DISRUPTIVE SCIENCE AND TECHNOLOGY
Volume 1, Number 2, 2012

DOI: 10.1089/dst.2012.0010

ORIGINAL RESEARCH ARTICLE

Improving Speech Recognition with the Robot Interaction Language

Omar Mubin,[1] Christoph Bartneck,[2] Loe Feijs,[3] Hanneke Hooft van Huysduynen,[3] Jun Hu,[3] and Jerry Muelver[4]

Abstract

This article presents the design and evaluation of a Robot Interaction Language (ROILA). This speech recognition friendly spoken artificial language is designed to be used by humans for interacting with robots. We evaluated the use of ROILA in a Dutch high school. The language was taught as a part of the science curriculum followed by a controlled experiment, where the language was compared against English. The results from the experiment showed that the ROILA performed better than English on account of both objective recognition accuracy and the subjective assessment by the students. We estimate the trade-off between this benefit in relation to the effort required to learn ROILA. In a regular usage scenario, it would pay off to use ROILA.

Key words: artificial; language; recognition; robot; ROILA; speech

Introduction

The number of robots in our society is increasing rapidly and already in 2008, the sales of service robots outnumbered the sales of industrial robots.[1] An easy way to communicate with service robots, such as Roomba or Nao, is natural speech. Prasad et al. even go as far as describing speech interaction with robots as a"Holy Grail".[2] However, the limitations in speech recognition technology for natural language is a major obstacle for the general introduction of speech interaction for robots. At times, current speech recognition technology is not good enough for it to be deployed in natural environments where the ambience influences its performance.[3] One of the problems that speech recognition technology is facing is the inherent properties of natural language. Examples are dialects, ambiguity in context, grammar irregularities, and homophones (words that sound the same but have different meanings).[4] As a consequence, miscommunication occurs frequently between the user and the robot, which leads to a considerable frustration for the user. The Palm company faced a similar problem with handwriting recognition for their handheld computers in the 1990s. To overcome the insufficient recognition accuracy, they invented an artificial alphabet: Graffiti (see Fig. 1). It is easy for users to learn and easy for the computer to recognize this new alphabet. Our Robot Interaction Language (ROILA) takes a similar approach by offering a speech recognition friendly artificial language that is easy for users to learn and easy for machines to understand.

An artificial language, as defined by the Oxford Encyclopedia, is a language that is deliberately invented or constructed, especially as a means of communication in computing or information technology. However, within the domain of artificial languages as mentioned in the definition earlier, both spoken and nonverbal languages such as formal languages or programming languages exist. However, our focus is on spoken artificial languages only; therefore, from now on, whenever we state artificial languages in this article we refer to spoken artificial languages only. Several artificial languages exist,[5] and Esperanto[6] might be one of the most acknowledged. To the best of our knowledge, none of the available spoken artificial languages (such as those mentioned in[5]) were optimized for human machine interaction but were rather created to facilitate human-human communication ("also known as universal languages"). Therefore, the focus of our research was to design a speech-recognition friendly artificial language that humans could use to talk to robots. First attempts at creating a speech-recognition friendly language have been made by constraining the use of a natural language. Rosenfeld, Olsen, and Rudnicky[7] argued that constraining language is a plausible method of improving recognition accuracy. The user experience of an artificially

[1]School of Computing, Engineering and Mathematics, University of Western Sydney (UWS), Penrith, Australia.
[2]University of Canterbury, HIT Lab NZ, Christchurch, New Zealand.
[3]Eindhoven University of Technology, Eindhoven, The Netherlands.
[4]North American Ido Society, Inc., Madison, Wisconsin.

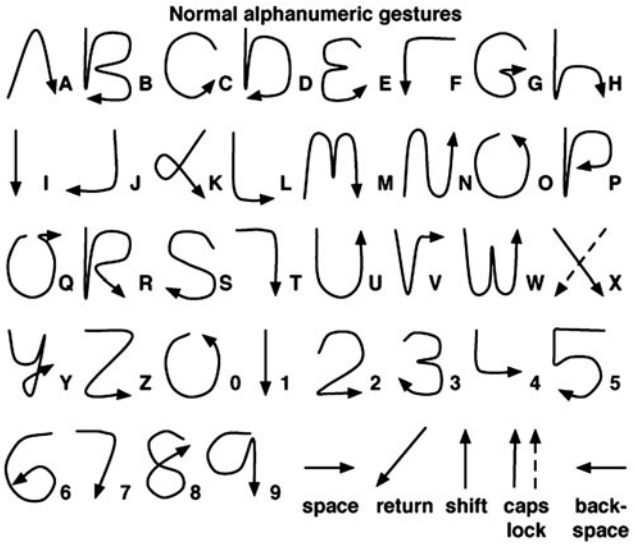

FIG. 1. Graffiti from Palm Inc.

constrained language was evaluated within a movie-information dialog interface, and it was concluded that 74% of the users found the constrained language interface to be more satisfactory than a natural language interface[8]. Another example is the constrained command language proposed in[9] that is intended to be used for interacting with various appliances.

Several attempts have been made to ensure efficient speech interactions between robots and human beings. Prasad, Saruwatari, and Shikano discuss the pros and cons of speech-enabled robots.[2] Besides efforts in improving verbal human–robot interactions, spoken languages for robot–robot communication have started emerging,[10] where the focus can also be to teach robots languages along the way of improving human–robot interaction in the long run.[11] Still, the main thrust of research in human–robot verbal interaction is focusing on providing controlled command languages to interact with robots, and first results have become available.[12,13] These constrained languages are inherited from natural languages, and, are, therefore, potentially easy to learn. In addition, constrained languages may limit the vocabulary, grammatical structures, or even go as far as modifying words.

Another stream of research that improves speech recognition is the use of a multimodal input, for example, by using cameras within the robots' environment to record the gestures of the user.[14–16] Therefore, the robot is not reliant on only one modality, and, consequently, there is a prospect of better human–robot communication. As reported in,[17] the robot tracks the gaze of the user when the object or the verb of a sentence in a dialogue is undefined or ambiguous.

However, little is known about how difficult it is to learn such constrained languages, and the available data on the efficiency of such afore-mentioned multimodal speech systems are still scarce. For example, in systems such as those described in,[18,19] speech recognition technology has been shown to work with robots in controlled settings, but their tests were only carried out with a limited number of speakers (in both cases, 3 speakers). Therefore, it is debatable as to whether such systems would work outside the lab and when exposed to a large group of users.

Given the high potential for speech interaction between users and robots, we explore a spoken artificial language that aims at achieving a new balance between, on the one hand, being easy to learn for users and, on the other hand, being easy to recognize for robots. Humans are incredibly adaptive, and an artificial language that utilizes this advantage might have the potential to outperform natural languages (in terms of recognition accuracy) in the interaction between users and robots. We, therefore, go beyond constrained languages by developing and evaluating a ROILA. This truly new artificial language is optimized for automatic speech recognition and, in theory, requires minimal learning effort from humans. There have been first efforts in designing speech-recognition friendly artificial languages to talk to machines,[20,21] but the results of these studies are fairly limited. In both studies mentioned,[20,21] the vocabulary size was small, and the languages had no grammar, that is, commands consisted of solitary words. Moreover, no results of formal evaluations with users are available.

It is a clear disadvantage that users need to learn ROILA to interact with robots. However, similar to learning to type with ten fingers, it might pay off in the long run. We, therefore, need to keep the efficiency of ROILA in mind by monitoring the learning effort and the potential gain in recognition accuracy. The main question of this study is whether we can find a new balance for speech interaction technology that offers a better trade-off than the current natural speech-recognition systems. We briefly discuss the design of ROILA before presenting an evaluation of ROILA.

The Design and Implementation of ROILA

The design and implementation of the language is described in greater detail in.[22] In this article, we only briefly summarize the design process. First, we reviewed the grammar and phonology (the sounds of the language) of major natural and artificial languages. The review revealed certain design patterns that we considered in the selection of phonemes and choice of grammatical constructs. Using the questions, options, and criteria technique,[23] design decisions were made for grammatical markings. The two most important factors in this process were the effort it would take humans to learn a language and the potential recognition accuracy of speech engines. We drafted a concise grammar that has no irregularities or exceptions. Its rules and markings are represented by adding isolated words (word markers) rather than inflecting the existing words of a sentence. A genetic algorithm was used to generate a vocabulary in which the words have the least likelihood of being confused with each other. The vocabulary went through several iterative cycles of construction and user evaluation. Semantics of ROILA were simply borrowed from Basic English on the basis of word frequency. Therefore, the shorter the word in ROILA, the more frequently occurring word from Basic English was assigned to it. A sample ROILA sentence would be "pito botama jifi webufo," translating to "I turned left," with a literal translation of "I turn <word marker past tense> left."

The first step in our implementation of ROILA was to review existing speech recognizers. We concluded that Sphinx-4[24] was the most suitable in terms of accuracy, open source license, and flexibility for the modification that was necessary for implementing ROILA. The next step was the

choice of an acoustic model, which determines how phonemes have to be pronounced. We had to decide to either use an existing model or create a specific ROILA model. New acoustic models are usually built by a large corpus of recordings from many native speakers.[25] This is not yet possible for ROILA, as we have no native speakers. We, therefore, decided to adopt the established North American acoustic model that is already a part of Sphinx-4, as it contains all the phonemes required for speaking ROILA. We extended Sphinx-4 by adding a phonetic dictionary and also by providing a new grammar. We used the Festival[26] Text-To-Speech (TTS) engine to enable the robot to speak ROILA.

We selected the LEGO Mindstorms NXT platform[27] as the prototyping platform for the first ROILA-enabled robot, and the LEGO company kindly donated 20 Mindstorms NXT sets to our project. We used the Java-based Lejos firmware, which simplified the communication between the java-based speech recognition software running on the computer and the robot. The complete setup consisted of a LEGO NXT robot, a Blue Snowflake USB microphone, and a Bluetooth-enabled laptop. The Sphinx speech recognizer and the Festival TTS were set up to run on the laptop.

Training of the Participants

To be able to run experiments with ROILA, we first had to train users. Any experiment would otherwise only measure the difference between a known and an unknown language. Obviously, a user would be unable to interact with a ROILA robot if they could not speak and understand ROILA. The Christiaan Huygens College in Eindhoven, The Netherlands, allowed their students to participate in our study. As a token of appreciation, we donated 10 LEGO Mindstorm kits to the school. Our ROILA study was integrated into their robotics module, which is a section of their science class. The students received credits for their participation and by virtue of points scored in the final ROILA exam. A ROILA curriculum was carefully designed for the students to aid them in their learning both in school and at home. The exam was based on the curriculum and tested the knowledge of ROILA of the students across vocabulary, pronunciation, and grammar. The students spent 3 weeks learning ROILA both at school and at home. In-school learning was more interactive and hands on as the students tested their ROILA skills by speaking to and playing with LEGO robots (see Fig. 2). In-home training was online readings and short tasks that supplemented the material studied in class. The online training was implemented using the Moodle software (www.moodle.com).

The ROILA lessons

The three lesson plans were carefully designed in collaboration with the science teachers. Each lesson comprised two parts: a theoretical part and a practical part, with the ROILA exam taking place at the end of the third lesson. In the theoretical part, students were introduced to the linguistic elements of ROILA. In total, they learned 50 words (For the vocabulary list used in the curriculum, see Table 1). In the practical part, we designed simple interaction exercises in which the students practiced what they had learnt by talking ROILA to the LEGO robots. Examples of such mini exercises were efficiently navigating the robot so that it does not fall from the table and a robot race between two or more robots. Therefore, the ROILA vocabulary comprised words primarily related to simple interaction scenarios with the robot such as greeting messages, navigation, shooting, and sensing colors. Some other words in the vocabulary were related to emotions and other everyday contexts. The 20 LEGO Mindstorm robots were divided equally in class; so, each robot was placed in groups of two or three children at the most. One lesson was offered per week, and each lesson lasted for 100 min. The language of instruction during the lessons was English. The science teachers were always present in class to help the students. The usual medium of instruction of their science classes was a mixture of Dutch and English. The interaction was not conducted in English (we only used ROILA), as we felt that this would create an unnecessary bias in the minds of the children when they would compare the recognition performance of ROILA against English before the controlled experiment that we wished to conduct at the end of the three lessons. Our assumption was that since the students already had prior training in English (in other contexts and

FIG. 2. Students interacting with the robots during class.

Table 1. Robot Interaction Language Vocabulary Used in the Curriculum

ROILA	*English*
bofute	start
kanek	go
koloke	forward
botama	turn
webufo	left
besati	right
nole	back, backward
jimeja	quickly
kipupi	slowly
buse	no, not
kufim	to, toward
jutof	like
make	see
bama	you
pito	I
kilu	one
seju	two
tewajo	three
tuji	many (plural marker)
jinolu	ball
jesime	step
saki	have
tifeke	red
wipoba	yellow
wekepo	color
kasok	big
kute	little
malula	size
wapisi	bucket
biwu	what
wopa	good, okay, right
pojos	zero
fibi	four
jitat	five
silif	six
kutuju	blue
koleti	green
tobop	put
lamab	now (at this time)
wekapa	error
bemeko	wrong
wolej nawe	other way
sowu	and
buno	or
jifi	past tense (marker)
jifo	future tense (marker)
bafop	into, in
nelete	pick (up), lift
jasupa	put down, drop (v.)
lujusi	box
bileki	carry, bring
bobuja	run
fosit	walk

ROILA, Robot Interaction Language.

scenarios however), we would only train them in ROILA during the lessons.

A lesson booklet was also provided to the students where they could write down notes from both the lessons in class and from their homework. The booklet became their diary, and we also asked the students to record how much time it took them to complete the homework. Before the commencement of the 2nd and 3rd lesson, we checked the booklets to see whether the time recording had been carried out. At the end of the 3rd lesson, the booklets were collected from the students. Overall, the students appeared motivated to learn ROILA and to interact with the robots. Some students created their own customized robots and scenarios during the lessons. For example, they set up obstacles in a maze-like fashion and made the robots navigate around them by giving ROILA commands (see Fig. 2).

Evaluation of ROILA

We conducted a controlled experiment that evaluated the recognition accuracy of ROILA against English and which also assessed the learnability of ROILA. Furthermore, we looked at the subjective impressions that the participants had of ROILA. The experiment was conducted during the week after the end of the ROILA lessons (4th week of the curriculum). The experiment was set up to address the following research questions:

1. Do the users perceive the use of ROILA to be easier than the use of English when talking to a speech system?
2. Is the speech recognition accuracy higher when users speak ROILA in comparison to when they speak English?
3. Under the assumption that ROILA has higher recognition accuracy, how long do the users have to interact with the speech system before their initial investment of learning ROILA pays off?

Participants

We selected Dutch students between the age of 12 and 15 to participate in the study. We selected non-native English speakers, as otherwise we would probably only be able to measure the difference between users speaking a native language and those speaking a non-native one. By selecting Dutch students, we could make sure that English was not their first language and in their age range, the students would have a sufficient level of speaking and understanding English as a second language as they had already undergone English language education for a number of years in school. In addition, none of the students or teachers expressed any difficulty with the level of English employed in the ROILA curriculum. Therefore, we could assume that the 50 ROILA words and their English meanings were comprehendible to the children. We also assumed that older students might be less enthusiastic working with LEGO. In total, we worked with 102 students who were spread across four classes, with each class having between 20 and 30 students.

Due to practical and logistical constraints, we were not able to include in the experiment all the students who had participated in the language-learning phase of the project. We, therefore, made a selection. Instead of using a purely random method of selecting a subset of students, we decided to take other criteria into account. We excluded students who were absent in class, as we could not be certain that their level of ROILA understanding would be sufficient. Students who exhibited a reluctant attitude toward the science class were also excluded after consultation with their teacher. In addition, we wanted to have a nearly equal representation of female and male students. A selection of 35 students were invited to the experiment, and all of them accepted

the invitation. We are aware that this selection is not random and that it might have introduced a certain bias. However, through this method, we were able to exclude other biases that are not directly related to ROILA.

To determine how strong a possible bias of our selection could be, we ran a between-subjects analysis of variance (ANOVA), where the between subject factor was whether a student was selected for the experiment or not. The dependent variable was the ROILA exam score for that student. Students who were selected for the experiment did not significantly (F (1, 95) = 2.8, p = 0.10) receive better scores in the exam. We can, therefore, exclude the possibility that we had accidentally selected ROILA specialists.

Measurements

The setup of the experiment was a within-subjects design in which the within factor was language (ROILA or English). The measurements were the Subjective Assessment of Speech System Interfaces (SASSI) scores,[28] number of commands, semantic accuracy, sentence accuracy, and word accuracy. The three accuracy measures and the number of commands were obtained with the help of video and audio recordings of the experiment. In addition, we recorded some measurements to control for possible biases.

The main measurements are defined in the following sections, with further details in section 5.2.

SASSI score. The SASSI questionnaire is a standardized questionnaire used in Speech Interaction analysis, and we used the Dutch version of the SASSI questionnaire.[29] The questionnaire comprises 34 items that are clustered into the following factors:

System Response Accuracy: Did the system recognize the user input correctly, and, hence, what did the user intend and expect?

Likeability: Was the system appreciated, and did the system induce some positive affect in the user?

Cognitive Demand: An indication of the perceived level of mental effort required to use the system and the feelings arising from this effort.

Habitability: The items in this factor deal with aspects such as, does the user know what to say and what the system is doing. High habitability would mean that there is a good match between the users' conceptual model of the system and the actual system.

Annoyance: General irritation and frustration pertaining to interactions with the system.

Speed: Performance of the system in terms of time taken to carry out a response.

Number of commands. The number of commands is the count of commands uttered by the participants per condition. We excluded commands that were not related to the interaction context. Some children, for example, started talking to the facilitator, or talked in Dutch. Since the microphone was continuously on, the system mistakenly thought a command was given.

Semantic accuracy. Semantic accuracy denotes whether the system interpreted what was meant correctly despite not having processed the command with 100% recognition accuracy. An example could be that when the participants said "Turn Left," the system only recognized the utterance as "Left" and instructed the robot correctly to turn left. This variable is also stated as concept accuracy in literature.[30]

Sentence accuracy. Sentence accuracy states how many sentences were recognized 100% accurately.

Word accuracy. Word accuracy is a standard accuracy measurement that is used for speech recognition[30] based on the Levensthein distance. The reverse of word accuracy is also known as word error rate (WER), which is another commonly used measure in Speech Recognition where Word Accuracy (%) = 100 − WER (%). WER is the number of word (byte) operations (insertions, substitutions, and deletions) required to convert the recognized sentence to the reference sentence, or the sentence that is uttered by the user.

Procedure

Before the experiments, the parents of the students signed an informed consent form for their children that allowed us to include the children in the study. All students were provided with a handout a few days before the experiment. We not only asked the children to refresh their knowledge of the vocabulary and grammar in the handout, but we also made the handout available during the experiment. The handout listed the available vocabulary in both, ROILA and English, which consisted of 25 words each. We designed an interaction scenario in which we could incorporate as many words from the vocabulary as possible. This gaming scenario is described in detail later on in this section.

After welcoming the participant, the experimenter seated the participant in a quiet room. The experimenter explained the game that the participant was expected to play with the robot. Afterward, the experimenter answered any question posed by the participant before taking a seat at the back of the room. Half of the participants would play a game first in ROILA and then in English and the other half in the reverse order. Each game session lasted for 10 min. The LEGO robot and the interaction required for the game were simple extensions from the class lessons. The objective of the game was to put as many balls as possible in four colored goals (Red, Green, Blue, and Yellow), which were spread in a room (see Fig. 3). The children would first have to navigate the robot to rest on a specific predetermined colored circle and once the robot had sensed the color via a color sensor, the children could shoot at the goal of the same color. The robot would call out the color in question at the start of every game round. For example, when playing the game in English, the robot would first say Toward Red and then the children would give navigation instructions so that the robot would move toward the red circle. Once the robot would stop on the red circle, the children could say See Red to let the robot sense the color. On correct sensing, the children were allowed to shoot, by saying Drop Ball. In some situations, they would first have to orient the robot toward the red goal by giving further navigation instructions before shooting. The process would be repeated when the robot would give another color recommendation, for example, Toward Green. The same scenario was repeated when the children would play the game in ROILA. A list of commands that could be used by the children and possible responses from the robot are summarized in Table 2.

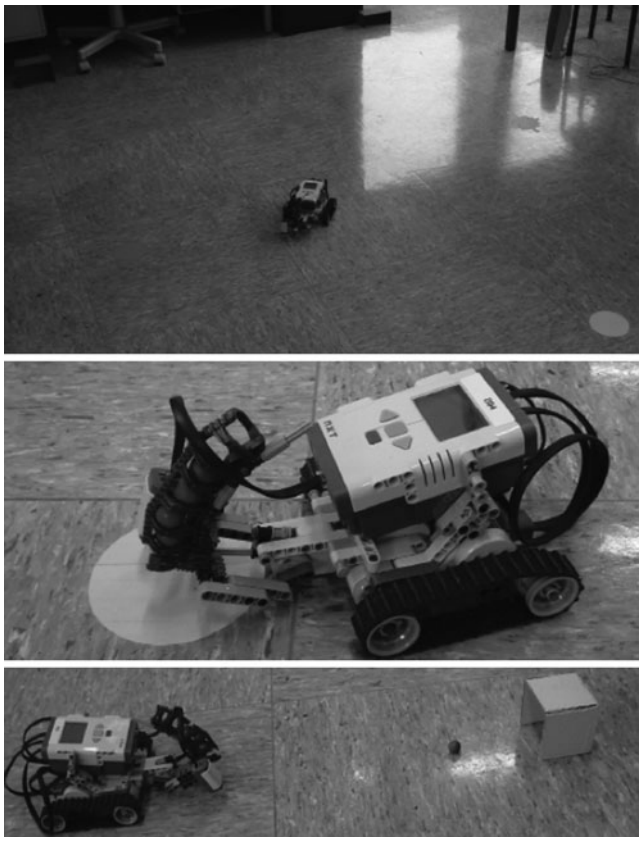

FIG. 3. Game setup.

After each session, the participants were asked to fill in the SASSI questionnaire.

Setup

The children were seated in front of a table on which the microphone was placed. The robot was placed one meter away from the table on the ground. A laptop was placed on a second table right next to the first table. The screen was turned away from the participant. It showed the recognized words. A video camera was placed behind the participant and pointed at the laptop screen. This allowed us to record the recognized words in relation to what the participant had said. In addition, we automatically saved the recognized words in a log file on the laptop.

Results

We excluded four participants from our analyses due to a software problem that resulted in shortening the experimental session to below 10 min.

Before starting the main analysis, we executed certain pretests to assess possible biases and to assess our measurement instruments. A reliability analysis within each of the six factors of the SASSI questionnaire resulted in a Cronbach's Alpha of above 0.7 for each of the six factors. Cronbach Alpha is a reliability coefficient that states the extent to which certain variables measure the same thing. Typically, a value of greater than 0.7 indicates a reasonable agreement.

Participants who achieved a higher game score in one language condition could subjectively rank the language more positively in the SASSI questionnaire compared with participants who scored lower. We, therefore, analyzed the game performance of the participants. We performed a paired sample *t*-test for the 31 participants. Participants in the ROILA condition shot on average more balls (ROILA: 1.00, English 0.79) and scored on average more goals (ROILA: 0.66, English 0.21), but neither difference was statistically significant. We, therefore, assumed that the participants' game performance did not significantly influence the SASSI scores.

SASSI scores

We performed a repeated-measure ANOVA in which the language (ROILA or English) was the independent variable, and the six SASSI factors were the dependent variables.

ROILA was perceived on average more positively on all six factors of the SASSI questionnaire (see Table 3). Three factors, namely System Response Accuracy, Annoyance, and Speed, were also significant in favor of ROILA. Likeability was approaching significance.

Table 2. Robot Interaction Language and English Interaction During the Game

Commands that could be used in the game		*What the robot could say*	
ROILA	*English*	*ROILA*	*English*
kanek koloke	Go forward	kufim tifeke	Toward red
kanek nole	Go backward	kufim kutuju	Toward blue
botama webufo	Turn left	kufim koleti	Toward green
botama besati	Turn right	kufim wipoba	Toward yellow
botama nole	Turn back	wopa	Good
bobuja	Run	wekepa	Error
kanek jimeja	Go quickly	bemeko wekepo	Wrong color
kanek kipupi	Go slowly	tobop jinolu	Put ball
buse kanek	Stop		
biwu wekepo	What color		
jasupa jinolu	Drop ball		
make tifeke	See red		
make kutuju	See blue		
make koleti	See green		
make wipoba	See yellow		

Table 3. Analysis of Variance and Mean-Standard Deviation Table for SASSI Main Effects

	Language type		*ROILA*		*English*	
Factorname	*F(1,30)*	*p*	*Mean*	*st.dev*	*Mean*	*st.dev*
Sys. Resp. Acc.	4.88	0.04	2.72	0.62	2.44	0.60
Likeability	3.56	0.07	3.36	0.45	3.10	0.70
Cogn. Demand	0.01	0.91	3.12	0.63	3.10	0.84
Annoyance	4.94	0.03	3.21	0.77	2.87	0.62
Habitability	0.29	0.59	3.04	0.86	2.95	0.79
Speed	10.44	<0.01	3.63	0.92	3.24	0.96

Recognition accuracy

We used the video recordings in combination with the log files to transcribe the interaction. We coded what the participant said, what the system recognized, and what action the robot took. This allowed us to calculate the different accuracy measurements. The microphone was permanently on and recorded whatever was said, including irrelevant utterances. We excluded utterances in any language other than the prescribed language for the given condition. We also excluded utterances directed toward the experimenter. Neither type of utterance is of relevance to the calculation of accuracy. The number of commands (NrCo) was measured as a count, while the other three accuracy measurements (semantic accuracy, sentence accuracy, and word accuracy) were calculated as percentages. Semantic accuracy was calculated as

$$SemAc = \frac{\text{Number Of Correct Actions}}{\text{Number Of Commands}} \quad (1)$$

Sentence accuracy was calculated under the assumption that a command equals a sentence:

$$SenAc = \frac{\text{Number Of Correct Sentences}}{\text{Number Of Commands}} \quad (2)$$

Word accuracy was calculated as

$$WoAC = 1 - WER \quad (3)$$

where the WER is the number of operations required to convert the recognized sentence into the reference sentence, or the sentence that was uttered by the user. This computation is, in fact, the Levenshtein distance. The operations can be of three types, namely insertions, deletions, or substitutions, and they all have the same cost. We state the following equation to compute WER, where S = substitutions, D = deletions, I = insertions, and N = number of words in the reference sentence.

$$WER = \frac{(S + D + 1)}{N} \quad (4)$$

During the transcription process, we discovered that due to a camera failure some footage from the sessions were missing for seven participants. We excluded these participants from the analysis on recognition accuracy measurements, bringing down the number of cases to 24.

We performed repeated-measures ANOVA in which the language (ROILA or English) was the independent variable. The number of commands and the three accuracy scores were the dependent variables.

The results in Table 4 show that the sentence accuracy and word accuracy in the ROILA condition were significantly above the accuracies in the English condition. There is no significant difference between the conditions with regard to semantic accuracy, but participants used significantly more commands in the English condition than in the ROILA condition.

Evaluating the efficiency and learnability of ROILA

To understand whether it is worth learning ROILA, we need to put the gained advantage of improved recognition accuracy in relation to the time invested into learning ROILA. We will now try to make an informed estimation about the possible efficiency of ROILA. First, we are going to estimate how much time ROILA would save a user before estimating the learning effort.

A literature review revealed that correcting one speech recognition error in a dictation task of 100 words would take approximately 3.5 sec.[31,32] This value might change slightly for different languages and speech recognizers. Ideally, we would have liked to confirm this value in the context of ROILA dictation tasks, but due to time constraints, we had to accept this estimation as is. Instead, we will try to integrate this estimation of 3.5 sec into the application context of the interaction used in our experiment.

The results of our experiment showed that ROILA had an 18.9% better word accuracy than English (see last row of Table 4), that is, on average, 18.9% more words were interpreted correctly in ROILA as compared with English. In a dictation task of 100 words, a speaker would, therefore, save $3.5 \times 18.9 = 66.2$ sec (approximately 1.1 min) when speaking ROILA instead of English. Based on the information in the

Table 4. Means and Analysis of Variance Table for Recognition Accuracy Analysis

			ROILA		*English*	
Measure	*F(1,23)*	*p*	*Mean*	*st.dev*	*Mean*	*st.dev*
NrCo	9.13	<0.01	74.42	15.84	82.67	15.93
SemAc(%)	1.03	0.32	69.41	12.26	65.07	19.61
SenAc(%)	8.65	<0.01	54.72	14.47	42.55	21.36
WoAc(%)	20.18	<0.01	63.58	15.14	44.74	26.76

lesson booklet from the children, we know that on average, the 24 participants considered in the recognition accuracy analysis spent 65.4 min to learn ROILA at home. In addition, they spent 300 min in the ROILA lessons at school, which brings us to a total of 365.4 min.

It is also essential to extrapolate the benefits and the costs of learning ROILA within the context of the interaction in our experiment. From our log files, we know that on average, the participants uttered 74.4 commands (see Table 2) and that one command consisted on an average of 1.93 ROILA words. During the 10 min of interacting with the robot, the participants, therefore, spoke on average $74.4 \times 1.93 = 143.21$ ROILA words. For every 100 words spoken, we estimated that ROILA would save 1.1 min, bringing the total benefit to1.58 min for 10 min of interaction. The children had already invested 365.4 min into learning ROILA; so, it would take them $365.4 \times 10 = 2312$ min (38:36 h) to break even.

Discussion

The higher SASSI ratings for ROILA in comparison to English lead us to believe that the Dutch high school students in our study perceived ROILA to be more user friendly than English. Three SASSI factors were rated as being significantly higher (System Response Accuracy, Annoyance, and Speed), and Likeability was approaching significance. These higher scores might be explained by the ROILA's 18.9% higher word recognition accuracy. However, we have to acknowledge that the overall word recognition accuracy was not particularly high (63% ROILA, 45% English). Trained speech recognizers in ideal conditions are able to perform better, but without any training, such recognition accuracy is expected from Sphinx on English test data.[33] Still, an improvement of 18.9% recognition accuracy of ROILA compared with English is encouraging.

The recognition accuracy results were more in favor of ROILA than the SASSI scores. This may be explained by the fact that the semantic accuracy does not differ significantly between ROILA and English. The robot would often execute the right action, although not all words were recognized correctly. The participants could not be aware of this difference, as they could only observe the robot's behavior. Hence, they might have rated the two languages less differently in the SASSI questionnaires. In essence, it would appear that the dialog manager of our system has been able to make up for the recognition errors to some degree. Having a robust dialog manager is one technique that is used for undermining (although to a small extent) the recognition errors of the speech recognizer.[34]

We also observed that the participants uttered more commands in English than in ROILA. This could be due to the fact that they were likely to be more proficient in English than in ROILA, but it could also be due the low recognition accuracy of English, which would have forced them to repeat their commands frequently. It is well accepted in speech recognition research that the probability of a further error doubles after a first recognition error.[35] We also did not observe the students having any difficulty in pronouncing English words, due to lack of proficiency in English, for example. The data available do not allow us draw final conclusions on this issue. In any case, we have to acknowledge that the participants were not native English speakers and their Dutch accent contributed to the low recognition accuracy for English and ROILA. Similar problems are expected for any untrained recognition engine and even the different regional English accents, such as Scottish or Australian, can cause considerable disruptions.[36] To repeat our argument based on what has been stated earlier, if we had used native speakers, then we would have only tested a native language against a second language. The results of such a study are predictable and would be less useful for evaluating ROILA. Our assumption was that using the standard untrained North American acoustic model of Sphinx would give both English and ROILA a fair base for comparison.

On a methodological note, we believe that our setup can be useful for other researchers who wish to analyze the recognition accuracies of their robot interaction systems. We have presented a simple yet effective setup methodology that can qualitatively and quantitatively allow for the comparison of distinct robot interaction systems. As highlighted earlier in our article, previous evaluations of speech-enabled robotic systems have neither involved a substantial sample size, nor have they tried to address real-life scenarios. In our opinion, collecting rich and contextual data via psychological methodologies is the way to go with regard to studying the state-of-art human–robot interaction. Empirical and experimental research is gaining popularity and ascendancy as far as the evaluation of human–robot interaction is concerned[37]; however, there are not many large-scale evaluations to report when it comes to studying speech-enabled robotic systems. In summary, the focus of HRI evaluations has not been primarily on the modality of speech.

As far as we are aware, the ROILA evaluation would also be a first relatively large-scale attempt of evaluating the learnability of artificial languages in the context of using them to talk to machines or robots. We hope our efforts would encourage other language developers to determine a benchmark that could be used for evaluating artificial languages as a whole and also in terms of their learnability. Linguistic research has used artificial language learning to study how humans learn languages in general, that is, more specifically what are the cognitive and biological processes involved in natural language learning; for example, see Folia et al. (2010).[38] What was missing was a methodology to evaluate artificial languages and a framework that can establish whether a particular artificial language is easy to learn.

Future Research

We would like to fully embed ROILA into the LEGO Mindstorms NXT. At this point, the speech recognition and speech synthesis component is running on a laptop computer. The technical constraints of the 2.0 version of the NXT that made this setup necessary. However, the further development of the NXT platform in combination with the availability of the Pocket Sphinx engine might make it possible to embed ROILA directly in the future. Once ROILA would become a part of Lejos or even the original LEGO firmware, we might be able to create a critical mass of ROILA empowered devices. Integrating ROILA into more advanced robotic platforms, such as NAO, would be possible nowadays; however, as we point out in due course, standard installers need to be developed. We would like to invite all robotic developers to consider integrating ROILA, which has been released under the

Creative Commons Attribution-Share Alike 3.0 Unported License. However, ROILA is not even limited to robots. Any speech system could take advantage of it, such as mobile phones or car navigation systems.

We are also still interested in what impact ROILA might have on the semantic level of speech understanding. The regular grammar and the elimination of homophones could improve general speech understanding. Artificial intelligence systems that use the results from the speech recognizer as their starting point for interpreting the meaning of an utterance could also benefit from ROILA.

Limitations

The results of our study need to be considered in the light of certain limitations. We did not employ the full vocabulary of ROILA, as this would have exceeded the ability of the students. Within 3 weeks, it was only possible to teach them 50 out of the nearly 1000 words. To guarantee a fair comparison, we also limited the English vocabulary in Sphinx to the same words that had been used in ROILA. The vocabulary of ROILA is based on Simple English and, hence, is not optimized for the use in human–robot interactions. As a matter of fact, simple English did not even include a word for robot. During the process of developing the ROILA vocabulary, we extended the vocabulary with such necessary words.

We also have to acknowledge that we did not use a fully random method of selecting the participants of the study. We tried to counter balance other biases, such as the effect of puberty in our selection. Our analysis showed that we did not accidentally select only the talented ROILA speakers and, hence, we tend to believe that a possible selection bias does not fully negate the validity of this study. We also still need to take into consideration that the speech recognition for a system in which the speaker can talk in his or her native language in combination with a trained recognition engine is likely to produce better results than the setup discussed in this article. An additional experiment would be necessary to compare ROILA against such an optimized system. For such a comparative study, we would need to have fully trained ROILA speakers and an acoustic model for ROILA.

The speech synthesis module of our system was well received in general. However, we acknowledge that we could have further coded our videos to determine in how many instances the children did not understand what the robot said to them. For example, this could have been accomplished by counting the number of times the children said "biwu" when playing in ROILA or "what" when playing in English. In conclusion, none of the children reported any difficulties in understanding the robot.

Disruptive Potential, Barriers to Overcome, and Conclusions

As for the efficiency of ROILA, we conclude that the benefits of using ROILA are in its continuous use. It would not pay off to use ROILA for an application that users only work with once a week. However, for a scenario in which users have a robot at home and interact with it several times a day for years to come, it could be beneficial to use ROILA. The deciding factor is to create a critical mass of interaction time. If there are more robots or computer systems that understand ROILA, the more it will pay off to speak it. Nowadays, it does not yet seem to be of great use, but our hope is that the benefits of ROILA will encourage robot enthusiasts and software engineers to integrate the freely available ROILA system into their applications. We fully acknowledge that the ROILA setup at the moment is a work in progress and requires some prerequisite labor to have it up and running. Ideally, we would like it to be completely generalizable to any platform so that robot developers could migrate the ROILA setup to their own robots. This is one of our future goals, that is, to design ROILA installers (an API for example), allowing developers to enable ROILA interaction with different robots (such as the Nao robot) in a short time. This would ultimately mean that ROILA does not need to be restricted to LEGO Mindstorms. By providing researchers and developers with a "plug and play" ROILA setup, in an instant, the potential applications of ROILA interaction grow exponentially. This will also enable us and fellow researchers to evaluate ROILA in other contexts and to develop it further. In addition, ROILA does not need to be restricted to robots; we foresee its use in not only household robots (such as the Roomba vacuum cleaner) but also different kinds of behavioral products. Our future goal of developing a platform-independent ROILA setup for robots is in parallel to our efforts in developing a ROILA "setup" for humans as well. We have accomplished this by writing a ROILA book[39] that details the grammatical rules and vocabulary of ROILA and is targeted at anyone who wishes to learn the ROILA language. The book will help in training prospective ROILA speakers.

Of course, we have to admit that we are somewhat idealistic in proposing yet another artificial language. Speech technology researchers have invested a lot of time and effort in improving algorithms that improve recognition accuracy; however, we present a reverse approach that is a bit provocative, to say the least. Motivating human users to learn ROILA is also one of the key challenges. However, we wish to believe that once people (and society in general) acknowledge the long-term benefits of a language that is geared toward facilitating human–robot interactions, they will be slightly more compelled to explore and learn ROILA. Arika Okrent presents a compelling history of artificial languages,[40] and we cannot exclude the fact that ROILA will become another failed attempt. However, our hope is that ROILA will be able to create a critical mass of speakers not through the availability of human speakers but through the presence of machine speakers. With only one update of Microsoft Windows, millions of computers could become native ROILA speakers, providing users with the implicit need and motivation to learn and use ROILA.

Acknowledgments

The authors would like to thank all the teaching staff and the children at the Christiaan Huygens College Eindhoven. They would also like to appreciate the support of LEGO Mindstorms.

Author Disclosure Statement

No competing financial interests exist.

References

1. IFR Statistical Department. World robotics survey. Technical report, 2008. www.IFR.org
2. Prasad R, Saruwatari H, Shikano K. Robots that can hear, understand and talk. Adv Robotics. 2004;18:533–564.
3. Shneiderman B. The limits of speech recognition. Commun ACM. 2000;43:63–65.
4. Forsberg M. Why is speech recognition difficult? Technical Report from Department of Computing Science Chalmers Univ of Technology, Sweden, 2003.
5. Large JA. The Artificial Language Movement. Oxford, UK: Blackwell Publishers, 1987.
6. Janton P. Esperanto: Language, Literature, and Community. Albany, NY: State University of New York Press, 1993.
7. Rosenfeld R, Olsen D, Rudnicky A. Universal speech interfaces. Interactions. 8:34–44, 2001.
8. Tomko S, Rosenfeld R. Speech graffiti vs. natural language: Assessing the user experience. In *Proceedings of HLT/NAACL*, Boston, MA, 2004.
9. Harris TK, Rosenfeld RA. Universal Speech Interface for Appliances. Technical report, Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, 2004.
10. Schulz R, Glover A, Wyeth, G, et al. Robots, communication, and language: An overview of the lingodroid project. In *Australasian Conference on Robotics and Automation (ACRA)*, Brisbane, Australia, 2010.
11. Seabra Lopes L, Chauhan A. How many words can my robot learn? An approach and experiments with one-class learning. Interact Stud. 2007;8:53–81.
12. Bos J, Oka T. A spoken language interface with a mobile robot. Artif Life Robotics. 2007;11:42–47.
13. Oka T, Abe T, Sugita K, et al. Runa: a multimodal command language for home robot users. Artif Life Robotics. 2009; 13:455–459.
14. Perzanowski D, Schultz AC, Adams W, et al. Building a multimodal human-robot interface. Intelligent Syst IEEE. 2001;16:16–21.
15. Fransen B, Morariu V, Martinson E, et al. Using vision, acoustics, and natural language for disambiguation. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 73–80, New York, NY, 2007. ACM.
16. Ishi CT, Liu C, Ishiguro H, et al. Head motions during dialogue speech and nod timing control in humanoid robots. In *Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction*, HRI'10, pages 293–300, New York, NY, 2010. ACM.
17. Hanafiah ZM, Yamazaki C, Nakamura A, et al. Human-robot speech interface understanding inexplicit utterances using vision. In *CHI'04 extended abstracts on human factors in computing systems* (p. 1321–1324). Vienna, Austria, 2004. ACM.
18. Marin, R., P. Vila, et al. (2002). Automatic speech recognition to teleoperate a robot via Web, International Conference on Intelligent Robots and Systems IEEE. Lausanne, Switzerland.
19. Valin JM, Yamamoto S, et al. "Robust recognition of simultaneous speech by a mobile robot." IEEE Trans Robotics. 2007;23:742–752.
20. Arsoy E, Arslan LM. A universal human machine speech interaction language for robust speech recognition applications. In: Sojka et al., ed. International Conference on Text, Speech and Dialogue. Berlin: Springer, 2004, pp. 261–267.
21. Hinde S, Belrose G. Computer Pidgin Language: A New Language to Talk to Your Computer? Technical Report. Bristol, UK: Hewlett-Packard Laboratories, 2001.
22. Mubin O, Bartneck, C, Feijs L. Towards the design and evaluation of roila: a speech recognition friendly artificial language. In: H. Loftsson (Eds.), Advances in Natural Language Processing (IceTAL). Berlin: Springer, 2010, pp. 250–256.
23. MacLean A, Young RM, Bellotti VME, et al. Questions, options, and criteria: Elements of design space analysis. Hum Comput Interact. 1991;6:201–250.
24. Lamere P, Kwok P, Gouva E, et al. The cmu sphinx-4 speech recognition system. Proc. of IEEE Conference on Acoustics, Speech and Signal Processing, Hong Kong, 2003.
25. Liu C, Melnar L. Training acoustic models with speech data from different languages. In: ISCA Tutorial and Research Workshop (ITRW) on Multilingual Speech and Language Processing. Lisbon, Portugal: ISCA, 2005.
26. The Centre for Speech Technology and Research. Festival, 2008. www.cstr.ed.ac.uk/projects/festival
27. LEGO. Lego mindstorms nxt, 2010. mindstorms.lego.com
28. Hone KS, Graham R. Towards a tool for the subjective assessment of speech system interfaces (sassi). Nat Lang Eng. 2001;6:287–303.
29. Turnhout K. *Socially Aware Conversational Agents*. PhD thesis, Eindhoven University of Technology, Eindhoven, Netherlands, 2007.
30. Boros M, Eckert W, Gallwitz F, et al. Towards understanding spontaneous speech: Word accuracy vs. concept accuracy. In *Proceedings of the Fourth International Conference on Spoken Language Processing (ICSLP 96)*, Philadelphia, PA, pp. 1009–1012, 1996.
31. VoCollect. Speech recognition technology choices for the warehouse. A VoCollect White Paper, Intermec, IP Corp., 2010.
32. Wald M, Bell JM Boulain P, et al. Correcting automatic speech recognition captioning errors in real time. Int J Speech Tech. 2007;10:1–15.
33. Samudravijaya K, Barot M. A comparison of public-domain software tools for speech recognition. In: Workshop on spoken language processing (ISCA). Mumbai, India: ISCA, pp. 125–131, 2003.
34. Lee C, Soong F, Paliwal K. Automatic Speech And Speaker Recognition: Advanced Topics. Boston, MA: Kluwer Academic Pub, 1996.
35. Oviatt S, MacEachern M, Levow G-A. Predicting hyperarticulate speech during human-computer error resolution. Int J Speech Commun. 1998;24:87–110.
36. Huang C, Chen T, Chang E. Accent issues in large vocabulary continuous speech recognition. Int J Speech Tech 2004;7:141–153.
37. Bethel CL, Murphy RR. "Review of Human Studies Methods in HRI and Recommendations." Int J Soc Robotics. 2010;2:347–359.
38. Folia V, Udden J, et al. "Artificial language learning in adults and children." Lang Learn. 2010;60:188–220.
39. Stedman A, Bartneck C, Sutherland D. Learning ROILA. CreateSpace, 2011. ISBN 978-1466494978
40. Okrent A. *In* The Land of Invented Languages: Esperanto Rock Stars, Klingon Poets, Loglan Lovers, and the Mad Dreamers Who Tried to Build a Perfect Language. Spiegel and Grau, New York, 1st edition, 2009.

Address correspondence to:
Omar Mubin
School of Computing, Engineering and Mathematics
Locked Bag 1797
University of Western Sydney (UWS)
Penrith NSW 2751
Australia

E-mail: omar.mubin@gmail.com