Project R-CASTLE: Robotic-Cognitive Adaptive System for Teaching and LEarning

Daniel Tozadore, Adam Pinto, João Valentini, Marcos Camargo, Rodrigo Zavarizz, Victor Rodrigues,
Fernando Vedrameto, and Roseli Romero

Abstract—**Robots are already present in people's lives, for example, as receptionists, caretakers and educational tools. In human-robot interaction, social behavior is not only expected but often associated with users' confidence. Despite the large amount of research, robot adaptation and the gap between the system and non-programming interaction designers still need to be studied. This paper presents a cognitive architecture applied in a humanoid robot that aims to offer a framework capable of being programed for controlling the robot's resources, addressing previously known and new contents in interactive educational activities. Furthermore, this system provides objective measures of the user's attention along the activity - and after it as a report - and adapt the robot's behavior according to these measures of users' engagement. Functions of visual classification, speech processing, autonomous web-search for new contents and attention detectors were tested and analyzed separately. This approach shows effectiveness in basic and medium condition levels from a set of sceneries for each module.**

Index Terms—**Robotic Cognitive Architecture, Adaptive Systems, Educational Robotics, HRI**

I. Introduction

In Human-Robot Interaction (HRI), it is notorious that more adaptive techniques and human-like communication enhances the users' experience with robots. Mainly, in educational activities, enjoyable interactions for the students can increase some factors that are strongly related to learning. When a robot plays a tutor role, specially with young children, it is expected some behaviors like good response time, deep knowledge and accomplishment of the activities without mistakes. This set of behavior is often associated to the users' confidence in a new system [1].

For this reason, the majority of these studies are being made with the Wizard-of-Oz (WoZ) technique [2], which consists in someone controlling the robot without the user's knowledge. Studies to develop robust systems that autonomously interact with humans (in this case, students) have been increased recently, as well as the research to investigate their application impact in the final tasks [3].

However, the robots usage in HRI experiments by itself may produce contrary results from those initially expected. The success in such applications depends on a series of variables to be set up according to the users' expectations, culture, age and also of the activity nature. Studies in the robot adaptation have also shown a significant importance in this applications.

Considering the raised points so far, this paper presents the objectives and ongoing studies for a Robotic Cognitive Architecture to be used as framework in HRI designs: the Robotic-Cognitive Adaptive System for Teaching and LEarning (R-CASTLE). R-CASTLE aims to deliver a new educational tool that allows any kind of designer (the programing and non-programing persons that want to apply this methodology) to easily plan interactive activities with electronic devices, in which in this case is applied in a humanoid robot, proving an autonomous and natural communication with the robot and adaptive skills for short and long-term interaction.

Natural communication of the robot is guaranteed through modules that perform audio-visual processing and robot's gestures manipulation, while the robot's adaptation behavior relies on multiple sensors and algorithms to collect and analyze the users' perceptions. All the variables configuration from this project are result of several technical and interactive (user-centered) studies, focused in specifics issues that, combined, produce the whole scenario of this research problem. Tests with all modules integrated or tests with users were not performed in this study.

This paper is organized as it follows. In Section II, we discuss the background of Educational HRI studies. Section III presents the project's technical scheme. Partial results are discussed in Section IV and already performed user studies are in Section V . Finally, in Section VI, conclusions and future works are presented.

II. Background

Studies to develop robust systems that autonomously interact with humans - which are the users of such systems - have been increased recently, as well as research to investigate their application impact in the final tasks [3].

Conversely, the search for autonomous systems has a highly increased in last decades, followed by the concerns with this type of interaction design [4]. These studies go from specific functionalities to fully automation degrees and, in the majority of them, the findings lead to keep working in this direction [5], [6].

Considering technical solutions for the automation, Cognitive Architectures has been shown itself as an acceptable option for more generic and complex automation design in robotics. A Cognitive Architecture consists of an intelligent system that benefits from memory, learning and actions to solve generic problems. It constantly takes the environmental parameters to process and update internal states, learning

The authors are with the Department of Computer Science, Institute of Mathematics and Computer Sciences, University of São Paulo, Brazil.

from past experiences to deal with incoming new problems. Basically, an Architecture Cognition is an intelligent system that disposes of memory, learning and decision for generic solutions [7]. Classical solutions, as the Soar [8] and ACT-R [9] architectures, are constantly employed to solve generic problems. The resulting systems in studies with these architectures generate good generalizations in the tasks as social robots [10], [11]. However, for specific and focused studies, this generalization may not be the best solution. Problems that are harder to have total control with classical architectures, such as encompassing the coordination of multiple sensory and motor modalities for the robot, the timing of proactive and reactive actions and the recognition of interacting human states (cognitive, affective, physical, etc) are pointed as the reason of the rise of new proposed architectures [12].

It is worthy to reiterate that the interactions may vary according to the users. For example, adults and children interactions with the robot are potentially very different due to childrens neurophysical and mental development being ongoing. Furthermore, the view of cognition is being extended, suggesting that cognition is no longer a exclusive domain of the individual, but the product of a fine-grained interaction between agents of any nature, be they people or machines [13]. This is one of the main reasons to perform several studies to understand the users profile before and while developing such systems.

Coming to HRI in children educational domain using cognitive systems, one example of autonomous solution is the EASEL [14] Project. The EASEL (”Expressive Agents for Symbiotic Education and Learning”) aims to deliver a new set of Robotic Based Tutoring Solutions: a Synthetic Tutoring Assistant. It considers the utilization of different subsystems as modules to automate the whole system. Speech recognition in noisy environments [15] and audiovisual scene analyzer [16]; dialog and behavior planning [17] and dialog handler and behavior adapter [18]; content and behavior generation using a Zeno [19] robot and a FACE [20] robot compose the EASEL modules. The user-centered studies involve children's perceptions of the robot [21], how children play with different robot shapes [22], how different plays change the type of interaction [23], how children adopt different behaviors in symbiotic cooperation tasks [24] and how to model personalities in robots for symbiotic interaction in the educational context [25]. However, these proposal does not allow people outside the project to design the interactions, neither provide accurate reports about the users accomplishment after the interactions.

III. R-CASTLE Technical Scheme

The proposed architecture aims to offer a framework capable to be programed for controlling the robot's resources, addressing previously known and new learning contents, in interactive educational activities. Summarizing, there are some open-source programed behaviors (as object and face recognition, web research, speech processing) and an interface where a non-programing designer can use all robots features to create a new activity A NAO robot was used as device in our tests due to making part of our lab tools but the architecture is easily adapted to control other devices as well.

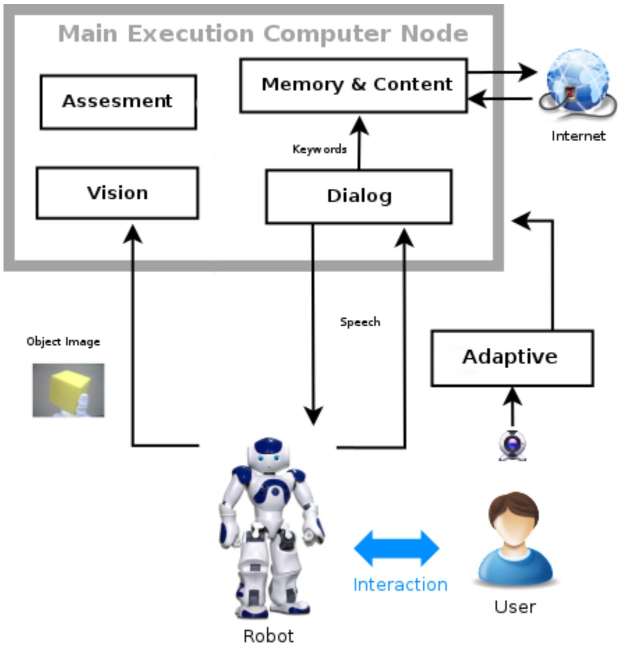


Fig. 1. An architecture overview summarizing the modules' operation.

The R-CASTLE is divided in modules according their functionalities: Vision, Dialog, Memory & Content and Assessment, as illustrated in Fig 1. The following subsections detailed the modules.

A. Vision

The visual system goal is to recognize and classify along the interaction the specific objects or images that were previously trained in the system. In this case, the input was the NAO's frontal camera and the images were collected and classified in specific moments. These techniques are from machine learning field. Several combinations were tested for geometric figures approach. This content was chosen due to be solid objects easy to be handled by children, which was a good start for initial tests.

Primarily, the vision functionality was implemented with both a Top-Down and Bottom-Up attention visual system to recognize 2D figures [26]. A 3^{th} Generation Neural Network, the Local Excitatory Global Inhibitory Oscillatory Network (LEGION) [27], was used to object segmentation, a Multi-Layer Perceptron (MLP) network provided a high level data classifier for objects recognition, a Self-Organized Map (SOM) to create a object-saliency map and an object selection module, which highlights the most important object in the image. The resulting system autonomously classified basic figures, such as triangles, squares, rectangles and parallelograms, playing with the children and teaching some concepts of flat geometry. With a small configuration, it was also possible to classify the same basic figures, but with overlapping [28].

Aiming to adapt this functionality to 3D recognition, the visual system was changed to a detection and classification method based on objects texture. The system was the VOCUS2 [29] algorithm for segmentation and background noise extraction, that uses the *SURF* method to extract the feature descriptors and *bag-of-words* method in their histogram to create a feature dictionary, and, finally, multiple SVM with these descriptors to annotate, or predict new input images according to the feature dictionary.

Currently, the classification method being tested is an ensemble solution which takes into consideration the prediction of three super-visioned machine learning methods: K-Nearst Neighbors (KNN), MultiLayer Perceptron (MLP) and Support Vector Machine (SVM). In each classifier, two approaches were performed: one using the image histogram and the other using raw pixels as input. It's known that, in some cases, these techniques by themselves can solve this problem, but as a more powerful alternative the vision system was implemented with an ensemble classifier technique [30], since the training time is considered low and the accuracy remained the same or increased in all scenarios.

B. Dialogue

The Dialogue Module's objective is to allow the interaction to be performed in a more human-like communication manner: the verbal conversation. It is designed to provide more intelligent two-way communication between the robot and the user. It is intelligent because the early version was only able to receive raw sentences from the users and send fixed speech to the robot's voice synthesis, both with a predefined dialogue. In this version the architecture analyzes the received sentences to have a better interpretation from a group of correlated sentences.

Regarding the methods, the system is a composition of three core functions: the speech recognition and text to speech functions; a combination of Natural Language Processing (NLP) techniques for basic sentence interpretation and a simple doubt/negation/affirmation analysis. The *Google's API* is used for speech recognition while the text to speech functions were provided as a native tool from the NAO robot. The Natural Language Processing techniques came as part of the *TextBlob* package. And also more python packages for NLP are being employed in R-CASTLE. Finally, the affirmation/doubt/negation analysis is based on key words defined by the designer that can indicate one of these statements. The check order is doubt, negation and affirmation to reduce the misinterpretations that some expression types may cause in the sentences. The keywords for each statement are held in specific files and can be changed before the interactions.

As an initial phase, the system classifies the sentences expressed by the users using *Google's API* in order to find keywords that can suggest affirmations, doubts/questions or negations. In case doubts/questions are found, the Dialog Module sends this information to the Assessment Module to be included in the reports and can interfere in the interaction flow, making the system repeat the exercise or the last sentence spoken. Furthermore, when a new topic is aimed to be learned during the interaction, the Dialog Module have to process the users sentences and send only the sentences' nouns to the Memory & Content Module, that will search for the noun definition in the Wikipedia.

C. Memory & Content

All the content regarding information aimed to address along the interaction is managed by the Memory & Content module. This module's algorithms allow to manually program contents in the system database or search on the Internet for questions along the interaction that were not programed by the designer. This procedure can be requested at any time by the Dialog Module. First, the system searches on its local database, composed of already questions answered. For new questions, it searches on the Wikipedia database through an API and the new knowledge is added to the memory. All the answered questions are preprocessed by the Dialog Module and only the resulting keywords, that are mainly nouns, are used for web search.

Previous research from the same research group [31] showed that young students tend to ask the robot some ”personal” questions. To make the answer more appealing while preprocessing, the Dialog module searches for words like ”you” and ”your”, for instance, ”How old are you?” or ”What is your favorite hobby?” and, if finds it, it adds a personal tag to the resulting keywords of this sentence. The personal questions tag triggers a special database where all the robot's ”personal” question is manually added. Questions that are not found are answered by a default phrase and all the keywords are later manually inserted.

D. Adaptive

An important part of the interaction design is to hold the users attention span and keep them interested as long as possible. Adaptation is the key to achieve this goal in all kind of interaction and its may occurs changing the robot's behavior. The robot can changing its behavior, assuming different roles, such as security guard or just guidance, error or perfect performer, strait or interactive instructor, tutor or learner, and in all of them, the variation produces a different result .

The adaptive module aims to change the robot behavior according to the users' indicators observation. For a better analysis, the indicators were divided in three main groups regarding the users' measures of Attention, Communication and Learning. Worthy to note that these 3 measures are qualitative. Thus, a set of objective measures are required to conclude a consistent qualitative result. The objective measures from each group are summarized in Table I. The groups are shown in the first line (with the corresponding denotation function in parentheses) follow by the respective indicators in their columns. The indicators are: Eye gaze and Posture for the Attention; number of Words spoked by the users and user's Emotions for Communication; and Right/Wrong answer and Time to answer to the proposed exercises. All the objective measures are being modeled as a vector to functions from each group. The functions are: α to Attention, β to Communication and γ to Learning.

TABLE I
OBJECTIVE MEASURES BY GROUP.

Attention (α)	**Communication** (β)	**Learning** (γ)
Eye gaze (Eg)	Number of Words (nW)	Right/Wrong answers (RWa)
Posture (p)	Emotions (Em)	Time to answer (Tta)

These functions will process the corresponding vectors of objective measures and generate outputs according to particular rules from each function. In this way, the resulting functions with the corresponding vectors are represented as $\alpha = (Eg,P)$, $\beta = (nW,Em)$ and $\gamma = (RWa,Tta)$. The functions outputs will have two utilizations: be saved and sent to the Assessment Module to produce late reports to the designers about the interactions with the users and to be used as input to the adaptive behavior function. The adapted robot's behavior, denoted as Ψ, is calculated by the F_{Adp} shown in Equation 1.

$$\Psi(\alpha,\beta,\gamma) = F_{Adp}(w_\alpha * \alpha, w_\beta * \beta, w_\gamma * \gamma) \tag{1}$$

The F_{Adp} is a function to adapt the resulting robot's behavior Ψ, trying to optimize the interaction engagement and leaning rate. The parameters α,β and γ are the group functions outputs and the w_α, w_β and w_γ are the corresponding weights, calculated by some super-visioned machine learning algorithm.

Although the intention is to achieve many behavior degrees, the resulting robot's behavior Ψ so far is a discrete scale between two independent behaviors that goes from 1 to 5, where 1 is considering the final behavior as totally one of the two extremes in the scale and 5 is totally considering the other one. Following the logic, if the result is the number 3, a mixed behavior with half of both scale behaviors is performed by the robot. This model is represented in 2, in which the Adaptive module receives the entries read by the sensors, process the F_{Adp} function according to the respective parameters, producing the behavior Ψ that will be placed in the behavior scale set by the designer.

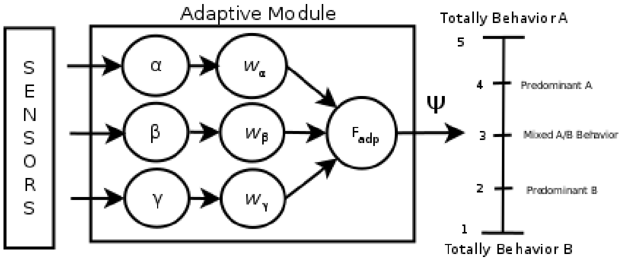


Fig. 2. Adaptive Module representation.

For instance, consider two independent behaviors to be placed in the robot's behavior scale and they are focused in ”Physical Interaction” and ”Content Learning”. In the first one, the robot will take more physical interaction with the student, like inviting him/her to dancing, performing some play or physical exercise, and so on. In the second, the robot will approach the programed content in a very strait manner. As said previously, if the F_{Adp} takes 1 as result the robot will assume the totally ”Physical Interaction” behavior, if the result is 5 it will take the ”Content Learning” behavior and if result is 3, a merged behavior between the conditions. In case of 2, the robot will take a predominant ”Physical Interaction” behavior, interchanging small content approaches along the play, whereas in the 4 point of the scale the robot will assume a predominant ”Content Learning” behavior, interchanging small interactions along the content approaching.

Adapting to a more continuous behavior scale, one solution being study is to group the robots performances by behavior (e.g. all the listed interaction from ”Physical Interaction” behavior are performances of this behavior). Thus, the behaviors in the scale are defined by its performances. The behavior Ψ will use the variable λ to influence the probability of the system to chose a specific performance from the predominant behavior. The λ is the probability of choose a performance from scale behavior A and the performances from behavior B will be chosen with probability $(1-\lambda)$. For example, taking the scale of Fig. 2, where $\Psi(5)$ = Totally Behavior A, the system will choose performances from behavior A with probability 1 and group B with probability 0, it means, only behaviors from group A would be chosen. The same goes to $\Psi(1)$, but with inverted values. To the mixed behavior, the system will choose performances from both groups with 0.5 of probability. To the values between the middle and the extremities, 2 and 4, the system will choose performances with probability of 0.75 from one group and 0.25 from the other, depending of the scale value. Therefore, it is possible to extend the behavior scale to have more than 5 behavior's values, calculated by the value λ. It is known that, the probability variable λ will not assume totally continuous values, being biased in discrete numbers (although a lot of them), but as already said, this solution is under investigation.

These performances may be proposed by the user, for example to play soccer. The user teach the robot to perform something and this performance will be saved in the Memory Module and can be used also in future interactions, which also characterizes cognition. Easy to note that this solutions will require designer creativity to create as many performances as possible , in order to build a large and varied set performances from each behavior.

Regarding the specifics algorithms to collect the objective measures, those that depends on users verbal communication (nW, RWa, Tta) will be provided by the Dialogue System, whereas the others are analyzed with Adaptive Modules own algorithms. The Eye gaze is under development (see section IV) and can already detect face deviation. The posture and emotion analyzers were develop in previous works in the same research group. Zuher and Romero [32] develop a system based on Microsoft kinect to detect key points in persons to extract posture measures, while Gian and Romero [33] considered a regular web cam to classify emotions based on face's features. The both systems need smaller adjustments that are already being handled. Furthermore, a emotion analyzer system using a Convolution Neural Network (CNN) is also being develop by this purpose.

Studies are being performed to create reliable data as the ground truth for training the weights $w_\alpha, w_\beta, w_\gamma$ of the F_{Adp}. With this, it will be possible to training super-visioned machine learning algorithms to optimize the F_{adp} and also compare their efficiency. Some of them are already done (as shown in V) or in development and the rest will be take place soon.

E. Assessment

The Assessment Module has the functionality of collect relevant information about the users along the interaction and display these reports to the designers any time after the interaction. Once the reports generated aim to work as interaction logs, all the measures extracted by the other modules would be gathered and processed to show how they evolved along the time in the interaction. It should be done in a very intuitive manner. Graphic Users Interfaces (GUI) are being developed to do so. These reports will help the designers to plan the next activities based on the difficulties presented by the users and detected by the Assessment Module. Its objective is to be a supportive tool in modeling the users learning profile.

IV. Partial Results

In this section is presented the results that are not published from some modules. The modules or functions that are note mentioned are still in development phase and will be addressed soon.

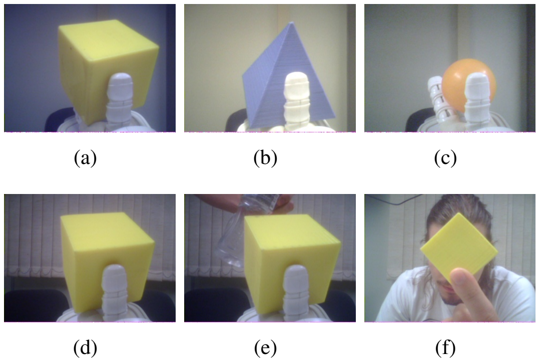
A. Vision

The proposal is that a designer could change the programed content and the change could be done hours before the interactions. Thus, a low training time is desirable in order to change the data base with few samples and training the system quickly. A ensemble classifier was implemented with methods that offers good accuracies with low training time as presented next.

The classification method employed is an ensemble solution which takes into consideration the prediction of three super-visioned machine learning methods: K-Nearst Neighbors (KNN), MultiLayer Perceptron (MLP) and Support Vector Machine (SVM), and in each classifier. There are two most common approach to treat the input data in the three mentioned classifiers for images: the image's raw pixel as array (PXL) and the image's histogram (HST). Therefore, the ensemble takes in account 6 combinations of classifier/approach.

In configurations matters it was set the KNN with fiveth near neighbors (K=5), the MLP with three layers with 100 neurons each and limited max step in 500 and a linear kernel to SVM. Attributes not mentioned here were used with default value. The images content is basic spatial geometry, regarding the figures cube, pyramid and sphere, which are the classes used to label the samples. In this way, the database was trained with 90 images of each of these three classes, resulting in a total of 270 samples collected from NAO's camera in an ideal scene, which means that the images have only the geometric figure in a monochrome background. The training times in this configuration were 0.49, 4.43 and 5.25 seconds, for KNN, MLP and SVM, respectively, and combining the three models, the time spent is on average 10.18 secs.

The test dataset was divided in two groups of ideal and noise images, each one with 90 samples separated in 30 samples for each class. The ideal images were similar to the ones in the training set, on the other hand, the noise groups are images with disturbed scenario in three levels: first only changing the scenario's background (light noise), then adding some other objects in the scene such as bottles, pens, backpacks (medium noise) and, finally, taking away the object from the robot's hand and showed it in different distances from the robot's camera (heavy noise). In Fig. IV-A some examples of used images, are shown.



(a) (b) (c)

(d) (e) (f)

Fig. 3. Images used for tests in the two conditions. First row is the ideal scenes of a (a) cube, (b) pyramid and (c) sphere. Second row is a cube in three noise levels: (d) light, (e) medium and (f) heavy.

In the noise group, 10 samples were collected for each class in each noise level. The accuracy of classifiers for the noise level is shown in Fig. IV-A. Comparing the results of the ideal scene with a noisy scene, as it is shown in Fig. IV-A, it is possible to notice a good accuracy of the classifiers in the group with noises, compared to the ideal group , where stand out the approach using ensemble with the voting of all classifiers, and the MLP classifier with a histogram approach.

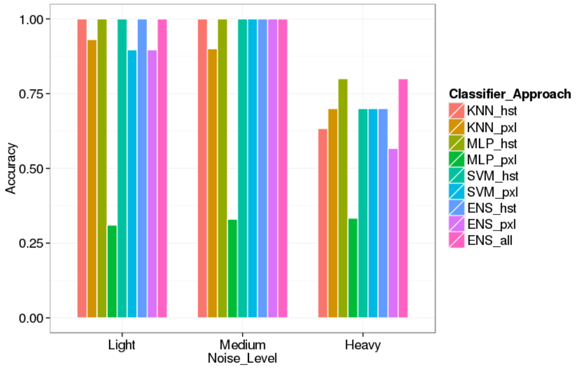


Fig. 4. Comparison between the classifiers separated by the noise levels from scene.

An unexpected contrast was found in the ensemble compared with the SVM with the pixel approach method where the SVM itself classified 100% of the samples versus 98.8% of the ensemble, which in this case is an undesirable situation in suggesting a processing waste. However, for the noise images the ensemble was better than this simple classifier. Analyzing the results, for a noisy scene the ensemble technique remains the score more than other classifiers. In this study, the best accuracy scored from a noisy scene is 93.25%, from MLP

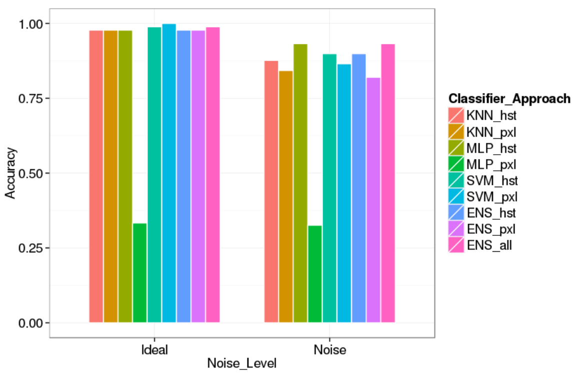


Fig. 5. comparison between the ideal scene and the noise scene with three level noises.

with histogram approach and ensemble with majority voting from all classifications and approaches. These results suggests that MLP classifier has a good generalization. In other words, it has a better accuracy for images that are more distinct from those used in training. We have noticed that most mistakes were made when the objects were far distant from the camera. During the tests, a problem was noticed with MLP classifier with the pixel approach (mlp_pxl), where classifications have become addicted. There is no information about the reason of the problem, more study is needed in the case to understand the reason.

B. Dialog

Only tests to the interpretation function were performed so far. Three different methods were evaluated to filter the given phrase and return a topic that was expected to be the core of the sentence and to be sent to the Memory $ Content Module.

The first method was a very simple effort to filter *stopwords* - words that have no meaning in the sentence, such as ”not”, ”the”, ”what”, etc.- from the input phrase. The second one was based on tagger analysis, translating each one of the phrase's words from Portuguese to English (once the users are Brazilian Portuguese speakers and the methods used so far work only for English language) and looking for nouns that could be interpreted as the main topic of the sentence. The third is a combination of the previous cited methods, where first it is filter the stop words and then the second solution is applied. It decreases processing time due to the less words that it needs to translate.

The three methods were tested using 10 sentences that were defined in three processing difficulty levels. The input set was divided in 3 easy sentences (with only one noun), 4 medium sentences (with two nouns) and 3 hard sentences (with three or more nouns). It is expected that, the more nouns the sentences have, the less accuracy in finding the right core words the methods will show. The results are shown in Table II. The first column is the number of nouns and number of sentences in that condition in parentheses. The second column is the total expected nouns to be found in each group of sentence level difficulty. The rest of the columns are the three methods efficiency discriminated in right answers (how many nouns were found correctly), wrong answers (words not found that were not nouns) and average time to process the sentence and return the topics. In our view, they suggest that two in the three methods can be efficient in terms of isolating nouns from the complete phrase.

The first method has an accuracy of 11 out 11 when tested with simple sentences composed by *stopwords* and one or two nouns. As the sentence's complexity level increases, the efficiency tends to decrease, this was an expected consequence whereas this method works as a filter and do not deal with the different grammar classes.

The second method does not present satisfactory results due to the large distance between the meaning of isolated words and the same words meaning when they come as a phrase part. For instance, the word ”Cup” can have a different meaning isolated than when it comes as ”World Cup”. Where, in this case, the words not cut in the *stopwords* filter may disturb the other words' meaning. Furthermore, this inefficiency is increased because in several times some isolated words in Portuguese have different meaning or taxonomy in English.

As a combination of the two previous methods, the third one has solid results as the number of individual to translate were reduced. By this approach, it is expected that some cases of ambiguity were also reduced, increasing the accuracy in more complex scenarios. According to the results, this method shows better performance when compared with method 1 and 2.

In conclusion, the more simple the method, the less time was spent in the process. Its suggests that, an intelligent algorithm can be applied to predict how much complex the conversation is going. Thus, this algorithm can switch between these methods to offer simple question and less processing time or sophisticated processing, resulting more time processing, according to the dialogue necessity.

C. Memory & Content

In this module evaluation it was first used the same sentences used for the Dialog Module assess. For each sentence, it was prepared five main words that are essential to have in a good answer, based on Portuguese dictionary. It was took the keywords given by the Dialog module to web search with the implemented method and evaluate the answers. An answer was considered very good when it had all the five main words and very bad when no main words were found.

As seen on Table III, the results are based on how well the Dialog Module perceives the question to find keywords to search. The average appearance of keywords per answer is 2.9, which is below the average.

To circumvent this issue, another method that does not depend on the Dialog module was used. It was asked to 10 people asking non personal questions composed only by keywords in the search. For example, ”What is a robot?” turns into ”robot”. It was requested to the questions to be divided into easy, medium and hard that have respectively one keyword, two keywords and three keywords. Each subject asked five questions, being two easy, two medium and one hard. After receiving the answer, the subject rated it from

TABLE II

RIGHT ANSWERS, WRONG ANSWERS AND AVERAGE PROCESSING TIME OF THE THREE NOUNS FINDER OF THE DIALOG SYSTEM, GROUP BY THE SENTENCE DIFFICULTY LEVEL.

		Method 1			Method 2			Method 3		
Number of Nouns (Number of Sentences)	Total Expected Nouns	Right Awnsers	Wrong Awnsers	AVG Time (SD)	Right Awnsers	Wrong Awnsers	AVG Time (SD)	Right Awnsers	Wrong Awnsers	AVG Time (SD)
1 (3)	3	3	0	1e-04(4,7e-05)	3	3	5.32(1.00)	3	0	0.62(0.29)
2 (4)	8	8	0	2e-04(9,4e-05)	5	4	6.43(0.48)	6	0	1.45(0.24)
3 (3)	9	7	4	2e-04(9,4e-05)	5	3	8.85(1.18)	6	3	4.42(0.59)

TABLE III

NUMBER OF KEYWORDS FOUND PER QUESTION.

	sun	**edges**	**vertex**	**sides**	**table**	**robot**	**triangle**	**planet earth**	**solar system**	**blue sky**	**world cup**	**where Atlantic Ocean**
Appears on Dialog	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Keywords	4	5	5	0	0	4	5	5	4	0	3	0

0 to 5 based on the usefulness and if it covered what he expected, where 0 mean ”not at all” and 5 mean ”definitively”. A notorious increased was found, where the average rate for the easy questions was 4.5, 3.7 for the medium ones and, for the hard, 3.1. The final average for all the questions together was 3.9, 1 point more than the previous assessment, what suggests that an improvement in the Dialogue processor would increase this module efficiency too.

However, the results shows that the easier the questions, the higher the approval and this result fit the set that the system is presented upon.

D. Adaptive

In this paper, only results from the Eye gaze (Eg) algorithm will be shown. Although its approach is quite simple, it takes a significant processing power. Thus, this module needs to run in a different computer node that the main process runs. When the main process is initialized, it sends a message to the assessment module that triggers the deviation counter and the emotion classifier algorithms start to run. A socket implementation is being used for nodes communication.

The method used to measure face deviation was the Haar Cascade [34], implemented in *opencv* library. This method is a machine learning based approach where a cascade function is trained from a lot of positive and negative images and then used to detect any object in other images. A XML file was used to load the weights used in the machine learning algorithm, to classify faces. Cascade method works performing convolutions from masks of edge, line and center in a given number of neighbors and this parameter is variable. A series of tests were conducted to analyze its impact.

After the deviation counter awakes, it runs capturing frames from a computer camera, counting the deviations number and time length until it receives a message from the main execution, asking for this information or asking to stop the execution. For each frame, if a face is detected, the current time is assigned in two variables as initial and final time, otherwise, it just updates the final time variable with the current time. When it detects a face again, it compares if the variables difference (final time - initial time) is greater than a chosen threshold and, if this is the case, counts it as a deviation, saving the time length of its occurrence. At the end of execution, the algorithm calculates the total number of deviations, the total time looking away and the total time with the face gazing in the camera’s direction.

The deviation counter algorithm allows some arguments to be changed, such as minimal time to count deviation, allowing more abrupt or soft deviations, furthermore, the higher or lower values of minNeighbors (arguments that is passed to detectmultiscale *opencv* function) may let the attention deviation detection be more or less sensitive. The minNeighbors is the minimal number of neighbors that the cascade method does the rectangle convolution.

For this test, a video if approximately 65 seconds duration was recorded with a voluntary turning his face at some different angles and then with different speeds. The method was configured to allow a tolerance of 20 degrees in four directions (left, right, up and down) without counting a deviation. In total, 48 head turns were recorded but only 8 of them were considered a real attention deviation for us. During the classification of this video, the scale factor (argument of detectmultiscale function) was initially changed, but this compromised the video duration, which disrupted the classification of time deviation.

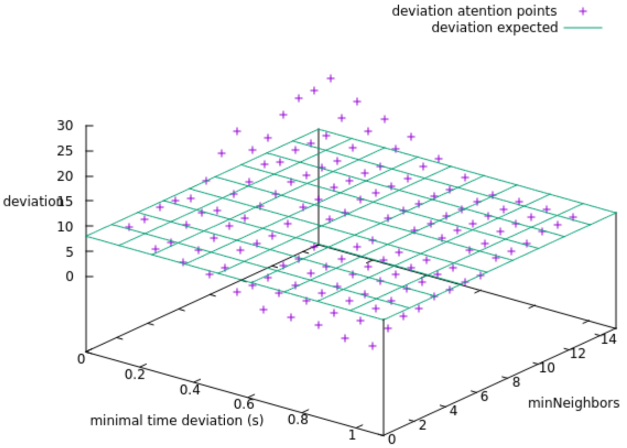


Fig. 6. Graph showing the points collected by varying two parameters (minNeighbors and minimal time deviation), and the plane with expected deviations

After that, the minimum deviation time was varied from 0.1 to 1 second, being increased from 0.1, and the minNeighbors from 1 to 14, being increased from 1, with the data collection the graph shown in Fig. 6 was generated.

As it can be seen in 6 the configuration with high minNeighbors and low minimal time deviation had a very large occurrence of Type II error (because there are more pointers on top of expected plane deviation attention), it means, errors that are false positives. On the opposite, it is possible to see a very large occurrence of Type I error (because are more pointers below the expected plane deviation attention). Two points near to the intersection of the planes was used, because this is the area of interest, where the errors of Type I and Type II are minimize. This points are (time = 0.6, minNeighbors = 9) and (time = 0.7, minNeighbors = 9). In the first and second point, respectively, the algorithm counted 7 and 8 deviation, 18% and 17% of time lost in attention deviation, with accuracy of 97,92% and 95,83%. Complementary data can be observed in the Table IV.

TABLE IV
Detections and not detection deviations of the algorithm using the values: point 1 = (0.6,9,8) and point 2 = (0.7,9,7)

		Deviation	No Deviation
Detected	Point 1	7	1
	Point 2	7	1
No Detected	Point 1	1	39
	Point 2	0	40

V. Studies in Educational HRI

A set of experiments was already conducted in order to evaluate some features of the R-CASTLE project. It was compared the impact of the robot playing a teacher role in educational environment, compared the variation on interactive behavior, the perception of kids with an autonomous robot versus a teleoperated one and children reaction to a same role played by a robot and a tablet. Most experiments was with public primary school students using the humanoid robot NAO.

A. Impact on Robot Behavior Variation

A WoZ experiment was purposed to evaluate children's reaction with a low and high interactive social responses of the robot NAO [31]. To the low interactive group only the voice was used, and the robot remained in the same position during all the 20 minutes session. In the other group, the robot had a repertoire of interactive behaviors as calling the student by it's name, moving it's body and hand shake. Apart from the behaviors, both groups listening the same tips and did exactly the same activity, which consisted of showing the geometric figure asked by NAO. NAO did not ask directly for the figure, but for some features, as number of faces or edges. Regardless the student's right or wrong answer, the robot always explained the figure asked.

It was used the Mean Opinion Score (MOS) and Continuous Audience Response (CAR) to evaluate a subjective opinion of children (MOS) and some judges (CAR) about all the interactive behavior of the robot. To evaluate the learning, an exam after the session with the robot was purposed. Children felt more comfortable, though the questionnaire easier and felt that learned better with high interaction robot. In fact, there were 28.7% better in the exams after the sessions with the robot. The judges of CAR realized that the attention of the children was kept during a longer period when the interaction was more dynamic and concluded that the children had a better reaction to name recognition task because it was the most personal task performed by NAO.

After this reinforcement class, it was performed an question game [35] with the same students, providing an second contact with the robot, with some improvements after the response in the first session. The game is an assessment methodology to investigate how much content children were able to keep from classes with variation in robot-children interaction, when they should give the right answer to "make the robot happy". The question was divided in three categories (easy, medium and hard), and NAO demonstrated some happiness or disappointment signs with every right/wrong answer.

It was used MOS to evaluate the game. Although the students felt that learned something, they said that the game was too hard, and in some times they did not understand what the robot was asking. But the children from the high interaction group in the first experiment as, in average, 17% and 32% better in the game on easy and hard questions. Also, they felt more challenged by their robot and 60% of them said that they studied at home to the game.

B. Autonomous vs WoZ

To investigate whether there was differences in users' acceptance in two conditions: an autonomous robot and a WoZ robot [36]. Two hypotheses was investigated in this study: "the autonomous system condition does not present significant difference from teleoperation condition according to users evaluation" (H1) and "After being informed about the teleoperation, students would reassess their experience with the robot and report a less enjoyable feedback than before knowing about it" (H2). In this study, 82 new students (it was not the same from previously experiments) randomly was divided in two groups and they not previously informed about the purpose of the research.

To evaluate, an Likert Scale form was used to assess the system from users' perspective. Questions was about the users' enjoyment, about the social skills of robot (is the robot rude, boring, intelligent), robot's time to response and if the student think there was someone controlling the robot. After the reveal, WoZ group was asked about the disappointment. Comparing the user enjoyment and the system time response, it was not found a significant difference, what means that the first hypothesis (H1) as confirmed. About the H2, although it was possible to notice a slight decrease in the user enjoyment in the questionnaire after the revelation, no significant difference was found.

The obtained results do not support the hypothesis that the disappointment was a reaction of the majority of the students after being informed. The fact that there was no disappointment is not surprising as children usually play with dolls and toys which they operate, using imagination without any need of autonomy. Also, the majority of students, as they do not have much contact with autonomous robots, do not believed in the autonomous behavior since the beginning of the session.

C. Tablets and Robots Comparison

This study aims to investigate the perception on different devices playing a teach condition [37]. 22 students participate of an english teaching storytelling behavior, 11 with an robot and 11 with an tablet. Both devices has exactly the same portuguese story about an "foreign friend", named John, who teach to the NAO/tablet several animal names in English. Excluding showing the animal's photo, all the resources (actuators) of both devices are fully used, and both as in WoZ condition.

In the evaluation phase, the analysis of three objective factors was taken: gaze deviation, number of stories (all children could stop the experiment in any of 7 possible stories) and a score obtained in the memory evaluation phase. As a result, no greater tendency to increase interest in maintain the story to any group. However, considering the gaze deviation, tablet retained more attention than the robot, perception of learning and comprehension. As the students could read the sentences in the tablet (with has subtitles), it was hypothesized that the tablet interface helped in kept the children attention. Also, some high expectations about the capabilities of the humanoid robot affect negatively the acceptance of NAO in a social role.

VI. Discussion

This paper presented the proposal, objectives, interactive studies, finished and ongoing developments of the R-CASTLE project: a robotic cognitive architecture that aims to deliver a new framework to assist in interactive educational activities design. Several contributions of this project can be highlighted.

For implementation matters, the proposed architecture will allow people without programing skills to easily design HRI tasks and these tasks will be done in an autonomous way. This type of intuitive framework is a good option for researchers coming from other areas, such as psychology and education, that are migrating to the HRI field, since this area has shown a significant growth and good growing potential in last decades. Regular teachers will benefit from its application since R-CASTLE can be used in different interactive devices, dispensing the need of expensive social robots;

For the final application's impacts, there are some points to note. Adaptive behaviors are important to achieve long-term goals and are viable observing user's physical signs, or indicators, along the interactions. Personal users' preferences saved by the system can be helpful to retake their attention in long-term interaction. Working together, the Memory & Content and the Adaptive Modules can provide this robot personalization, which can provide the feeling of rapport between human and robot, optimizing the engagement along the tasks. Combining communicative interaction and content approaches has shown to increase students' learning rates.

So far, the conceptual and technological developments are resulting in a architecture capable of capturing contextual information from audiovisual sensors, approach and learn new concepts, create individual personalization and rapport with students. As the quality of interaction can also vary according cultures, the presented studies performed in Brazil and those that are still ongoing are helping to understand the perception of users from developing countries that have less familiarization with high-tech robots.

Acknowledgment

The authors gratefully acknowledge grants to Brazilian organizations CAPES, CNPq and FAPESP

References

[1] R. Artstein, D. Traum, J. Boberg, A. Gainer, J. Gratch, E. Johnson, A. Leuski, and M. Nakano, "Listen to my body: Does making friends help influence people?" 2017.

[2] L. D. Riek, "Wizard of oz studies in hri: a systematic review and new reporting guidelines," *Journal of Human-Robot Interaction*, vol. 1, no. 1, 2012.

[3] G. Dorais, R. P. Bonasso, D. Kortenkamp, B. Pell, and D. Schreckenghost, "Adjustable autonomy for human-centered autonomous systems," in *Working notes of the Sixteenth International Joint Conference on Artificial Intelligence Workshop on Adjustable Autonomy Systems*, 1999, pp. 16–35.

[4] V. Montreuil, A. Clodic, M. Ransan, and R. Alami, "Planning human centered robot activities," in *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*. IEEE, 2007, pp. 2618–2623.

[5] M. Wald, "Using automatic speech recognition to enhance education for all students: Turning a vision into reality," in *Frontiers in Education, 2005. FIE'05. Proceedings 35th Annual Conference*. IEEE, 2005, pp. S3G–S3G.

[6] S. Lemaignan, F. Garcia, A. Jacq, and P. Dillenbourg, "From real-time attention assessment to with-me-ness in human-robot interaction," in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 2016, pp. 157–164.

[7] P. Langley, J. E. Laird, and S. Rogers, "Cognitive architectures: Research issues and challenges," *Cognitive Systems Research*, vol. 10, no. 2, pp. 141–160, 2009.

[8] J. E. Laird, *The Soar cognitive architecture*. MIT press, 2012.

[9] ACT-R, "Act-r architecture version 6," 2017. [Online]. Available: http://act-r.psy.cmu.edu/

[10] P. Baxter, T. Belpaeme, L. Canamero, P. Cosi, Y. Demiris, V. Enescu, A. Hiolle, I. Kruijff-Korbayova, R. Looije, M. Nalin *et al.*, "Long-term human-robot interaction with young users," in *IEEE/ACM Human-Robot Interaction 2011 Conference (Robots with Children Workshop)*, 2011.

[11] G. Trafton, L. Hiatt, A. Harrison, F. Tamborello, S. Khemlani, and A. Schultz, "Act-r/e: An embodied cognitive architecture for human-robot interaction," *Journal of Human-Robot Interaction*, vol. 2, no. 1, pp. 30–55, 2013.

[12] P. Baxter, S. Lemaignan, and J. G. Trafton, "Cognitive architectures for social human-robot interaction," in *Human-Robot Interaction (HRI), 2016 11th ACM/IEEE International Conference on*. IEEE, 2016, pp. 579–580.

[13] T. Belpaeme, P. Baxter, J. De Greeff, J. Kennedy, R. Read, R. Looije, M. Neerincx, I. Baroni, and M. C. Zelati, "Child-robot interaction: Perspectives and challenges," in *International Conference on Social Robotics*. Springer, 2013, pp. 452–459.

[14] D. Reidsma, V. Charisi, D. Davison, F. Wijnen, J. van der Meij, V. Evers, D. Cameron, S. Fernando, R. Moore, T. Prescott *et al.*, "The easel project: Towards educational human-robot symbiotic interaction," in *Conference on Biomimetic and Biohybrid Systems*. Springer International Publishing, 2016, pp. 297–306.

[15] S. Fernando, R. K. Moore, D. Cameron, E. C. Collins, A. Millings, A. J. Sharkey, and T. J. Prescott, "Automatic recognition of child speech for robotic applications in noisy environments," *arXiv preprint arXiv:1611.02695*, 2016.

[16] A. Zaraki, D. Mazzei, M. Giuliani, and D. De Rossi, "Designing and evaluating a social gaze-control system for a humanoid robot," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 2, pp. 157–168, 2014.

[17] M. ter Maat and D. Heylen, "Flipper: An information state component for spoken dialogue systems," in *International Workshop on Intelligent Virtual Agents*. Springer, 2011, pp. 470–472.

[18] H. Van Welbergen, D. Reidsma, and S. Kopp, "An incremental multimodal realizer for behavior co-articulation and coordination," in *Intelligent virtual agents*. Springer, 2012, pp. 175–188.

[19] D. Cameron, S. Fernando, A. Millings, M. Szollosy, E. Collins, R. Moore, A. Sharkey, and T. Prescott, "Congratulations, its a boy! bench-marking childrens perceptions of the robokind zeno-r25," in *Conference Towards Autonomous Robotic Systems*. Springer International Publishing, 2016, pp. 33–39.

[20] D. Mazzei, N. Lazzeri, D. Hanson, and D. De Rossi, "Hefes: An hybrid engine for facial expressions synthesis to control human-like androids and avatars," in *Biomedical Robotics and Biomechatronics (BioRob), 2012 4th IEEE RAS & EMBS International Conference on*. IEEE, 2012, pp. 195–200.

[21] D. Cameron, S. Fernando, E. Collins, A. Millings, M. Szollosy, R. Moore, A. Sharkey, and A. Prescott, "You made him be alive: Childrens perceptions of animacy in a humanoid robot," in *Lecture Notes in Computer Science*. Springer Verlag, 2017.

[22] D. Cameron, S. Fernando, E. Cowles-Naja, A. Perkins, E. Collins, A. Millings, M. Szollosy, R. Moore, A. Sharkey, and A. Prescott, "Children's age influences their use of biological and mechanical questions towards a humanoid," in *Proceedings of the 18th Towards Autonomous Robotic Systems (TAROS) Conference*. Springer Verlag, 2017.

[23] D. Cameron, S. Fernando, A. Millings, R. Moore, A. Sharkey, and T. Prescott, "Childrens age influences their perceptions of a humanoid robot as being like a person or machine," in *Biomimetic and Biohybrid Systems*. Springer, 2015, pp. 348–353.

[24] V. Charisi, D. Davison, F. Wijnen, J. Meij, D. Reidsma, T. Prescott, W. Joolingen, and V. Evers, "Towards a child-robot symbiotic co-development: a theoretical approach," 2015.

[25] D. Cameron, S. Fernando, A. Millings, M. Szollosy, E. Collins, R. Moore, A. Sharkey, and T. Prescott, "Designing robot personalities for human-robot symbiotic interaction in an educational context," in *Conference on Biomimetic and Biohybrid Systems*. Springer, 2016, pp. 413–417.

[26] A. H. M. Pinto, A. X. Benicasa, L. O. Oliveira, R. C. G. Meneguetti, and R. A. F. Romero, "Attention based object recogniton applied to a humanoid robot," LARS/SBR 2014, 2014, So Carlos. Proceedings of LARS/SBR 2014, 2014. p. 136-141, 2014.

[27] D. Wang and D. Terman, "Image segmentation based on oscillatory correlation," *Neural Computation*, vol. 9, pp. 805–836, 1997.

[28] A. H. M. Pinto, L. O. de Oliveira, A. X. Benicasa, R. C. G. Meneghetti, and R. A. F. Romero, "Inserção de um robô humanoide no ensino de objetos geométricos 2d sobrepostos," in *Anais do Simpósio Brasileiro de Informática na Educação*, vol. 25, no. 1, 2014, pp. 632–641.

[29] S. Frintrop, T. Werner, and G. M. García, "Traditional saliency reloaded: A good old model in new shape," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 82–90.

[30] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[31] D. C. Tozadore, A. H. Pinto, and R. A. Romero, "Variation in a humanoid robot behavior to analyse interaction quality in pedagogical sessions with children," in *Robotics Symposium and IV Brazilian Robotics Symposium (LARS/SBR), 2016 XIII Latin American*. IEEE, 2016, pp. 133–138.

[32] F. Z. M. S. Cavalcante, "Reconhecimento de movimentos humanos para imitacao e controle de um robo humanoide," Master's thesis, Univerisade de Sao Paulo, Sao Carlos, 2012.

[33] G. L. Libralon and R. A. F. Romero, "Mapping of facial elements for emotion analysis," in *Intelligent Systems (BRACIS), 2014 Brazilian Conference on*. IEEE, 2014, pp. 222–227.

[34] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[35] A. H. M. Pinto, D. C. Tozadore, and R. A. F. Romero, "A question game for children aiming the geometrical figures learning by using a humanoid robot," Proceedings of LARS/SBRp. 253-257, 2015.

[36] D. C. Tozadore, A. H. Pinto, G. Trovato, and R. A. Romero, "Wizard of oz vs autonomous: children's perception changes according to robot's operation condition," ACCEPTED in 26th IEEE International Workshop on Robot and Human Interactive Communication, 2017. ROMAN 2017., IEEE, 2017.

[37] D. Tozadore, P. Adam, C. Ranieri, M. Batista, and R. Romero, "Tablets and humanoid robots as engaging platforms for teaching languages," in *In proceedings of XIV Latin American Robotics Symposium and V Brazilian Robotics Symposium (LARS/SBR)*. IEEE Press, 2017.

Daniel Tozadore Daniel Tozadore is a PhD Student in Computation Department at ICMC of the University of São Paulo (USP/Brazil), since 2016.He received the M.S. degree in computer science, in 2016, and B.c. degree, in 2014, from the same institution. He works in the Pedagogical Social Robotics research, using Artificial Intelligence techniques in humanoid robots with the purpose of making Human-Robot interaction as natural as possible for the humans. Its application focuses mainly on early childhood education. His work also includes Machine Learning and Bio-Inspired Computing.

Adam H. M. Pinto Adam Henrique Moreira Pinto received the B.c. degree in 2011 from Goiás Federal University and the M.S. degree in computer science, in 2015, from Institute of Mathematics and Computer Science, ICMC - USP São Carlos, where he is currently pursuing the Ph.D. degree. His current research interests include educational social robotics and EEG signal analysis.

Marcos Camargo Marcos Camargo is a Bachelor Student in Computation Department at ICMC of the University of São Paulo (USP/Brazil), since 2016. He works in Artificial Intelligence techniques classify images in Human-Robot interaction for educational purposes.

João Valentini João Valentini is a Bachelor Student in Computation Department at ICMC of the University of São Paulo (USP/Brazil), since 2016. He works in Natural Language Processing (NLP) to allow human-robot verbal communication for educational purposes.

Rodrigo Zavarizz Rodrigo Zavarizz is a Bachelor Student in Computation Department at ICMC of the University of São Paulo (USP/Brazil), since 2016. He works in cognitive techniques for robot learning applied to HRI designs.

Victor Rodrigues Victor Rodrigues is a Bachelor Student in Computation Department at ICMC of the University of São Paulo (USP/Brazil), since 2016. He works in algorithms to detect users attention deviation applied in HRI context.

Fernando Vedrameto Fernando Vedrameto is a Bachelor Student in Computation Department at ICMC of the University of São Paulo (USP/Brazil), since 2016. He works in algorithms analyze users emotions and adapt the robot's behavior.

Roseli Romero Roseli Romero received her Ph.D. degree in electrical engineering from the University of Campinas, Brazil, in 1993. She is an Associate Professor in Department of Computer Science at ICMC of the University of São Paulo (USP), since 1988. Her research interests include artificial neural networks, machine learning, fuzzy logic, robotics and computational vision.