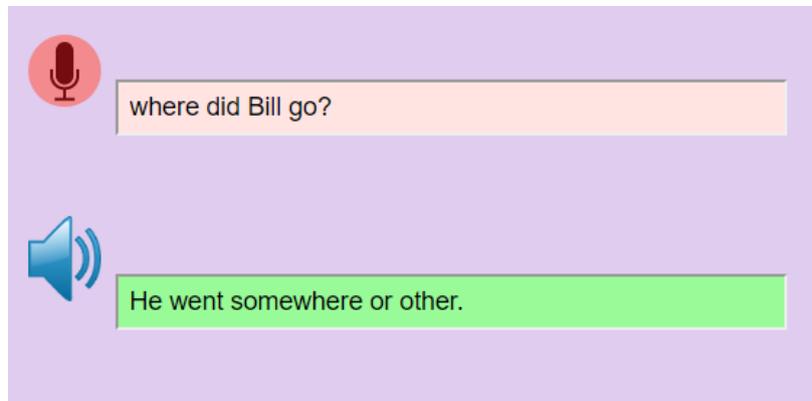


Figure 2. Main screen for “q-answer”.

Figure 3 shows the settings screen which appears when the tools icon is clicked. “Edit vocabulary” brings you back to the initial screen again. “Toggle sound” turns off/on the simulated speech reading of the answer, and “Toggle text” turns off/on the text display of the answer.

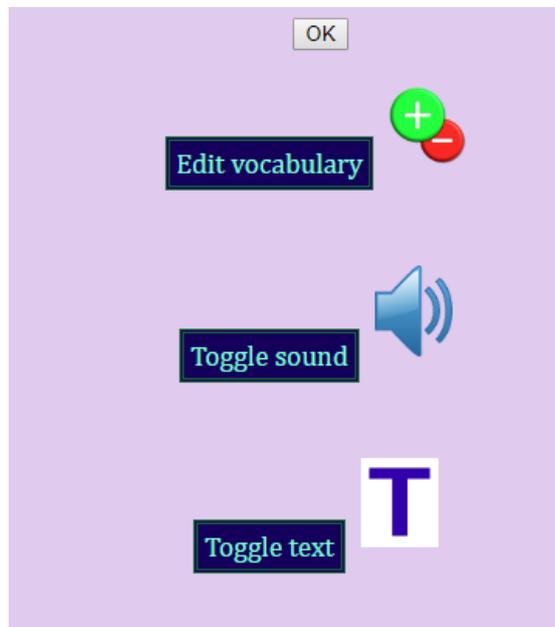


Figure 3. Settings screen for “q-answer”.

“q-ask” also has 3 interfaces. Figure 4 shows the initial setup screen, which comes up at the start. This allows you to check the available vocabulary and the checked question type options, and edit it if you wish. Clicking on both “OK” allows you to accept and proceed to the main interface.

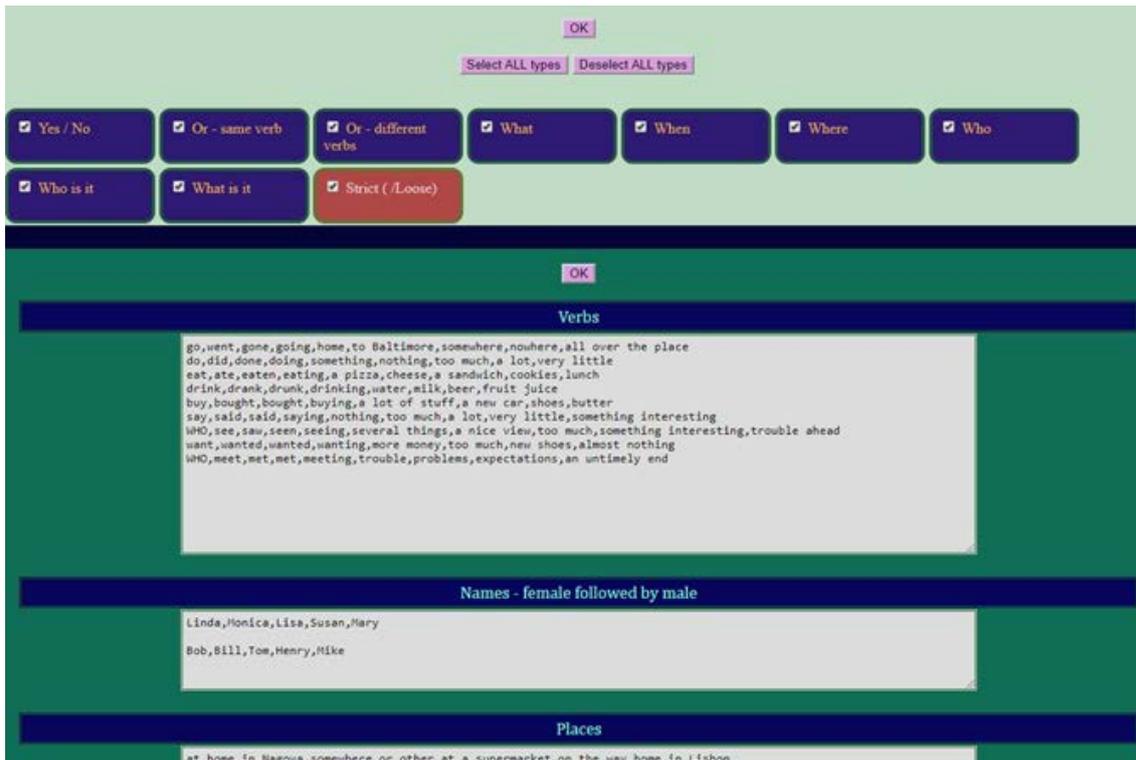


Figure 4. Initial screen for “q-ask”.

Figure 5 shows the main screen. Clicking on “New” generates a question. The microphone icon allows you to speak your answer, or else you can type it. The “Check” button allows you to check your answer. If incorrect, a correct example will be shown. Selecting “Cheat” displays the correct example before inputting your answer. By clicking on the speaker icon, you can hear the question again. Lastly, clicking on the tools icon allows you to adjust settings.

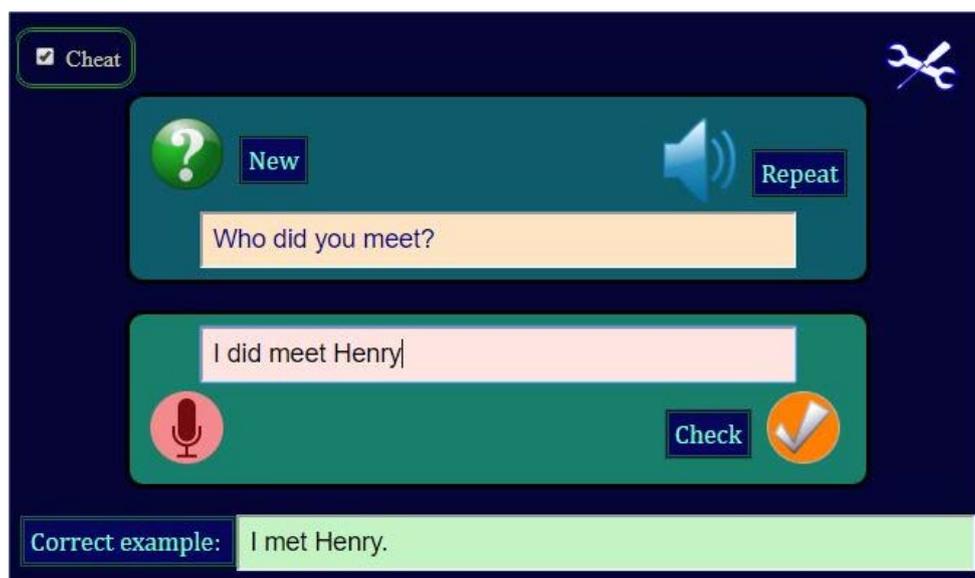


Figure 5. Main screen for “q-ask”.

Figure 6 shows the settings screen from clicking the tools icon. “Edit vocabulary” or “Edit question types” brings you to the setup screen again. “Toggle sound” turns off/on the simulated speech reading of the question, and “Toggle text” turns off/on the text display of the question.

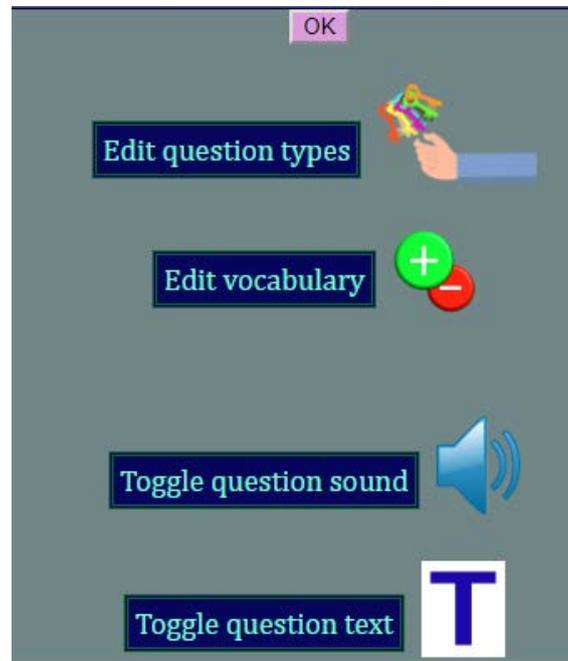


Figure 6. Settings screen for “q-ask”.

### 3. Scope of the system

#### 3.1. “q-answer”

The user’s question is first analyzed to see which pattern it fits, and the relevant algorithm is triggered to produce a suitable answer. Answers are chosen randomly from the possible alternatives. The types of questions handled, with some sample answers produced when using the vocabulary shown in Figure 1, are as follows:

#### ◆ *yes / no* questions

- Did Tom see Chris? *No, he didn’t.*
- Will you help me, please? *Yes, I will.*
- Should Linda buy a cat? *No, she shouldn’t.*
- Am I annoying you? *Yes, you are.*

Note that only the first two words, the auxiliary and subject, are relevant in determining the answer, so words outside the specified vocabulary, including nonsense words, may be used in the rest of the question.

#### ◆ *either / or* questions

- Did you drink tea or coffee? *I drank coffee.*

- Will Linda drink tea or eat chocolate? *She'll eat chocolate.*
- Should he eat a pizza or something healthy? *He should eat something healthy.*
- Is Mike going today or next week? *He's going today.*

Again, words outside the specified vocabulary may be used, but the verbs must be specified ones to elicit a proper response.

◆ **what questions**

- What were you eating? *I was eating lunch.*
- What did Bill buy? *He bought shoes.*
- What did Mary say? *She said very little.*
- What must Linda eat? *She must eat a sandwich.*

In this case, the algorithm will choose a random item from the list of objects shown in the initial screen.

◆ **when questions**

- When did Lisa eat the cake? *She ate it at one o'clock.*
- When will Mike go to France? *He'll go there after lunch.*
- When should I eat the grapes? *You should eat them in the morning.*
- When were you drinking coffee? *I was drinking it on Monday.*

In this case, the algorithm will choose a random time from the set-up list, so it is important to include only items which will give a sensible answer regardless of tense. E.g., “next week” might give rise to an answer like “She went there next week.”

◆ **where questions**

with “go”:

- Where did Mary go yesterday? *She went to Kyoto.*
- Where were you going? *I was going home.*

other:

- Where did Susan eat? *She ate somewhere or other.*
- Where can Mike buy the fish? *He can buy it in Lisbon.*

The “places” vocabulary list is divided into “go” and “do”, so that the vital basic verb “go” can be handled properly.

◆ **who questions**

Name as subject:

- Who did Tom see? *He saw Mary.*

Name as object:

- Who wanted to see Linda? *Mike wanted to see her.*

No name:

- Who was eating the cake? *Tom was eating it.*
- Who is (was) it that...:
- Who was it that ate the cake? *It was Monica.*

With the above types, it was hoped to facilitate practice of a wide variety of examples of question and answer forms in a stimulating way, since the learner asks their own questions, albeit within a limited framework.

### 3.2. “q-ask”

Although “q-answer” provides a type of interactive practice with question and answer, it is all one-way traffic with the learner passively receiving the answers and leaving the tricky grammatical transformations to the computer. This may be useful up to a point, but more complete practice of communication is desirable, so “q-ask” reverses the roles. It handles almost exactly the same range of types as “q-answer”, though the “name as object” and “no name” types of “who” questions have not yet been implemented, and a “What is it that...” (analogous to “Who is it that...”) type has been added.

The number of options on the initial screen is also much greater with “q-ask”. With “q-answer”, the user decides what questions to ask, so practice is automatically tailored to whatever structures it is desired to use. The default mode of “q-ask” is random, so that what question type will be asked each time is unpredictable. But the initial settings allow any subset of types to be selected, and the questions will be limited to those types. There is also a choice of “Strict” or “Loose” mode. The default Strict mode enforces practice with designated vocabulary items, whereas Loose mode allows greater freedom. With Strict mode, items from the listed vocabulary must be used for objects, names, places and times. Loose mode checks only the basic sentence structure and grammatical transformations. So, taking the question “What did Mike eat?” as an example, “He ate a sandwich.” will be acceptable in either mode if “a sandwich” is one of the objects listed after “eat”. If “a sandwich” is not listed, Loose mode will accept the answer, but Strict mode won’t. Both modes will reject “He eat a sandwich”. Loose mode will also accept the likes of “He ate GBtt%\$fqp skGgm” because the basic structure is correct, despite the nonsensical object. Thus, the learner is afforded flexible communicative practice in that, not alone are the vocabulary lists editable, but items outside the listed ones may be used.

The speech recognition capability is another aspect of the system which contributes to its communicative character. As in the system described in a previous paper (O’Brien, 2017), speech recognition was implemented using Google speech recognition. This is the only ASR engine that offers a web-based API, so it can be easily integrated into a browser-based system like the present one. Furthermore, in comparisons with other systems, it has demonstrated the most accurate speech recognition results (Ashwell & Elam, 2017; Daniels & Iwago, 2017; Këpuska & Bohouta, 2017).

#### 4. File structure of the system

“q-answer” and “q-ask” are separate systems, but almost identical in structure. They consist of a single HTML file, a folder containing images for the interface elements, and four text files containing vocabulary items, one file each for verbs, names, times, and places. Figure 7 depicts the structure diagrammatically.

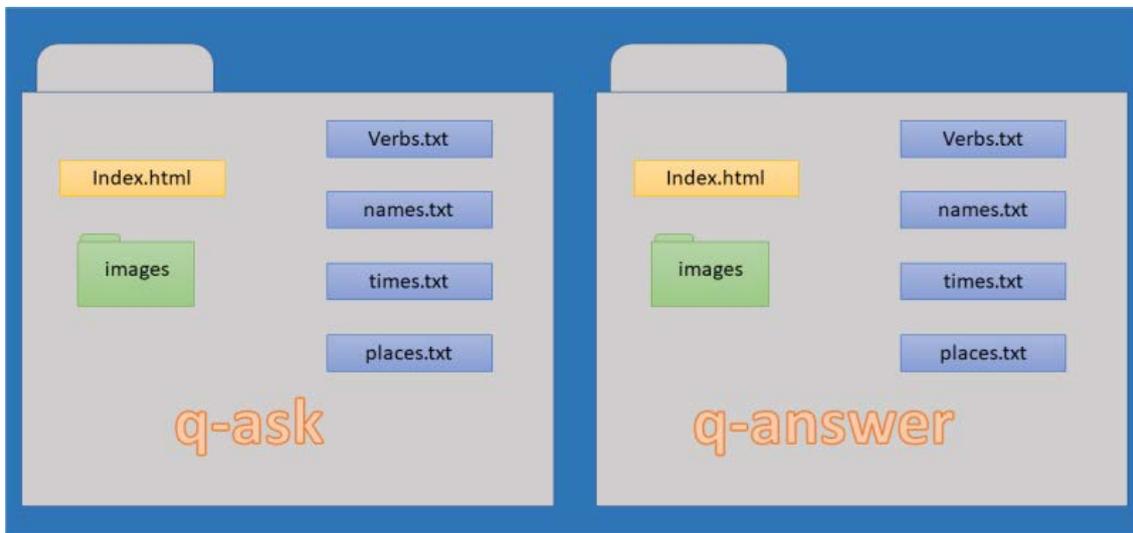


Figure 7. File structure of the entire system.

Both parts are freely available for download. Each one is in the form of a zip file, which expands into a folder as shown in Figure 7. If the folder is uploaded to a website, it should operate immediately without any need for setup or adjustment. In many cases, the uploader may want to edit the vocabulary items, rather than just using the default ones supplied. This is easy to do just by editing the comma-separated lists in the supplied plain text files. The user can also edit the items online while using the system, but the changes will not persist after the browser window is closed. The image files for the interface elements are in the “images” folder, and may be replaced by alternative images if the same filename is used.

An important point to note is that, for speech recognition to work, HTTPS must be enabled. HTTPS support is becoming more common as a free option on hosting services. If it is not available by default, an SSL certificate to enable it can be obtained from a Certificate Authority such as Let’s Encrypt, <https://letsencrypt.org>, which is a non-commercial organization offering a free, automated service.

#### 5. Pilot study

A small pilot study to assess user reaction was carried out with ten Japanese nursing students, of mixed English ability. The functioning of both parts of the system was explained and demonstrated to two students at a time, and then they were asked to try out both parts. First one student asked 5 or 6 questions to the “q-answer” system, using voice input, then the other student did the same. The same procedure was repeated once, before moving on to answering questions asked by the “q-ask” system in a similar, alternating way. The students sat beside each other and communicated freely during the process. The

reason for doing the trial in pairs was that it had been noticed before in informal testing with individual students that they found it quite mentally tiring to practice for more than a few questions (especially when answering rather than asking) while being observed by a teacher. Practicing in pairs was found to make the process less stressful and tiring, as it reduced the individual workload and lightened the atmosphere, facilitating mutual help and encouragement.

After the trial session, they were asked to complete a short web questionnaire in Japanese. In English translation, the survey questions asked which activity, asking or answering questions, they thought most likely to be helpful to their English study (“Can’t say” was a third alternative). Then, about the system in general, there were three 5-point scale items, “Is likely to help your study of English?”, “Is easy to use?”, “Is enjoyable to use?”, and three free input items, “Good points”, “Points needing improvement”, “Any other comments”. While the researcher observed their use of the system, providing assistance if needed, the questionnaires were completed anonymously in private, to avoid causing pressure.

The responses for the 5-point scale items were quite positive: an average of 3.9 for “Is easy to use”, 4.4 for “Is likely to help your study of English?”, and 4.8 for “Is enjoyable to use?”. In the free input section, positive points mentioned were, “good for pronunciation training” (most frequent), “makes one careful about singular/plural, articles & pronouns”, “conversation-like practice”, “good for enjoying study with a friend”, and “getting correct answer is enjoyable and builds confidence”. Among the negative points, by far the most common was the difficulty of getting one’s intended spoken words recognized correctly. Suggestions for improvement included having sound as well as text for correct examples, and a hint facility, e.g., an option to see the beginning of the correct answer.

The trial suggests that the system has the potential to be a useful tool for language study. There was a big difference between the students in how well the system correctly interpreted their intended utterances. Those who were less successful with the ASR expressed frustration, though they were very pleased if they eventually succeeded without resorting to correction by typing. Some were able to adjust their way of speaking to gain distinct improvement, to their great satisfaction. Some special features of Japanese speakers’ pronunciation of English were evident. For instance, names containing the “l” sound, “Bill” and “Linda”, caused difficulty with ASR, whereas “Monica” was recognized almost 100% of the time. “Would you...” caused great difficulty, whereas “Can you...” went much more smoothly.

## 6. Discussion

Even from the participants in the pilot study who had less success with ASR, the reaction was generally positive, so the system seems to have potential as a useful tool for the language teacher. It is being made available in the hope that it will be used in both classroom and self-study environments to the benefit of both teachers and learners.

Though in its essence, the system is a facilitator of pattern practice, it is hoped that it can be regarded as a highly enriched version of same, with elements of genuine communicative practice. Interestingly, its two separate parts can be seen as embodying two important concepts in Second Language Acquisition Theory: q-answer provides the user with *comprehensible input* (Krashen, 1981), and q-ask requires the user to

produce *comprehensible output* (Swain, 1985). While the learner's need for input is so obvious that nobody would question its central importance in SLA, the importance of producing output has not been universally accepted, with Krashen and others continuing to downplay its significance because of its strong element of conscious learning (Krashen, 1998; Ponniah & Krashen, 2008; Ponniah, 2010; Jarvis & Krashen, 2014). However, Swain has maintained her view, and been supported by others (Swain & Lapkin, 1995; Iwanaka & Takatsuka, 2007; Mennim, 2007; Donesch-Jezo, 2011; Berkner, 2016), so that her output hypothesis remains a seriously-considered concept among the many which have been proposed in SLA theory.

Although the system, with its emphasis on explicit practice focused on grammatical structures, may not appeal to supporters of the *natural approach* (Krashen & Terrell, 1983), it at least provides one of the recommended features of that approach, a stress-free learning environment. This is because it is designed with self-access in mind, though it is also suitable for classroom use. Also, there is much evidence for the usefulness of explicit practice in language learning (DeKeyser, 2010; Lyster & Sato, 2013; Jones, 2018).

The system could probably be classified as dialogue-based CALL, as defined by Bibauw, François & Desmet (2015), though its very limited scope makes it difficult to fit comfortably into any of their four categories. Indeed, many more sophisticated and capable systems have been developed (Xu & Seneff, 2009; van Doremalen, Boves, Colpaert, Cucchiari, & Strik, 2016; Bodnar, Cucchiari, Penning de Vries, Strik, & van Hout, 2017; Sydorenko, Smits, Evanini, & Ramanarayanan, 2018). However, a distinguishing point of the current system is that it is lightweight and can be freely obtained and deployed by anybody who has upload access to a website. They can use it as is, or edit the vocabulary items to their liking. It can then function as one resource among many, in what Gimeno-Sanz (2016) calls "atomized CALL" in contrast to Warschauer's (1996) "integrative CALL".

## **7. Download**

The two parts of the system can be tested online and downloaded as zip files from <https://www.mylesobrien.com/q-answer/> and <https://www.mylesobrien.com/q-ask/>. Each zip file includes the HTML file and images folder, which contains all the image files used in the user interface. These may be left unchanged, though users are free to make their own customizations by editing the HTML file or replacing any image files with their own. The vocabulary text files used in the online examples are also included. The teacher will need to edit these, or make new ones, to make persistent adjustments to the vocabulary. The author also gives permission for users to make modified versions of the system for educational use by editing the HTML or JavaScript, provided they do not claim the original or modified system as their own work.

## **References**

Ashwell, T. & Elam, J.R. (2017). How accurately can the Google Web Speech API recognize and transcribe Japanese L2 English learners' oral production? *JALT CALL Journal*, 13(1), 59-76. Retrieved from <https://journal.jaltcall.org/storage/articles/JALTCALL%2013-1-59.pdf>

Berkner, V. (2016). Revisiting Input and Output Hypotheses in Second Language Learning. *Asian Education Studies*, 1(1), 19-22. Retrieved from <https://pdfs.semanticscholar.org/c06d/ee67d68c61bbe79b57a6d8e1062a185a3bd3.pdf>

Bibauw, S., François, T., & Desmet, P. (2015). Dialogue-based CALL: an overview of existing research. In F. Helm, L. Bradley, M. Guarda, & S. Thouèsny (Eds), *Critical CALL – Proceedings of the 2015 EUROCALL Conference*, Padova, Italy (pp. 57-64). Dublin: Research-publishing.net. Retrieved from <http://dx.doi.org/10.14705/rpnet.2015.000310>

Bodnar, S., Cucchiarini, C., Penning de Vries, B., Strik, H. & van Hout, R. (2017). Learner affect in computerised L2 oral grammar practice with corrective feedback, *Computer Assisted Language Learning*, 30(3-4), 223-246.

Daniels, P., & Iwago, K. (2017). The suitability of cloud-based speech recognition engines for language learning. *JALT CALL Journal*, 13(3), 211–221. Retrieved from <https://journal.jaltcall.org/articles/220>

DeKeyser, R. (2010). Practice for Second Language Learning: Don't Throw out the Baby with the Bathwater. *International Journal of English Studies*, 10 (1), 2010, 155-165. Retrieved from <https://digitum.um.es/xmlui/bitstream/10201/23558/1/114021-452651-1-PB.pdf>

Donesch-Jezo, E. (2011). The role of output and feedback in second language acquisition – a classroom-based study of grammar acquisition by adult English language learners. *Journal of Estonian and Finno-Ugric Linguistics*, 2(2), 9-28. Retrieved from <http://ojs.utlib.ee/index.php/jeful/article/view/15323/10301>

Gimeno-Sanz, A. (2016). Moving a step further from “integrative CALL”. What's to come? *Computer Assisted Language Learning*, 29:6, 1102-1115.

Iwanaka, T. & Takatsuka, S. (2007). Roles of Output and Noticing in SLA: Does Exposure to Relevant Input Immediately After Output Promote Vocabulary Learning? *Annual Review of English Language Education in Japan*, 18, 121-130. Retrieved from [https://www.jstage.jst.go.jp/article/arele/18/0/18\\_KJ00007108492/pdf/char/en](https://www.jstage.jst.go.jp/article/arele/18/0/18_KJ00007108492/pdf/char/en)

Jarvis, H. & Krashen, S (2014). Is CALL Obsolete? Language Acquisition and Language Learning Revisited in a Digital Age. *TESL-EJ*, 17(4), 1-6. Retrieved from <http://tesl-ej.org/pdf/ej68/a1.pdf>

Jones, C. (Ed.) (2018). *Practice in Second Language Learning*. Cambridge: Cambridge University press.

Këpuska, V., Bohouta, G. (2017). Comparing Speech Recognition Systems (Microsoft API, Google API and CMU Sphinx). *Int. Journal of Engineering Research and Application*, 7(3), 20-24. Retrieved

from <https://pdfs.semanticscholar.org/2e7e/bdd353c1de9e47fdd1cf0fce61bd33d87103.pdf>

Krashen, S.D. (1982). *Principles and Practice in Second Language Acquisition*. London: Pergamon. Retrieved from [http://www.sdkrashen.com/content/books/principles\\_and\\_practice.pdf](http://www.sdkrashen.com/content/books/principles_and_practice.pdf)

Krashen, S.D. & Terrell, T.D. (1983). *The Natural Approach: Language Acquisition in the Classroom*. San Francisco: Alemany Press. Retrieved from [http://sdkrashen.com/content/books/the\\_natural\\_approach.pdf](http://sdkrashen.com/content/books/the_natural_approach.pdf)

Krashen, S. (1998). Comprehensible Output. *System*, 26, 175-182. Retrieved from [http://www.sdkrashen.com/content/articles/comprehensible\\_output.pdf](http://www.sdkrashen.com/content/articles/comprehensible_output.pdf)

Lyster, R., & Sato, M. (2013). Skill Acquisition Theory and the role of practice in L2 development. In M.G. Mayo, J. Gutierrez-Mangado, & M.M. Adrian (Eds.), *Contemporary Approaches to Second Language Acquisition* (pp. 71–92). Amsterdam: John Benjamins.

Mennim, P. (2007). Long-term effects of noticing on oral output. *Language Teaching Research*, 11(3), 265-280.

Ponniah, R. J. and Krashen, S. (2008). The Expanded Output Hypothesis.

*The International Journal of Foreign Language Teaching*, 4(2), 2-3. Retrieved from [https://www.academia.edu/738658/The\\_expanded\\_output\\_hypothesis](https://www.academia.edu/738658/The_expanded_output_hypothesis)

Ponniah, R. J. (2010). Insights into Second Language Acquisition Theory and Different Approaches to Language Teaching. *Journal on Educational Psychology*, 3(4), 14-18. Retrieved from <https://files.eric.ed.gov/fulltext/EJ1102366.pdf>

O'Brien, M. (1997). A computer program to provide practice in questions and answers for learners of English. *Computer Assisted Language Learning*, 10(3), 299-305.

O'Brien, M. (2017). A freely-available authoring system for browser-based CALL with speech recognition. *EUROCALL Review*, 25(1), 16-25. Retrieved from <https://polipapers.upv.es/index.php/eurocall/article/view/6830>

Swain, M. (1985). Communicative competence: some roles of comprehensible input and comprehensible output in its development. In S.M. Gass and C.G. Madden (Eds.) *Input in second language acquisition* (pp. 235–253). Rowley, MA: Newbury House.

Swain, M. & Lapkin, S. (1995). Problems in Output and the Cognitive Processes they Generate: A Step Towards Second Language Learning. *Applied Linguistics*, 16(3), 371-91. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.888.9175&rep=rep1&type=pdf>

Sydorenko, T., Smits, T., Evanini, K., & Ramanarayanan, K. (2018). Simulated speaking environments for language learning: insights from three cases. *Computer Assisted Language Learning*, 32(1-2), 17-48.

van Doremalen, J., Boves, L., Colpaert, J., Cucchiarini, C., & Strik, H. (2016). Evaluating automatic speech recognition-based language learning systems: A case study. *Computer Assisted Language Learning*, 29(4), 833-851.

Warschauer, M. (1996). Computer Assisted Language Learning: An Introduction. In S. Fotos (Ed.), *Multimedia language teaching* (pp. 3-20). Tokyo: Logos International. Retrieved from <http://www.ict4lt.org/en/warschauer.htm>

Xu, Y. & Seneff, S. (2009). Speech-Based Interactive Games for Language Learning: Reading, Translation, and Question-Answering. *Computational Linguistics and Chinese Language Processing*, 14(2), 133-160. Retrieved from <https://www.semanticscholar.org/paper/Speech-Based-Interactive-Games-for-Language-and-Xu-Seneff/29c0398871d6b34e9a220e134f90f5cec6dfb86a>

---