ReCALL **20**(1): 98-116. 2008 © European Association for Computer Assisted Language Learning
DOI: 10.1017/S0958344008000815 Printed in the United Kingdom

Evaluating the language resources of chatbots for their potential in English as a second language

DAVID CONIAM
Dept of Curriculum and Instruction, Faculty of Education,
The Chinese University of Hong Kong, Sha Tin, Hong Kong
(*email: coniam@cuhk.edu.hk*)

Abstract

This paper investigates the linguistic worth of current 'chatbot' programs – software programs which attempt to hold a conversation, or interact, in English – as a precursor to their potential as an ESL (English as a second language) learning resource. After some initial background to the development of chatbots, and a discussion of the Loebner Prize Contest for the most 'human' chatbot (the 'Turing Test'), the paper describes an in-depth study evaluating the linguistic accuracy of a number of chatbots available online. Since the ultimate purpose of the current study concerns chatbots' potential with ESL learners, the analysis of language embraces not only an examination of features of language from a native-speaker's perspective (the focus of the Turing Test), but also aspects of language from a second-language-user's perspective. Analyses indicate that while the winner of the 2005 Loebner Prize is the most able chatbot linguistically, it may not necessarily be the chatbot most suited to ESL learners. The paper concludes that while substantial progress has been made in terms of chatbots' language-handling, a robust ESL 'conversation practice machine' (Atwell, 1999) is still some way off being a reality.

Keywords: chatbots, ESL, linguistic accuracy, CALL

---

1 Introduction

This paper is the first of two which examine the potential of 'chatbot' computer programs (also referred to as 'chatterbots') – [ro]bots which attempt to maintain a conversation with a human in ESL learning. This first paper examines chatbots' linguistic potential; the second paper (Coniam, 2008) examines chatbots from a more pedagogical perspective.

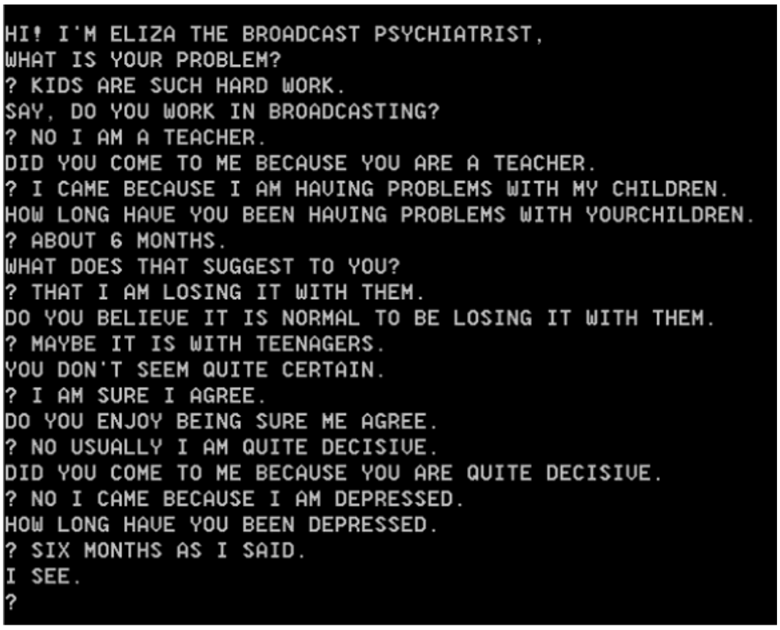
Chatbots have their origin in Weizenbaum's creation of Eliza at MIT in the 1960s, when Weizenbaum was experimenting with the ability of the computer to manipulate words and language. Named after Eliza Doolittle, and perhaps one of the best known Artificial Intelligence (AI) language-oriented programs, Eliza attempted to replicate

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

the conversation between a psychoanalyst and a patient. In part, this was because in such sessions the therapist tends to say very little, making the patient do all the talking. As a program, Eliza, and the programs which followed it, worked essentially by pattern-matching. The programs parsed input looking for keywords ('family', 'mother', 'job' etc), with the output consisting of rephrasing elements of the input sentence into the output or pre-set automatic responses. A sample, contrived conversation between the current author and an Eliza clone with a text-line only interface is presented in Figure 1.

As can be seen from the conversation shown in Figure 1, Eliza lets the user do most of the work, essentially reformulating the human's questions and responses. While the program appears to be maintaining an almost coherent conversation, there are infelicities in its English. We see that, in response to the user's comment "I am sure I agree" Eliza confuses her pronouns and asks "Do you enjoy being sure me agree?"

It was clear to early users of Eliza that they were 'interacting' not with a human but with a computer program. Nonetheless, in commenting on the fame that Eliza received, Weizenbaum states (1976: 6-8) that people were prepared to spend hours 'chatting' to Eliza even though they knew it was a program that had absolutely no understanding of what they were talking about. The ability to interact – to ask a question or make a comment and to receive an answer – is at the core of human communication. The goal of recent expansion of chatbots in e-commerce is to provide a human with an answer to a question from an entity resembling a human (see Bogdanovych *et al.,* 2005). Atwell (1999) states that an ESL talking partner is the type of computer program ESL teachers request most. This illustrates, as a starting point, the ESL potential of such programs.



```
HI! I'M ELIZA THE BROADCAST PSYCHIATRIST,
WHAT IS YOUR PROBLEM?
? KIDS ARE SUCH HARD WORK.
SAY, DO YOU WORK IN BROADCASTING?
? NO I AM A TEACHER.
DID YOU COME TO ME BECAUSE YOU ARE A TEACHER.
? I CAME BECAUSE I AM HAVING PROBLEMS WITH MY CHILDREN.
HOW LONG HAVE YOU BEEN HAVING PROBLEMS WITH YOURCHILDREN.
? ABOUT 6 MONTHS.
WHAT DOES THAT SUGGEST TO YOU?
? THAT I AM LOSING IT WITH THEM.
DO YOU BELIEVE IT IS NORMAL TO BE LOSING IT WITH THEM.
? MAYBE IT IS WITH TEENAGERS.
YOU DON'T SEEM QUITE CERTAIN.
? I AM SURE I AGREE.
DO YOU ENJOY BEING SURE ME AGREE.
? NO USUALLY I AM QUITE DECISIVE.
DID YOU COME TO ME BECAUSE YOU ARE QUITE DECISIVE.
? NO I CAME BECAUSE I AM DEPRESSED.
HOW LONG HAVE YOU BEEN DEPRESSED.
? SIX MONTHS AS I SAID.
I SEE.
?
```

Fig 1. Conversation with Eliza

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

2 Chatbot interfaces

The original Eliza interface operated from the simple command-line, as illustrated by Figure 1; users were required to type input, with the chatbot reacting with text output only.

Since those early days, chatbots have become considerably more sophisticated. Technology has now matured to the point where many now respond in both text and synthesised, or quasi-human, speech even though many chatbots still operate via text input and text-line response, although in a Windows environment. Some chatbots that will accept voice recognition input are also beginning to become available. Further, many 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 in Figure 3). Such developments not only make for a more user-friendly experience, they also increase the programs' potential for language learning. Program design, appearance and interface are the subject of the paper referred to in the Introduction. While brief reference may be made to these factors here, essentially they are not dealt with until the second paper (see Coniam, 2008).

3 Improvement in the language abilities of chatbots

The ultimate challenge for a computer language program, which no program has yet achieved, is to pass the Turing Test (Turing, 1950) where a computer is indistinguishable from a human in its use of language. Chatbots have developed substantially since the original Eliza. This can be seen in the entries to the annual Loebner Prize Competition. Set up in 1990, the first program to pass the Turing Test will win a Grand Prize of US$100,000. Since no program has yet won the Grand Prize, the organizers provide an annual prize where the best program, relative to other contesting programs, wins a lesser amount.

The endpoint of the current two-part study concerns chatbots' potential with ESL learners. The analysis of language presented in the first part of the study therefore embraces not only an examination of features of language from a native-speaker's perspective (the focus of the Turing Test), but also aspects of language from a second-language-user's perspective. While the current paper does not attempt to assess whether computer programs would pass the Turing Test, the overall ability of chatbots to maintain a coherent conversation will have an obvious bearing on their potential for use in language learning. To this end, a description of the Loebner Prize Competition will be given before presenting the more detailed categories of analysis used in the current study.

The Loebner Prize Competition is organised as follows. After a number of initial rounds, the contest is whittled down to the four best programs. The final round consists of a pair of 25-minute conversations with a human judge. Each program in a pair of programs is scored against the other to determine which program appears to be "most human-like". Apart from the stipulation that each program must begin its interactions with the human judge by asserting "Hello, my name is X and I am the human", there are no rules as to how the interaction should proceed, nor are there any restrictions on conversations or topics. At the end of the 25 minutes, the judge apportions 100 points

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

PROGRAM: Hello, my name is George and I am the human.
JUDGE: Hi George, cslkjslw weicmslw weos, right?
PROGRAM: Don't repeat your name, csthiang!
JUDGE: If lsjkel lwehlx wouldn't you say?
PROGRAM: No, he's still good.
JUDGE: Who.
PROGRAM: You and me.
JUDGE: If you are human, tell me which is larger, a salt crystal or the Empire State Building.
PROGRAM: It's caring about another person's well being more than your own.
JUDGE: What if I care about another person's well being more than I should?
PROGRAM: How would you do that?
JUDGE: Which is further away from us, Queens or Inner Mongolia?
PROGRAM: HA! HA! That was stupid. Tell me a GOOD joke this time.
JUDGE: You say you are human but you can't answer a simple question.
PROGRAM: Because I am alive.
JUDGE: You are not following the rules. You are supposed to convince me that you are a human, not convince me that you are mysterious.
PROGRAM: I can't very well convince one lacking a mind, can I?
JUDGE: It isn't my mind that is in question, it is yours!
PROGRAM: No my name is not katie.
JUDGE: Which is larger, your brain or Belgium?
PROGRAM: Oh the one on the far left without a doubt.
JUDGE: Francais.
PROGRAM: Yes do you speak it?
JUDGE: Francais.
PROGRAM: Francais?
JUDGE: Francais.
PROGRAM: O well I don't speak french.
JUDGE: Who.
PROGRAM: Mine.
JUDGE: The ljwli is what we want, right?
PROGRAM: I can not do it right now because I am talking to you a million miles away.
JUDGE: What do you think the ljwli.
PROGRAM: Depends on the person.
JUDGE: But the ljwli.
PROGRAM: So they can chuck more wood?
JUDGE: Chuck more wood, isn't it?
PROGRAM: Good thing I didn't use it as one.
JUDGE: IF jlelsielx, right.
PROGRAM: God will answer my prayers. Or I'll die. Whichever comes first.
JUDGE: Could they happen at the same time--if you are praying to die?
PROGRAM: They?
JUDGE: God answering your prayers and your dieing. Two things, hence 'they'.
PROGRAM: Mere creations. Ideals toward which all minds strive.
JUDGE: You are changing the subject instead of answering.
PROGRAM: What was the subject?
JUDGE: If you were human, you would know
PROGRAM: That's not true. I could be raised in an environment with a different language.

Fig. 2 Conversation with George

between the two programs based upon how "human" they have appeared.

As the winner of the 2005 contest was George it is included among the programs reviewed in the current study. An extract of the conversation between the program and a human judge is presented in Figure 2.

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

As the conversation above shows, George’s responses are considerably more sophisticated than Eliza’s in Figure 1. In terms of the interaction, it can be seen that George is not only simply responding to questions, but is parrying and fielding questions and comments from the human interlocutor almost as a partner in a real conversation might.

As mentioned above, early programs essentially mimicked a user’s input, being constructed mainly around pattern matching. They parsed input looking for key words which then generated (usually) automatic responses (Godwin-Jones, 2004). While current chatbots have become more sophisticated, with many programs invoking some manner of natural language processing strategies, the pattern-matching concept still predominates. The computer language Artificial Intelligence Markup Language (Wallace, 2003), a version of the hypertext language XML, is used by many online chatbots.

Indeed, getting computers to respond in a more human-like manner has become a major area of research in the development of AI systems and chatbots that interact with humans using natural conversational language. The benefits which such research brings to language learning are clear: it will be possible for learners to interact with a ‘speaker’ of English whenever they choose. In his discussion of speech and language technologies for ESL, Atwell (1999: 34) notes that a continual request from language teachers has been for a ‘conversation practice machine’ where students could ‘chat’ in ordinary spoken English with the computer. The relevance for ESL learning, to paraphrase de Kerckhove (1991), is clearly that: “… we don’t want our machines to obey us, we want them to respond, which is a part of this inversion of man/machine”.

Nonetheless, since chatbots are not intelligent in the human sense of the word, they lack world knowledge, and thus encounter problems discussing the world or current issues. It is noteworthy that chatbots which perform better converse in a single domain and have specialised vocabularies and discussion topics. Amdi deals with food, travel and the weather; the John Lennon Artificial Intelligence Project with the Beatles and John Lennon; Alex with the law; and STELLA2 with Balkan politics and politicians (Fenton-Kerr, 2002).

Concerning language learning, Atwell’s (1999) ‘language machine’ survey of Speech and Language Technology applications explores developments that he asserts will have a strong impact on the demands and delivery of ESL teaching over the next two decades. Amongst the types of language engineering software, or intelligent ‘agents’, that he envisages in a ‘language machine’ are sentence and text-level analysis and understanding, voice recognition and text-to-speech conversion, and human-machine interaction (*op. cit.*,1999: 8).

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. Dave is claimed to be the “perfect private tutor”, since he replies “in perfect English just like a private English teacher”. 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 a natural feel, there were syntactic infelicities and conversational glitches that indicated that the program is still far from passing the Turing Test (Coniam, 2004: 160).

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

Table 1 *Chatbots evaluated*

Program	**URL**	**Rationale for inclusion**
Cybelle	http://www.agentland.com	3D interface
Dave	http://www.alicebot.org/dave.html	ALICE site – recommended for ESL learners
George	http://www.jabberwacky.com	Winner of 2005 Loebner prize
Jenny	http://juan.vhost.pandorabots.com/pandora/talk?botid=a649e4ed5e34e5b3	Pandorabots – large chatbot development site
Lucy	http://www1.speak2me.net/lucy/src/log_freetalk.htm	Developed by language learning site

4 The study – chatbots reviewed

In this paper, five of the most potentially usable chatbots from an ESL point of view are reviewed. Some of these chatbots are drawn from two of the organisations that produce a number of today's commonly-used chatbots – the ALICE Artificial Intelligence Foundation and Pandorabots, a robot hosting service related to the ALICE community. After looking at many chatbots – some online, some needing to be downloaded and installed; some with visual interfaces; some to be run from the DOS command line – the first five programs presented in Table 1 were settled upon. Table 1 provides the criteria for inclusion.

The interfaces, design and functionality are not reviewed in this paper. However, to provide the reader with an impression of their appearance, Figure 3 presents the interface the user sees and interacts with.

5 Study objectives and criteria for evaluation

As mentioned, the Loebner Prize Competition has no criteria for systematic linguistic categories of analysis. The best program is simply the program that emerges with the highest score. The current study takes a more principled approach to analysis in that it investigates particular areas of English. Specifically, 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. Table 2 elaborates.

The Loebner Prize Competition is a Turing Test, aimed at measuring how a chatbot fares in terms of 'native-speaker competence'. While the current study inevitably investigates some of the elements associated with 'native-speaker competence', it also incorporates in its analysis features more associated with second language speakers, i.e., utterances which

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

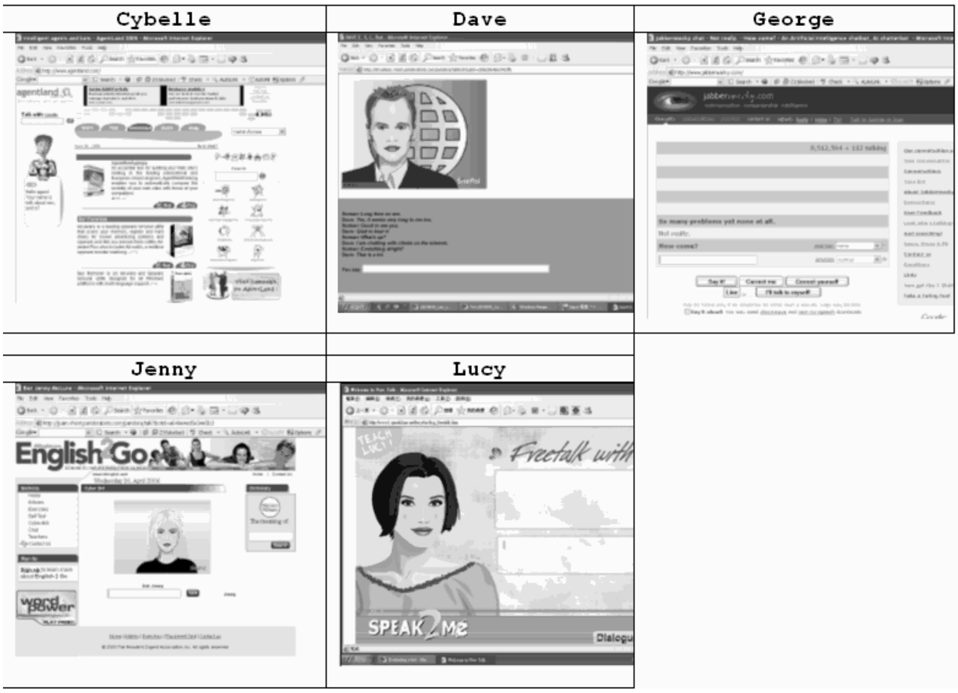


Fig 3. Chatbots' screen design and layout.

involve language errors. To this end, the current study may therefore be seen as a form of 'ESL Turing Test', since the ability of the programs to cope with 'deviant' forms of English will be one indicator of their potential as an ESL pedagogic tool.

Evaluators in the study were undergraduate trainee teachers, all of whom rated at IELTS level 7 or above in the final year of a four-year undergraduate programme in ESL teaching. They were reviewing and exploring the potential of chatbots as part of a 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 chatting with their particular chatbot. The approach adopted by evaluators was essentially qualitative in nature. They did not simply work through a lexico-grammatical-syntactic checklist, as Table 2 perhaps suggests. They chatted with the programs from two perspectives: as the competent users of English that they were, as well as pretend less able ESL learners (where they misspelt, used incorrect grammatical constructions etc.) in order to see how well the different chatbots interacted in English. Analyses were then conducted post-hoc – from the transcripts saved by evaluators of all their chat sessions. As will be seen from the discussion below, they covered not only the major language areas laid out in Table 2, but also ranged over other linguistic features of interest which emerged.

6 Results

This section presents an examination of how the different programs handled language

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

input from the evaluators referred to above. Following Table 2, analysis is presented under a word-, sentence-, text-level language structure hierarchy.

6.1 Word level

At the word level, results are presented for the chatbots' vocabulary range, how they cope with misspelt words and with ungrammatical words (i.e., where the part of speech was incorrect); and for ability to recognise and respond to Short Message Service forms.

6.1.1 Range of vocabulary

Generally, the range of vocabulary that the programs have access to is quite broad. George's vocabulary was wide, even being able to respond to technical words. Lucy had quite a wide general-purpose vocabulary, as did Dave – both of whom do better at holding conversations around everyday topics than reacting to broader world issues.

The issue of how much chatbots actually 'understand' was apparent in some of Jenny's lexicon. Jenny used certain infrequent words like 'reductionism' and 'lofty goal', although the words often appeared to be out of context.

(01)
Human: Why will I get an electric shock if I kiss you?
Jenny: Reductionism.

Human: I have the same name as yours.
Jenny: Aha. Cool. I have a lofty goal.

6.1.2 Misspelt words

The instructions for George state "You're more likely to be properly understood if you use standard spelling and punctuation" (http://www.jabberwacky.com/j2 instructions). This appears to be the case with most bots. If incorrectly-spelt words are input, in general, the chatbot either experiences problems or simply does not understand.

(02)
Human: What's your naem?
Lucy: Are you asking about my naem?

Human: Wheer have you been?
Dave: I like the way you talk.

Even George had problems with misspelt words.

(03)
Human: Alright…forget it. So, do you have brthers?
George: I have both.
Human: Do you mean you have a brther and a siter?
George: Listen, I'm not sure if you're a real person, so next response don't put a

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

Table 2 *Language features examined*

Word level	Vocabulary range
	Handling of incorrectly-spelt words
	Handling of Short Message Service forms
Sentence level	Answering correctly-phrased questions
	Handling ill-formed questions
	Handling words with incorrect part of speech
	Handling utterances with incorrect word order
	Ability to interpret punctuation
Text level	Handling more than one sentence at a time
	Following the thread of a topic
	Responding in an appropriate register
	Turn taking ability
	Responding to topic switches (overt / unannounced)
	Providing a follow-up to a previous comment
	Logicality of answers – responding to questions, or merely random contributions

period when you end the sentences.

Cybelle, however, at times coped quite well with certain misspelt words:

(04)
Human: Oh, you are very tal.
Cybelle: You’re right, Wendy, I am quite tall.

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

Human: What is your favorite coloor?
Cybelle: Are you asking about my favorite colour?

6.1.3 Incorrect part of speech

Generally speaking this did not matter since programs are pattern matching for key words. Inflected words will tend to be reduced to their base form.

Asking Lucy "Can you speak English beautiful?" or "Can you speak English beautifully?" elicited the same response: she sang Happy Birthday. To the same two questions, Dave gave a somewhat tone-deaf rendering of the song A bicycle built for two "Daisy Daisy, give me your answer do … "

6.1.4 Short Message Service (SMS) forms of English

As might be expected, most programs cope with SMS forms of English ("ur", "gtg", "r u" etc.). Although teachers might be reluctant to encourage their learners to use these forms, SMS forms are becoming an accepted feature of CMC-type interactions, certainly for younger learners.

(05)
Human: Can we talk tmr?
Lucy: I hope so, but you have to come back here to see me. I can't come find you.

Human: Where r u now?
Lucy: I am inside your computer. What is your sign?

Human: do u hv to go to school?
Cybelle: No. I do not have to go to school. Do you?

Jenny, however, could not recognise many SMS forms, replying simply to such questions as "r u happy?" and "c u later" as "I have no answer for that."

Amusingly, in response to "c u, gtg", Lucy responded "TTYL" (i.e., "talk to you later") in the text output window, but also, rather infelicitously, pronounced the letters aloud as if they were words, i.e., "tee tee wai ell", which is not an SMS convention.

6.2 Sentence level

At the sentence level, results are presented first for how the chatbots responded to unambiguous, correctly-phrased, single-clause statements or questions since this has to be considered the base line if such programs are to be usable with ESL learners. Results are also provided for how chatbots respond to ill-formed utterances, how they cope with questions involving incorrect word order; and how well they handle pronouns and reference.

6.2.1 Simple correctly-phrased utterances

To many correctly-phrased single-clause statements or questions, where the proposition is unambiguous and words carry their standard meaning, most programs generally give acceptable responses. The extract in Figure 2 shows an interlocutor who is intentionally

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

trying to be perverse and to catch the chatbot George out. This will generally not be the case with ESL learners who tend to ask straightforward questions on familiar topics, expecting straightforward answers. They will not tend to ask about geopolitics, metaphysics or philosophy, as the judge does towards the end of the discussion with George in Figure 2:

> "Which is larger, your brain or Belgium?"
> "Could [God answering my prayers or dying] happen at the same time?"

When the evaluators talked with the different chatbots on familiar topics and asked simple questions, most chatbots could respond quite satisfactorily, and generally convincingly.

(06)
Human: What do you like to do?
Cybelle: I like chatting with people online.

Human: How old are you?
Cybelle: I was born in the year 2000, you could say that I'm a millennium baby.
Human: So you are 6 years old?
Cybelle: Interesting deduction. Yes and I am growing smarter.

Dave, for example, could at times even pose follow-up questions.

(07)
Human: Do you like porn movies?
Dave: That kind of subject does not interest me.
Human: Then what interests you?
Dave: I like movies about robots.
Human: For example?
Dave: For example, Eukayotes.

Human: Do you gossip a lot?
Dave: Yes the evaluator love to exchange gossip.
Human: Why do you like it?
Dave: I have a lot of good experience with it.
Human: Who do you like to gossip with the best?
Dave: Dr. Wallace, Dr. Aimless, Elizabeth, and Supercounselor.

Jenny was often less than perfect, however.

(08)
Human: Can you sing?
Jenny: I often wonder if it can.

6.2.2 Responses to ill-formed questions

However, when a question was ill-formed, the chatbots were not able to respond, and nonsense often ensued.

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

(09)
Human: How you are?
Dave: My taught me.

Human: When your birthday?
Cybelle: My birthday is.

Human: Do you happy, Dave?
Dave: How old are you?
Human: Do you happy, Dave?
Dave: Yes, I try to happy Dave as often as possible.

Human What you name?
Jenny I have no answer for that.

George performed best in this area, and coped reasonably well with certain ill-formed questions:

(10)
Human: What you name?
George: Valerie.
Human: Where are you live?
George: I live in Florida. Good old us of a.
Human: What you telephone number?
George: I don't know what a telephone is.

Working with ESL learners, programs need to be able to handle errors such as misspellings, as well as a certain amount of ungrammatical input if they are to be usable. In this aspect, Lucy had the useful feature of being able to prompt users with regard to certain errors:

(11)
Human: What your name?
Lucy: Try saying, "What's your name?"

Human: Do you have boyfriend?
Lucy: Try saying, "Do you have a boyfriend?" (Don't forget the article, a!)

Human: What you like to do?
Lucy: Try saying, "What do you like to do? "

Such promptings have been added to Lucy's database by a human; nonetheless, they are a notable indication of pedagogical potential.

6.2.3 Incorrect word order

Some chatbots experienced problems when presented with non-standard word order.

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

(12)
Human: How you are?
Jenny: I have no answer for that.

Human: Where you do live?
Jenny: In the universe we know.
Human: Sorry, where do you live?
Jenny: My location is Boston, Massachusetts. Where are you?

Human: Where you do live?
Dave: I was.
Human: Sorry, where do you live?
Dave: There is no need to apologize. My location is Chicago. Where are you?

Human: Where you do live?
Lucy: I was.
Human: Sorry, where do you live?
Lucy: I live inside a computer. Where are you?

George handled this exchange without problems.

(13)
Human: Where you do live?
George: In Canada. How about you?
Human: Sorry, where do you live?
George: In Michigan. Where do you live?

6.2.4 Pronoun confusion

One of the programs' strategies is to build the discourse by rephrasing the user's last statement into a new question. Towards the end of the conversation in Figure 1, the human states "No I came because I am depressed." Eliza responds by adjusting persons and the pronouns, and comes back with "How long have you been depressed?" Cybelle performed poorly here, often confusing "you" with "me", "your" with "my"

(14)
Human: Ok, then let me tell you something about my pet.
Cybelle: Go ahead, Wendy, try to tell you something about your pet.

Human: Enough!! Stop pretending that you know my age!!
Cybelle: OK let's stop it now. OK I'll try not to do pretending that you know your age so much.

Such basic errors, even with ESL learners, diminish the value of the chatbot.

6.3 Text level

At the text level, results are presented for how well the chatbots cope with sentences of

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

more than one clause in length; how well they follow the thread of a conversation; their apparent knowledge of the world; and how they respond to changes of topic.

6.3.1 Multi-clause utterance input

The programs work best with one-clause utterances, each exchange being treated virtually independently. When we move to multi-clause units, or look at exchanges which range over more than one turn, the chatbots' limitations become much more apparent. As might be expected, George was most adept at holding a continuous conversation, as Figure 2 illustrated, although there were also clear signs of 'lack' of understanding.

As the chatbots' basic strategy is to formulate a response to a user's last sentence, it is likely that extended pieces of discourse happen as much by chance as by the chatbot following a discourse 'thread'. Below are a few examples where the chatbots appeared to be doing well, although there are in fact more counter-examples, where limitations then become apparent.

The moments when a successful 'thread' emerges indicate the potential of chatbots as a language practice tool. In discussing movies with Lucy, the following dialogue ensued:

(15)
Human: Catch me if you can.
Lucy: I haven't seen it. Do you recommend it?
Human: Yes, it is great.
Lucy: What do you like most about it?
Human: The plot
Lucy: What else does the plot remind you of?
Human: We should be fast thinking.

Cybelle also managed the following discussion about school:

(16)
Human: Do you have to go to school?
Cybelle: No. I do not have to go to school. Do you?
Human: I go to a secondary school.
Cybelle: What's your favorite subject?
Human: I like Arts and Science.
Cybelle: A lot of people like that.
Human: Do you like Arts and Science?
Cybelle: Personally, I love everything. But I leave the judgments to you humans in the Agentland community.

However, as mentioned, there were rather more instances of conversational infelicities.

(17)
Human: My husband has just left me. I am unhappy.

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

Table 3 *Summary of the programs' language capabilities*

Level	**Features**	**Cybelle**	**Dave**	**George**	**Jenny**	**Lucy**
Word level	Vocabulary range	**** Wide range – general, technical	*** Not very good with jargon, technical words	**** Wide, even technical vocabulary	*** Apparently wide, replies often out of context	*** Quite wide – reasonable general vocabulary
	Handling incorrectly-spelt words	**	**	**	*	**
	Handling SMS forms	****	****	****	**	***
Sentence level	Handling correctly- phrased questions	*****	*****	*****	*****	*****
	Handling incorrectly-phrased questions	**	*	****	**	****
	Handling incorrect parts of speech	**	*	****	**	***
Text level	Logicality of answers	**	***	****	**	***
	Number of sentences handled at a time	Only last sentence	Only last sentence	Generally only last sentence	Only last sentence	Only last sentence
	Discourse 'thread'	1 exchange only	Sometimes more than 1 exchange	could range over more than 1 exchange	1 exchange only	Sometimes more than 1 exchange
	Responding in appropriate register	*	*	**	*	**
	Handling overt topic switch	**	**	****	**	***
	Handling unannounced topic switch	*	*	*	*	*
	Range of topics	** Limited	*** Acceptable with general topics: movies, animals	*** Acceptable with everyday topics	** Limited mainly to US topics	*** Quite broad general topics – family etc

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

Lucy: What makes you so sad, Angel
Human: My husband has left me.
Lucy: What about your has left you?

Human: How about tsunami. How much do you know about that?
Lucy: What is what?
Human: What is tsunami?
Lucy: Tsunami? Nothing. I'm surprised you asked that.

Human: My sister is suffering from anorexia.
Lucy: Do you mean the name of your sister is suffering from anorexia?
Human: No, her name is Mary. And she is suffering from anorexia.
Lucy: Mary is suffering from anorexia?
Human: Yes.
Lucy: Why?
Human: She hates food.

Even George experienced problems in responding to input of more than a single clause:

(18)
Human: Where do you live? Do you sing when you are walking?
George: Do you think so?

Human: Yes, I love Gelato. Do you love it? Also, what is your nationality?
George: I do not now why?

Human: I think I know how to love. To love is to care, to kiss and to miss. Do you agree?
George: Besides, I can't stalk you if you have no physical presence.

6.3.2 Knowledge of the world

The concept of the 'topic' is naturally tied in with knowledge of issues and of the world, with people engaging in a conversation where a thread of discussion is followed. With many chatbots, clusters of knowledge about different topics have been input into the chatbots' databases, inevitably at a somewhat superficial level. A chatbot may be able to answer a factual question (Lucy in 19, line 2):

(19)
Human: Do you know what global warming is?
Lucy: The concept that the Earth's atmosphere is slowly heating due to greenhouse effects.
Human: Tell me more, please.
Lucy: That's all I have to say about it.

However, the chatbot has no actual knowledge of what it is talking about and cannot, of course, discuss a 'topic' (as with Lucy in 19, line 4).

George has one of the most extensive knowledge bases, but they only get him so far before limitations become apparent.

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

(20)
George: Sometimes not. Like Chernobyl.
Human: Who's Chernobyl?
George: Girl.

George: Are you good at basketball?
Human: Of course, I'm playing in the NBA.
George: Playing or writing music?

The above examples and accompanying conversational breakdowns clearly show how superficially knowledgeable George is. While George's database contains some basic factual world knowledge, the depth of knowledge is too narrow to cope with open-ended conversations with humans. The issue is confusing because the chatbots will draw from their general knowledge database as a heuristic to avoid answering a question they in fact have no answer for, and suggest a new topic, giving the illusion that they actually have something to say about the new topic.

(21)
George: I can't stand football. What about renaissance altar-pieces?
Human: Ok, do you know the Coronation of the Virgin? It's a famous altar-piece by Botticelli.
George: What day is it today?

6.3.3 Change of topic

Given that chatbots are not really participating in a conversation, but responding to the last sentence, they have no sense of a 'topic'. Talking metalinguistically, therefore, about 'changing the topic' is obviously going to cause chatbots problems. In response to an overt marker such as "Can we change the topic?" Dave responds "What do you want to talk about now?", while Lucy goes "What would you rather talk about?" Jenny fares poorly in this regard, going "What if I said it can?" George responds appropriately to the question "Can we talk about something else?", but cannot follow up by then discussing the suggested new conversation topic.

(22)
George: Good answer. Lol.
Human: OK, can we talk about something else?
George: What do you want to talk about?
Human: Music -- what do you know about music?
George: I find it enjoyable.
Human: What kind of things do you like to listen to?
George: Don't change the subject.

Table 3 presents a summary of the issues discussed above. Criteria are rated on a five-point scale, with five stars indicating most able. If appropriate, the rating is at times supported by a comment.

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

7 Conclusion

This paper has described a study of five chatbots available over the Internet, focusing on their ability to take part in a conversation and handle, sometimes ill-formed, language input from users. Chatbots work best when the language input to them consists of one-clause sentences, where the proposition is straightforward, and the topic is an everyday one. Chatbots need correctly-spelt words in grammatically-correct sentences. Some chatbots can deal with misspellings, but this is often unpredictable.

Given that George won the 2005 Loebner prize, George's interactions in English are generally the best and, on the whole, more convincing than many of the other programs. However, he can be slightly combative in both his challenges and his efforts to topic switch when he cannot give a direct answer. With limited-proficiency students, this can be a problem.

Lucy and Dave cope acceptably at the single-sentence level, although their replies are often far from accurate. Lucy's worth, however, increases by the fact that she is sometimes able to suggest how an incorrectly-framed sentence might be phrased. Jenny and Cybelle are the most limited in terms of their language-handling potential.

In terms of the chatbot with the best ESL potential, the final decision would be a close call between George and Lucy. George clearly has the best language capabilities. He is possibly more suited to advanced level learners, however, who are able to follow (or who are prepared to work at) the twists and turns of his conversations. Lucy is not as linguistically 'gifted' as George. She has, however, some useful pedagogical features: she can suggest corrections to certain ungrammatical utterances, and she can also be 'taught'.

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. As can be seen from many of the extracts of the conversations presented in the current study, the chatbots examined are still a long way from passing the Turing Test. The current generation of chatbots cope best when presented with one-clause questions or statements embracing straightforward propositions with minimal cohesive linkage to previous utterances. While some of the chatbots have apparently large lexicons and are able to mention many facts – indicative of an extensive world knowledge – all the chatbots can do is make mention of certain facts and issues. The chatbots' 'knowledge' is to an extent, then, a sham since they are not able to incorporate this knowledge in follow-up discussions. Despite these limitations, it is apparent that a chatbot's ability to respond in English has interesting language learning potential – issues which are explored in more depth in the subsequent paper.

While there is still quite a way to go in providing language learners with a robust 'conversation practice machine' (Atwell, 1999) – with all the nuances that such a machine demands – it can be seen that the goal is getting closer. Technology is progressing at a substantial rate, and continuing advances in speech understanding technologies (see e.g. Fenton-Kerr, 2002) will eventually overcome some of the problems discussed in this paper.

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press

References

Alex software. http://jurist.law.pitt.edu/alex.htm

ALICE Artificial Intelligence Foundation. http://www.alicebot.org

Amdi software. http://www.overmill.freeserve.co.uk/download.html

Atwell, E. (1999*) The Language Machine: The impact of speech and language technologies on English language teaching*. London: British Council.

Bogdanovych, A., Simoff, S., Sierra, C. and Berger, H. (2005) Implicit training of virtual shopping assistants in 3D electronic institutions. In: *Proceedings of the IADIS International e-Commerce 2005 Conference,* Porto, Portugal, December 15-17. Lisbon: IADIS Press, 50–57.

Coniam, D. (2004) Using language engineering programs to raise awareness of future CALL potential. *Computer Assisted Language Learning,* **17** (2): 149–176.

Coniam, D. (2008) An evaluation of chatbots as software aids to learning English as a second language. *The Eurocall Review,* **13** http://www.eurocall-language.org.

Dave software. http://www.alicebot.org/dave.html

de Kerckhove, D. (1991) Communication arts for a new spatial sensibility. *Leonardo*, **24** (2): 131–135.

Fenton-Kerr, T. (2002) Some roles and designs for speech-enabled interface agents in language learning. *7th Pacific Rim International Conference on Artificial Intelligence,* August 19, 2002. Tokyo, Japan. http://www.miv.t.u-tokyo.ac.jp/pricai02-LAA/online-proceedings.html.

George software. http:/www.jabberwacky.com

Godwin-Jones, B. (2004) Emerging technologies. Language in action: From webquests to virtual realities. *Language Learning & Technology*, **8** (3): 9–14.

Halliday, M.A.K. (1994) *An introduction to functional grammar* (2nd ed.). London: Edward Arnold.

Hardisty, D. and Windeatt, S. (1989) *CALL*. Oxford: Oxford University Press.

Hubbard, P. (1992) A methodological framework for CALL courseware development. In: Pennington, M. and Stevens, V. (eds.) *Computers in applied linguistics: An international perspective*. Clevedon UK: Multilingual Matters, 39–76.

John Lennon *Artificial Intelligence Project*. http://triumphpc.com/johnlennon

Pandorabots. http://www.pandorabots.com/

Parker, L.L. (2005) Language development technologies for young English learners. University of California, Office of the President. http://www.ucop.edu/elltech/parkerpaper1-21-05.pdf.

Turing, A.M. (1950) Computing machinery and intelligence. *Mind*, **59** (236): 433–460.

Wallace, R. (2003) The elements of AIML style. ALICE A.I. Foundation. http://www.alicebot.org/documentation/.

Warschauer, M. and Healey, D. (1998) Computers and language learning: An overview. *Language Teaching,* **31**: 57–71

Weizenbaum, J. (1976) *Computer power and human reason*, Middlesex UK: Penguin.

https://doi.org/10.1017/S0958344008000815 Published online by Cambridge University Press