An Evaluation of Chatbots as Software Aids to Learning English as a Second Language

Abstract

'Chatbot' programs are pieces of software that can hold a conversation, or interact, in English. This paper explores the potential of chatbots for ESL (English as a Second Language) learning from a pedagogical perspective. From the command-line days of Eliza, chatbots have matured considerably – to the point where many chatbots now involve an avatar interface, with speech recognition also becoming available as a feature. The paper evaluates six chatbots currently available either online or for purchase. The evaluation examines chatbots from the perspective of their interfaces as a human-looking or sounding partner to chat with, as well as their usability as pieces of software suitable for ESL learners. To put some of these issues in perspective and provide insights into their use, the paper also reports on the use of some chatbots in the ESL classroom. The paper concludes with an analysis of chatbots currently available, noting that while chatbots have matured considerably since the early days of Eliza, they still have a long way to go before they can interact with students in the way that researchers such as Atwell (1999) envisage.

Keywords: chatbots, ESL, software analysis, CALL

1. Introduction

This paper is the second of two which examine chatbot programs – computer programs which attempt to maintain a conversation with a human. The previous paper reviewed five chatbots (the first five in Table 1 below), exploring how they handled and produced language from an essentially Hallidayan systemic functional grammatical perspective of performance at the level of word, clause and text [Note 1]. Essentially, chatbots perform best when the language input consists of one-clause sentences, the proposition is straightforward, and the topic is an everyday one (A summary is presented in Table 2
below before the details of the current study are presented). The previous paper also
provided considerable background to chatbots in terms of their history and
development. This will not be restated here. The reader is referred to Coniam
(forthcoming) for background historical information.

The current paper moves forward from a linguistic focus to examine chatbots from the
perspective of their features, functionality and usability in pedagogical terms.

2. Background

If we are to view chatbots as CALL software, a useful starting point is Salaberry's
(1996) discussion of what constitutes a “minimal theoretical framework” with regard to
effective CALL applications. Salaberry posits that CALL’s effectiveness as a medium in
language learning depends on three elements: the role that the message plays in the
medium within the language-learning situation; the content of the materials; and the
ways medium and content interact within the whole language-learning experience.

It is also worthwhile considering Warschauer and Healey’s (1998) delineation of the four
'stages' of CALL since this helps to further situate chatbots within the larger CALL
framework. In their overview of the development of CALL, Warschauer and Healey
(1998) describe four CALL 'stages' comprising: 'Behaviouristic' CALL, 1960s and 70s;
'Communicative' CALL, 1970s and 80s; 'Integrative' CALL, mid 1990s onwards; and
'Intelligent' CALL, 2000s onwards. The rationale for the 'Intelligent' CALL stage
(Warschauer and Healey 1998: 65-66) – where the current paper positions itself – is
that the intelligent power of the computer should maximise interaction with the material
to be learned, with 'intelligent' programs providing meaningful feedback and guidance to
students, and with comprehensible information presented through multiple mediums to
fit the learning style of individual students.

A useful extension to Warschauer and Healey's Intelligent' CALL stage is found in
Atwell's (1999) survey of Speech and Language Technology applications. Atwell
investigates different types of language engineering software from the perspective of a
'language machine'. He also investigates how this will impact on ESL teaching over the
next two decades with regard to applications such as sentence and text-level analysis
and understanding; voice recognition and text-to-speech conversion; and human-
machine interaction (1999: 8).

3. Chatbots in foreign language teaching and learning

The value of chatbots for language learning is clearly apparent: they have the potential
to provide a convenient chatting environment for learners to conduct authentic
conversations in the target language either by writing or by speaking. Further, the chat
transcripts (as in Figure 2, for example) provided by some programs can serve as a
source of the target language.

Various researchers have commented on the potential of chatbot as a conversation
simulator conducting conversations with users on various topics. Parker (2005: 49), for
example, discusses the use of chatbots with young English language learners. She
states that while chatbots’ ability to converse is still limited, they have potential for
second language learning support – although little use appears to have been made of
chatbots for this purpose thus far (2005: 49).

Although chatbots have not been totally ignored in the language teaching literature,
there is very little literature on their use and potential. In their innovative book on
conducting CALL activities with (at the time) generally-available software, Hardisty and
Windeatt (1989: 75), for instance, suggest an activity simulating a job interview with
an Eliza -type program. The Monthly ESL Newsletter of the ESL ALL!
International website contained a feature in its November 2005 newsletter on
conversing with a 'robot' as a way for learners to improve their writing or listening skills
(http://www.eslall.com/newsletter/Newsletter_nov_05.html). There is no evaluation of
the bots listed, however. The newsletter simply describes the features of 10 online bot
sites.

In a study with undergraduate ESL teachers in training, Coniam (2004) describes an
evaluation of the ALICE Artificial Intelligence Foundation site’s Dave (also evaluated in
the current study). It is claimed that Dave is the “perfect private tutor”, since he replies
“in perfect English just like a private English teacher” (http://www.alicebot.org/dave.html). Teachers’ attempts at conversing with Dave indicated that, conversationally, Dave showed considerable development from the restricted pattern matching employed by Eliza. However, while many of Dave’s conversational strategies had quite a natural feel, there were syntactic infelicities and conversational glitches which indicated that the program is unlikely to pass the Turing Test in the near future (Coniam, 2004: 160).

The developers of the Lucy chatbot (http://www.speak2me.net) – quite rightly recognising the potential of a speaking partner – have recently adapted elements of the online Lucy chatbot into a standalone piece of software, Lucy’s World: Smallt@lk (Ladder Digital Education Corp., 2005). This utilises the speech and interactive elements of the Lucy chatbot, but attempts to control the situation by restricting the topics and situations that Lucy is able to converse about.

In areas other than ESL, Chatterjee and Meijer (2005) describe the DeMeCoT Chatbot, a conversation trainer, for learners of Dutch. As a supplement to classroom speaking activities, this chatbot provides students with the opportunity to conduct dialogues – to which they receive response and feedback. However, Chatterjee and Meijer observe that trials and assessments of a range of chatbots demonstrated the limitations of such chatbots as language training instruments. They note that interaction is liable to break down, in part because chatbots lack the necessary knowledge for topic-based conversations; as well as being incapable of eliciting specific language use; and with changes in register usually being somewhat inconsistent. DeMeCoT, as a teaching vehicle, is therefore deliberately limited in its function – which is to enable students to practice certain linguistic structures and forms in simulated and limited conversations.

4. The study – chatbots reviewed

In this paper, six of the most usable chatbots from the perspective of potential ESL pieces of software are reviewed. These were selected on the basis of being strong either in terms of their ability to interact from a language perspective, or having an attractive interface – preferably both. In terms of the interface, it should be noted that chatbots have become considerably more sophisticated since the early days of Eliza, which operated from the command line with text output only. Technology has now matured to the point where, while many chatbots still operate via text input and a text-line response, many also respond in both text and synthesised, quasi-human, speech. Chatbots which will accept voice recognition input from a user are also beginning to appear.

Of the chatbots analysed in this paper, a number are drawn from two of the stables which are producing a number of today’s commonly-used chatbots – the ALICE Artificial Intelligence Foundation (http://www.alicebot.org) and Pandorabots, a robot-hosting service related to the ALICE community (http://www.pandorabots.com). After examining many chatbots – some online, some needing to be downloaded and installed; some with visual interfaces; and some to be run from the command line – six programs were settled upon. Five of these are available over the Internet; the sixth (the Ultra Hal Assistant) needs to be installed on a standalone computer. It is included in the current study since it innovatively incorporates voice recognition into the chatbot. Table 1 lists the six chatbots, and presents the criteria for inclusion.
<table>
<thead>
<tr>
<th>Program</th>
<th>URL</th>
<th>Rationale for inclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cybelle</td>
<td><a href="http://www.agentland.com">http://www.agentland.com</a></td>
<td>3D interface</td>
</tr>
<tr>
<td>Dave</td>
<td><a href="http://www.alicebot.org/dave.html">http://www.alicebot.org/dave.html</a></td>
<td>ALICE site – recommended for ESL learners</td>
</tr>
<tr>
<td>George</td>
<td><a href="http://www.jabberwacky.com">http://www.jabberwacky.com</a></td>
<td>Winner of 2005 Loebner prize</td>
</tr>
<tr>
<td>Jenny</td>
<td><a href="http://juan.vhost.pandorabots.com/pandora/talk?botid=a649e4ed5e34e5b3">http://juan.vhost.pandorabots.com/pandora/talk?botid=a649e4ed5e34e5b3</a></td>
<td>Pandorabots – large chatbot development site</td>
</tr>
<tr>
<td>Lucy</td>
<td><a href="http://www1.speak2me.net/lucy/src/log_freetalk.htm">http://www1.speak2me.net/lucy/src/log_freetalk.htm</a></td>
<td>Developed by language learning site</td>
</tr>
<tr>
<td>Ultra Hal Assistant</td>
<td><a href="http://www.zabaware.com">http://www.zabaware.com</a></td>
<td>Incorporates voice recognition</td>
</tr>
</tbody>
</table>

Table 1. Chatbots evaluated in the study.

Table 2 below (reproduced from the previous paper) presents a summary of the language issues analysed in the first five chatbots (the Ultra Hal Assistant was not included in the previous study because its speech recognition is dependent upon third party software such as Dragon NaturallySpeaking (Nuance, 2006) or Microsoft’s Speech Application Programming Interface Speech SDK 5.1 (http://www.microsoft.com) being installed –see below), with the different criteria rated on a five-point scale (five stars indicating good). Some ratings are supplemented by a comment.
<table>
<thead>
<tr>
<th>Level</th>
<th>Features</th>
<th>Cybelle</th>
<th>Dave</th>
<th>George</th>
<th>Jenny</th>
<th>Lucy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Word</strong></td>
<td><strong>Vocabulary range</strong></td>
<td>****</td>
<td>***</td>
<td>****</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td></td>
<td><strong>Wide range – general, technical</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Not very good with jargon, technical words</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Wide, even technical vocabulary</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Apparently wide, replies often out of context</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Quite wide – reasonable general vocabulary</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Handling incorrectly-spelt words</strong></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>*</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td><strong>Handling SMS forms</strong></td>
<td>****</td>
<td>****</td>
<td>****</td>
<td>**</td>
<td>***</td>
</tr>
<tr>
<td><strong>Sentence</strong></td>
<td><strong>Handling correctly-phrased questions</strong></td>
<td>*****</td>
<td>*****</td>
<td>*****</td>
<td>*****</td>
<td>*****</td>
</tr>
<tr>
<td></td>
<td><strong>Handling incorrectly-phrased questions</strong></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>*</td>
<td>***</td>
</tr>
<tr>
<td></td>
<td><strong>Handling incorrect parts of speech</strong></td>
<td>**</td>
<td>**</td>
<td>****</td>
<td>**</td>
<td>****</td>
</tr>
<tr>
<td><strong>Text</strong></td>
<td><strong>Logicality of answers</strong></td>
<td>**</td>
<td>***</td>
<td>****</td>
<td>**</td>
<td>***</td>
</tr>
<tr>
<td></td>
<td><strong>Number of sentences handled at a time</strong></td>
<td>*</td>
<td>*</td>
<td>**</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td><strong>Only last sentence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Generally only last sentence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Only last sentence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Only last sentence</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>1 exchange only</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Sometimes more than 1 exchange</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Could range over more than 1 exchange</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>1 exchange only</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Sometimes range over more than 1 exchange</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Responding in appropriate register</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Handling overt topic switch</strong></td>
<td>**</td>
<td>**</td>
<td>*****</td>
<td>**</td>
<td>***</td>
</tr>
<tr>
<td></td>
<td><strong>Handling unannounced topic switch</strong></td>
<td>*</td>
<td>*</td>
<td>**</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td><strong>Range of topics</strong></td>
<td><strong>Limited</strong></td>
<td><strong>Acceptable with general topics – movies, animals</strong></td>
<td>**</td>
<td>**</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td><strong>Acceptable with everyday topics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Acceptable with everyday topics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Limited mainly to US topics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Limited mainly to US topics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Quite broad general topics – family etc</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Summary of the programs' language capabilities.
The review in the previous paper concluded –as can be seen from Table 2– that chatbots had quite wide vocabularies, at least in terms of everyday topics. When input to them was in the form of correctly-phrased sentences, they coped reasonably well. Chatbots generally required correctly-spelt words in grammatically-correct sentences to function well. Some chatbots were able to deal with misspellings, but this was often unpredictable. The majority were really only able to handle input in the form of one proposition at a time, and preferably as a single-clause sentence. The range of topics they were able to respond to varied, although in general they performed much better with common topics. In summary, then, the previous paper concluded that chatbots coped best when the language input to them consisted of one-clause sentences, with a straightforward proposition, and an ‘everyday’ topic. With regard to the different linguistic abilities of the different chatbots, the previous paper concluded that George was by far the most able, although he could be slightly combatative both in his challenges and his efforts to topic switch when unable to provide a direct answer. Lucy and Dave coped acceptably at the single-sentence level, even though their replies were at times far from accurate. Lucy’s value, however, increased because she was sometimes able to suggest how to rephrase an incorrectly-framed sentence –an issue that will be revisited in the current paper. Jenny and Cybelle were the most limited in terms of their dealing with language.

5. Study objectives

The current paper evaluates chatbots in terms of their appearance and functionality as pieces of software. The functionality/usability perspective draws on Hubbard’s (1992) thorough framework for analysis of CALL courseware, and examines issues such as:

Screen design and layout
The quality of the sounds and graphics
The appearance and attractiveness of the chatbot (for example, is there an avatar?)
The program’s speed of response
Ease of use of the program – instructions, help etc.
Features offered by the program:
Is it possible to save conversations?
Can the chatbot be taught?
Does the chatbot remember a user from previous visits?
How much of the conversation is presented?
Does the chatbot offer different emotion?

As detailed in the previous paper, evaluators were trainee ESL teachers (all of whom rated at IELTS level 7 or above [Note 2]) in the final year of a four-year B.Ed. programme. They were reviewing and exploring the potential of chatbots as part of a 39-hour (three-unit) CALL course on which they were enrolled in their ESL teacher education programme. Each chatbot was evaluated by at least one evaluator, who, on average, spent between 8 and 10 hours working at their particular chatbot. Subsequent to the analysis, some of the evaluators taught ESL lessons where they encouraged students to work with the chatbots they had been evaluating. All analysis was conducted using computers generally less than one year old running Windows XP and via broadband Internet connections.

6. Results

This section now examines the different programs with regard to their appearance and operation as pieces of software, and how they handled language input from users.

6.1 The interface – screen design and layout

As mentioned above, many chatbots have advanced from a purely textual interface to an interface which incorporates an audio and/or visual interface, incorporating an avatar –a virtual, animated personality (see Lucy, Jenny below). Such developments not only
make for a more user-friendly experience, they also increase the programs’ potential in terms of language learning.

Figure 1 below provides screen captures of the six programs. As can be seen, the *Lucy* and *Dave* screens are devoted solely to the purpose of chatting, as is the *Ultra Hal Assistant*. These three are visually more appealing than the other three in that the screen is not cluttered with extraneous objects.

Although *George* has various options—including emotion-setting and a help function—the screen is cluttered by many buttons, links and adverts. The *Cybelle* website is also very cluttered with numerous adverts and other links, all of which distract from its appearance.

While *Jenny* is also somewhat cluttered in that there are links to various other websites, the screen is somewhat clearer than that of *Cybelle*. The large button to *Jenny*’s right is a link to the Merriam-Webster online dictionary—a useful function for an ESL learner.

As can be seen from Figure 1, *Lucy*, *Dave*, *Jenny* and the *Ultra Hal Assistant* have more attractive interfaces. These incorporate avatars which smile, move their heads to follow cursor movement, and move their lips in synch with the spoken output produced. *Lucy*’s smiling head and shoulders take up the left third of the screen, with the right-hand side taken up by input and output windows. *Dave*’s trendy-looking head and shoulders occupy the top quarter of the screen with the bottom half containing the ensuing conversation. *Jenny* is somewhat smaller but positioned clearly in the middle of the screen.
Cybelle, although an interesting-looking three-dimensional avatar, was squashed into a small column on the left-hand side of the screen. Unless users know what they are looking for, Cybelle could easily be overlooked.

George does not have an avatar, although a three-dimensional version is in preparation. The Ultra Hal Assistant default avatar is an unattractive lizard or robot. This can be customised, however, and the more acceptable human-looking Erica invoked instead (Figure 1 above).

6.1.1 Audio quality
The default for George is the text line. It is possible to link up Shockwave, but the speech quality thus generated is mechanical-sounding, and does not sound very human-like. Jenny’s speech quality is also artificial. Cybelle has no spoken output at all. The output produced by the Ultra Hal Assistant is somewhat staccato in its delivery.

The best speech quality comes from Dave and Lucy, who invoke quite natural-sounding spoken English generators, incorporating acceptable stress, rhythm and intonation. With Lucy, Dave, Jenny and the Ultra Hal Assistant, the avatar’s lips also move in synchrony with the spoken output, although Dave’s at times appear to be slightly out of synchrony. Nonetheless, this is a feature which augments the chatbots’ realistic image.

6.1.2 Conversation presentation
Lucy has a large window for the user’s input and output – both in a large easily-readable font. Although only the most recent turn is visible onscreen, the ‘Transcript’ button allows users to see the conversation thus far, via a scroll bar. The ability to view – and save – the conversation in this way makes for a useful pedagogical feature. Figure 2 presents a sample.

![Figure 2. Transcript from Lucy.](image)

The bottom half of the screen below Dave’s face shows the last eight lines of the conversation. Beyond these eight lines, however, nothing earlier can be viewed, nor can anything be saved.

The dialogue in George scrolls upwards, as in a chatroom, and therefore does not appear exactly as a conversation might be expected to in an English language teaching situation. There is, however, a link to view the conversation so far.

Jenny, Cybelle and the Ultra Hal Assistant show the most recent response only. Nothing prior to this can be accessed or saved.

6.2 Speed
In terms of response times, George has the fastest reactions, with almost instantaneous replies. Dave’s responses are generally quite acceptable. Cybelle, Jenny and Lucy can
be somewhat slow at times, sometimes leaving the user to wait for 10 seconds or more before a response is generated.

The *Ultra Hal Assistant* is naturally fast as it is installed as a standalone program.

### 6.3 Reviewing, Saving, E-mailing

The conversation in both *Lucy* and *George* can be e-mailed, a useful feature. It is not possible to save or e-mail the conversation in any of the other chatbots, however.

### 6.4 Help / Instructions

Generally speaking, where the interface in uncluttered (as with *Lucy* and *Dave*, for example) program operation is quite intuitive. *George* is the only program with a help feature on how to get the program to operate, or to get the best out of the program in terms of how it handles language and how it may respond. This is only in English, however, and for lower-level English language learners might not be easily understood.

### 6.5 Other characteristics

*George* has 'reaction' and 'emotion' buttons which supposedly increase the 'emotional intelligence' of the conversation. In practice, however, these buttons seem to contribute little to the quality of the conversation. They are an attempt to begin to deal with register as a linguistic feature, although they are not managing very well as yet.

*George* and *Lucy* have features whereby the chatbots can be 'taught'. This obviously permits the chatbots' language databases to be enlarged.

In terms of innovation, the *Ultra Hal Assistant* scores highly because it incorporates voice recognition as input, and for this reason alone it has been included in the current study. The program must, however, first be trained to a single user's speech –using either commercially-available software such as *Dragon NaturallySpeaking* (Nuance, 2006) or Microsoft's Speech Application Programming Interface *Speech SDK 5.1* ([http://www.microsoft.com](http://www.microsoft.com)). The quality of recognised speech is, however, not particularly accurate. Since it is voice recognition accuracy is generally lower with second-language speakers of English than with first-language speakers (Coniam, 1998), the *Ultra Hal Assistant* is some way off being usable as an ESL tool with second language learners of English. Consequently, while the program has been included in the software evaluation review in the current paper, its linguistic performance was not analysed in the previous paper.

### 6.6 Cost

All the online chatbots reviewed are free with the exception of *Dave*, which costs US$9.99 as a one-off subscription. A paid version of *George* (US$30 per year) is available, which apparently permits the user to teach George, to adapt George to "your personality - your character" and to incorporate a version of that chatbot on one's own website ([http://www.jabberwacky.com/yourbot](http://www.jabberwacky.com/yourbot)). The basic cost of the Zabaware's *Ultra Hal Assistant* is US$29.99.

A summary of the issues discussed above is now presented in Table 3 below.
Features | Cybelle | Dave | George | Jenny | Lucy | Ultra Hal Assistant
--- | --- | --- | --- | --- | --- | ---
Screen design | Cluttered, Cybelle in one small part of screen | Simple, face only | Somewhat cluttered; many links | Cluttered screen | Attractive; large face of Lucy with chat box | Simple, uncluttered
Appearance of avatar | Small, but 3D; turns body, waves; personal details file | Trendy-looking, moves head, lip synch slightly irregular | None; 3-D avatar in preparation | Small, eyeball movement; lip synch | Moves head, smiles, eyes follow cursor; lip synch good | Large attractive avatar, moves head and earrings
Sound quality | None | Good sound quality | Poor – needs Shockwave | Speech somewhat artificial | Good sound quality | Somewhat staccato
View conversation | Last utterance only | Conversation scrolls down screen – last 8 lines only | Conversation scrolls up screen; last 8 lines only | Current exchange only | Whole conversation in back screen | Current exchange only
Save conversation | No | No | Can save | No | Can email | No
Speed | Slow | Quite fast | Fast | Slow | Can be slow | Fast – standalone installation
Help / instructions | None | None | Instructions, but complex | None, but link to online dictionary | None | None
Correcting chatbot | No | No | Yes | No | Yes | No
Speech Recognition | No | No | No | No | No | Yes

Table 3. Summary of the features of different programs.

In summary, *Lucy* has the most attractive interface in that the screen is uncluttered and the avatar is large and has an attractive, friendly appearance. *Lucy* also responds in very acceptable spoken English, with good lip-synching. It is also possible to view and save (i.e., e-mail) conversations. One slight problem with using *Lucy* is that response time from a user’s input to her answering can occasionally be slow. Nonetheless, it is free. From the perspective of a CALL piece of software, therefore, *Lucy* emerges as the program with most potential from an ESL learning perspective.

### 7. Chatbots' pedagogical potential

Thus far we have examined chatbots’ appearance, functionality and ability with language. I would now briefly like to consider issues from a pedagogical perspective, and discuss their possible use in the ESL classroom.

Subsequent to their evaluations, as part of teaching practice, the student-teacher evaluators taught ESL lessons, where they encouraged their ESL students to interact with the chatbots they had been evaluating. The lessons were not open-ended, but generally had a language focus.

I will now briefly report on the classroom activities of some of the trainee teachers in the current study, who, having participated in the evaluation of the chatbots, used them in their teaching for specific language activities.

In one class, students had to use various wh- questions to get as much personal information from the chatbot as they could in order to write a description of the chatbot (a sample appeared in Figure 2). In another, the teaching point was the modal verb “should”. In this lesson, after first working out questions centring around “should” that
they might ask, students had to try out their questions on the chatbot. Finally, they had to report back to the class the questions they had put to the chatbot and how it had answered them.

I will report in more depth on one teacher who was teaching reported speech. This teacher asked the students to spend time talking to George and to find out things about him. Students had to first note the questions they put and the answers they received and then write these up as a standard language practice drill. Table 4 elaborates the work produced by one student in the class.

<table>
<thead>
<tr>
<th>Exchange</th>
<th>Human</th>
<th>George</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Conversation: How are you today?</td>
<td>I'm OK, thanks, just a bit bored.</td>
</tr>
<tr>
<td></td>
<td>Reported as: I asked her how she was that day.</td>
<td>She told me she was OK, just a bit bored.</td>
</tr>
<tr>
<td>2.</td>
<td>Conversation: Why are you bored?</td>
<td>I'm at work.</td>
</tr>
<tr>
<td></td>
<td>Reported as: I asked why she was bored.</td>
<td>She told me she was at work.</td>
</tr>
<tr>
<td>3.</td>
<td>Conversation: Where do you work?</td>
<td>I work at home.</td>
</tr>
<tr>
<td></td>
<td>Reported as: I asked where she worked.</td>
<td>She told me she worked at home.</td>
</tr>
<tr>
<td></td>
<td>Reported as: I asked her where she lived.</td>
<td>She told me she lived in Loughborough.</td>
</tr>
<tr>
<td>5.</td>
<td>Conversation: Are you the computer?</td>
<td>No. I'm a woman.</td>
</tr>
<tr>
<td></td>
<td>Reported as: I asked her that is she the computer.</td>
<td>She told me she was a woman.</td>
</tr>
</tbody>
</table>

Table 4. Work produced by one (less able) student.

For homework (to be subsequently handed in to the teacher), students had to write up their 'interview' in the form of a short article. One advantage was that different students' language levels could be taken account of. More able students were able to ask more complex questions, while less able students (as in Table 4 above, for example) asked somewhat simpler questions. All students were actively engaged in the task despite certain problems with George (see below) and despite the fact that they were not interviewing real people. The majority also submitted some interesting articles as homework.

The teacher polled students of this lesson with a short questionnaire in order to gauge their reactions to the chatbot and to chatting with it. Students' responses to the lesson were mixed: while some responded that they had enjoyed the lesson (as with the student whose work is shown in Table 4 above), others (the more able ones) were less positive because they could not always follow George's conversational threads, and (according to their responses) "wasted time on arguing" with the chatbot when he responded unpredictably. Most students also commented rather unenthusiastically on George's rather dull interface, with its lack of an avatar, and no spoken output.

8. Conclusion

This paper has described a study of certain chatbots currently available over the Internet. The study has involved an evaluation of those chatbots in terms of their appearance and functionality as programs; their ability in terms of handling language
and taking part in a conversation; and has provided a snapshot of their potential in the ESL classroom.

Since the early days of Weizenbaum's *Eliza*, chatbots' appearances have developed and improved substantially. Many now have an almost human-looking avatar as their interface with their interlocutor. The best of these avatars have an attractive appearance, move their head in response to mouse movement, blink their eyes as would a human, and move their lips in synch with the speech they produce. *Lucy* and *Dave*, although only two-dimensional, are the most attractive in terms of appearance. *George* does not have an avatar, which in terms of use with ESL learners (as learners using him commented) is a serious disadvantage.

Many of the chatbots have also incorporated spoken output. *Lucy*'s speech is of quite high quality, as is *Dave*'s. While still computer generated, it is approaching the qualities of human speech with good stress, rhythm and intonation patterns. *George*'s speech output, via the program Shockwave, is not as good, sounding much more mechanical than that of *Lucy* and *Dave*.

Screen design also merits mention. *Lucy*'s and *Dave*'s screens are meant for ESL learners and are uncluttered. *George*, and *Cybelle* especially, have buttons, links and adverts all over the screen. This is distracting as well as confusing for potential learners.

For a program to be worthwhile as a second language tool, the learner needs easy access to their transcripts of their conversations, since these can then be used as follow-up in the classroom or as homework. *Lucy* provides this in a very accessible manner, as does *George*.

In terms of the chatbot with the best ESL potential, the final decision would be a close call between *George* and *Lucy*. *George*, the winner of the 2005 Loebner prize, clearly has the best language capabilities. He is possibly more suited to advanced level learners, who are able to follow (or who are prepared to work at) the twists and turns of his conversations.

*George* scores poorly, however, in terms of cluttered screen layout, no avatar, and poor audio quality of the spoken output. *Lucy* is not as linguistically gifted as *George*, but her screen is clear, she has an attractive appearance and good quality sound output. There is easy access to the ongoing conversation, which can be easily printed or saved. She also has some useful pedagogical features; she can be taught, and she can suggest corrections to certain ungrammatical utterances.

Special mention must go to the *Ultra Hal Assistant*. This is a bold attempt at incorporating voice recognition into the chatbot engine. However, while it gives an indication of what we may hope for in five years' time, currently, it makes so many errors of recognition and incoherent follow-up that its potential for use with ESL learners is not really a possibility.

In terms of exploring and investigating their ability to respond accurately and sensibly in English, it will have been noted that while chatbots have potential, they still have many limitations. Essentially, they cope and respond best when they are given simple one-clause questions or statements which embrace straightforward propositions and little in the way of cohesive links to what has gone before in the conversation. To restate a point from the previous paper, the current generation of chatbots are still unlikely to pass the Turing Test in the near future.

Successful simulated conversation places large demands on the capabilities of natural language parsers/grammars and on speech systems, with the effectiveness of conversational exchanges relying on the ability of artificial intelligence to understand human language and to generate logically consistent and realistic responses.

Nonetheless, technology is progressing at a substantial rate, as may be seen from programs such as the *Ultra Hal Assistant*. Continuing advances in speech understanding technologies, (Fenton-Kerr, 2002) for example, will eventually overcome some of the problems discussed in this paper. And while we still have a long way to go in terms of providing language learners with a robust 'conversation practice machine' (Atwell, 1999) –with all the nuances that such a machine requires– we are getting closer.
Notes
Note 1. The study's analytic perspective is modelled on a Hallidayan (1994) systemic functional grammar view of language structure as its baseline (i.e., words – clauses – text)

Note 2. IELTS (the International English language Testing System; http://www.ielts.org) measures the ability to communicate in English for people who intend to study or work where English is the language of communication. Originally the UK and Australia's tertiary-level English language test, it is now taken by over half a million test takers yearly. It rates test takers on a nine-point scale, with level six generally being the standard required by UK universities for entrance purposes.

References


Coniam, D. (forthcoming) Evaluating the language resources of chatbots for their potential in English as a Second Language learning.

David Coniam
The Chinese University of Hong Kong (Sha Tin, Hong Kong)