4th International Conference on
Development and Learning and on Epigenetic Robotics
October 13-16, 2014. Palazzo Ducale, Genoa, Italy



# Emergence of the Use of Pronouns and Names in Triadic Human-Robot Spoken Interaction

Grégoire Pointeau, Maxime Petit, Guillaume Gibert, Peter Ford Dominey
Robot Cognition Laboratory, INSERM U846, Bron FRANCE
Email: {name.surname}@inserm.fr

***Abstract*—We present here a system capable of learning to extract the correct comprehension and production of personal pronouns and proper nouns during Human-Robot or Human-Human interactions. We use external 3D spatial and acoustic sensors with the robot iCub to allow the system to learn the proper mapping between different pronouns and names to their properties in different interaction contexts. The properties are Subject (*Su*), Speaker (*Sp*), Addressee (*Ad*) and Agent (*Ag*). A fast mapping system is used to extract correlation between the different properties. After a learning phase, the robot is able to find the missing property when only 3 out of 4 are known, or at least to discriminate which word cannot be used to be the lacking property. We present results from a set of experiments that provide some insight into aspects of human development.**

***Index Terms*—Embodied robotic, functional language learning, real-time learning, fast mapping, human-robot interaction.**

## I. Introduction

The future of social robotics will be written in the understanding of complex relations, where robots will interact not only with one user, but also with multiple agents. The classical learning of language through one-to-one spoken interactions has been studied for some time ([1]–[4]), but it has been shown that these interactions are insufficient to learn or use correctly personal pronouns [5]. According to Oshima-Takane, learning to use personal pronoun like "You" and "I" is done through observation and with the involvement of the student with several agents ([6], [7]). Gold and Scassellati have made several models using fast mapping for "You" and "I" [8]–[10], but here, we propose a system able to extend the learning from personal, to both personal and impersonal pronoun. This understanding is also an important step for the emergence of self [11].

If we want robots able to be in the middle of humans, behaving as one of them, we need these robots to understand human interactions. In fact, human interactions can be very complex and robots need a robust system able to understand and to acquire the knowledge of directed human interaction in order to be part of the interaction. The goal is not only to extract knowledge and to be able to create new knowledge, but also to use this knowledge at the proper moment for example to understand the type of relationship between different persons. In the future, robots should be able to behave in a human environment, and to get clues about the different relationship between the people present like humans do. This is why we decided to focus about the development of children, and try to apply it to our system, to get a robot with the same developmental results.

To do so, we present here a fast mapping system able to understand the use of different pronouns during a classical interaction either dyadic or triadic. We use a fast mapping system between the use of a pronoun and the context in which it has been used to classify the different subjects of the pronounced entences. The model is based on child development. The pre-required for the understanding of a triadic interaction is to be able to detect the *Sp (speaker), Ad (addressee), Ag (agent)* and *Su (subject)* of an action. Corkum & Moore [12] have shown that at about 9 months, children can detect the direction of an adult's gaze. This age has been put in evidence by Tomasello [13] as the *"Nine month revolution"* and is the starting point of a full understanding of a complex interaction, including triadic interaction. Thus, this will be the starting point of our system. We will give the robot the possibility to detect these interaction properties.

The section II will explain the method and system used (the physical and software architecture), the section III will explain the learning mechanism through fast mapping. The section IV will summarize the experiments and the conditions we tested, and the results obtained. Finally, we will have a discussion part (section V) about the results and limitations of the system and what are our next steps in this field of research, before concluding (section VI).

## II. Method and System Used

### A. Physical Architecture

In this section we will present the system we used for our work. The following study has been realized on the robot iCubLyon01 [14] at the INSERM Robot Cognition Laboratory in Lyon, France. The physical architecture is centered on the robot iCub, and the Reactable (an interactive table), to allow face-to-face physical interaction [15]. We use

a first Kinect above the head of the robot to detect the movements of the present agents. A second Kinect will be used to detect the orientation of each agent, and binaural microphones placed on each ear of the robot to localize the origin of the sound (see Figure 1). This second Kinect needs to be placed at less than 1m from the subject for good results. This is the reason of the use of 2 Kinects.

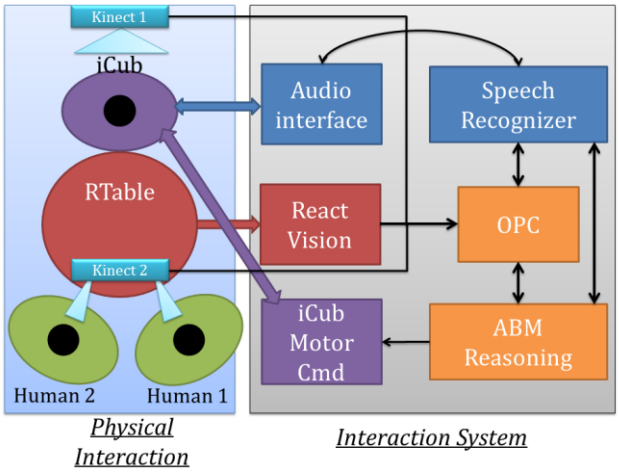


**Figure 1: Physical and software architecture of the system used.**

The software architecture is centered on an Objects Properties Collector (OPC) which can be considered as a working memory, and that represent the state of the world at a given time. In this OPC the contextual data from the different sensors will be stored and all this information will be stored in an episodic-like memory (ELM) and be parsed by a reasoning module (abmReasoning) to create some knowledge relative to the heard sentences, and store it in a semantic memory (SM). The ELM and SM will form the auto-biographical Memory (ABM). More information about ABM can be found in [16], [17].

### B. *Software Architecture and Sensors*

The audio interface and speechRecognizer used are based on Microsoft speech recognizer SAPI5.1. The system, given a grammar, can detect the semantic role of each word in a sentence. For example, we used sentences like: *"Peter put the cross to the left"* or *"You point the circle"*. In these sentences, the first word will be extracted as the subject (or pronoun), the second as a verb, and the last one as the object of the sentence.

The rigid head motion of several human partners can be estimated using the Random Forest algorithm developed by [18]. In fact, a depth camera (Asus Xtion) was placed on the Reactable close to the robot and facing the human partners. Given the depth image provided by the sensor, the Random Forest Head tracking algorithm provided the position and orientation of the human partners' head movements at 30 Hz. This information was used to estimate where the visual attention of each partner was directed to.

The speaker's location was determined using the acoustic signals coming from the binaural microphones placed in the robot's ears. The acoustic signals were retrieved in real-time thanks to the Jack library (http://jackaudio.org/). The Interaural Intensity Difference (IID) was computed. In fact, sound coming from the right has a higher intensity in the right ear microphone than on the left ear one. This difference allowed us to determine if the sound was coming either from left or right. In the triadic setup, the robot was able to determine who the speaker was (see Figure 2).


**Figure 2: System running when the robot is interacting with 2 agents around the Reactable.**

### C. *Method*

In our study, the robot will have a set of training data. For each set of learning data, we will use the pronouns: "*I*" and "*You*", and the proper names: *"Peter"*, *"Maxime"* and *"Grégoire"* (will be referred as "known names"). *"John"* and *"Mark"* will never appear during the learning phase (only the testing phase) and will be referred as "unknown names".

Then for each of the four possible modalities (*Sp*, *Ad*, *Ag*, *Su*) we give a random but doable (ie: *Sp* different of *Ad*) set of the three other modalities, and ask for the fourth one. For example, we will give the system: *"Su = "I"; Sp = "Peter" ; "Ad = iCub"* and the system should return: *"Ag = "Peter""*. Another example would be: we give the system: *"Sp = "iCub" ; Ad = "Maxime" ; Ag = "Maxime" "* and the system should return: *"Su = "You""*.

We will test different learning conditions (that we can easily simulate see Section IV) in order to i) determine how the system learn with constrained conditions, potentially simulating constrained real solution and ii) investigating what kind of interaction is needed for the child to learn personal and impersonal pronoun use. We will test a set of 7 different conditions. These conditions involved two or three agents, the robot can be either spectator or actor. The precise conditions will be explicated in the section IV, at the beginning of each sub section. An agent can talk to someone or not, and an agent can talk about the action of someone (himself included) or not.

As we said earlier, Tomasello has shown in [13] that the child is not able before the *"nine month revolution"* to fully

understand a spoken interaction where he/she is not involved (neither speaker nor addressee). This is what Oshima-Takane calls the "Addressee Condition" [6], and will be summarized by: *Ad= iCub*. Another kind of condition that we tested is the case of *«blind»* children. The particularity of *«blind»* children is that they can only detect the actions related to them: *Ag = iCub* [19]. We will have two conditions with respectively two or three agents, where the *Ag* of the action is always the iCub. The Figure 3 is an example of a tested condition with the corresponding legend .

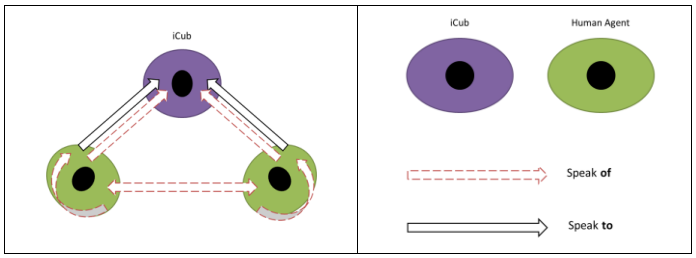


**Figure 3: Example of all available interaction for one of the learning conditions (left). Right: black arrows mean that the origin agent SPEAKS TO the target agent (origin = *Sp*, target = *Ad*). Red arrows mean that the origin agent SPEAKS OF the target agent's action (target = *Ag*).**

## III. Learning Mechanism

### A. Data formalization

For each encountered interaction, we can have access to the four contextual properties of interest: Speaker (*Sp*), Addressee (*Ad*), Agent (*Ag*) and Subject or Pronoun (*Su*). We take the *Su* as higher level of property. For each *Su*, we build a cubic $N^3$ matrix (three dimensions: one for *Sp,* one for *Ad* and one for *Ag*), where N is the number of label encountered. A label can be any *Ag*, *Ad* or *Sp* encountered (i.e.: *"Maxime"*, *"Greg"*…). A label corresponds to a way to refer to a person for the robot. In most cases, the label will correspond to the proper name of the person (e.g. *Peter*), but it can be something generic (e.g. *Agent_5*). However, *Su* include personal pronouns (*"I", "You"*) and proper names that have been used in a sentence. The matrix is then filled with the number of events encountered. The notation (1) gives us the number of events encountered with a particular set of *Su*, *Sp*, *Ad* and *Ag*.

$$M_{Su}(Sp,Ad,Ag) \qquad (1)$$

For example, in the case of the sentence: "*You* point the toy", where: *Su="You"*, *Sp="Greg"*, *Ad="Maxime"*, *Ag="Maxime"*, we will add 1 in the matrix: $M_{You}(Greg,Maxime,Maxime)$, and for "John pushes the cross", where: *Su="John"*, *Sp="Peter"*, *Ad="Greg"*, *Ag="John"*, we will add 1 in the matrix: $M_{John}(Peter,Greg,John)$. We can expect to have only zero in the case of *Sp* is *Ad*, because we consider the case where one doesn't talk to himself. In the case of the apparition of a new label, the matrix will grow and fill the new case according to the number of utterances.

After the learning phase, we will have as many matrices as we have of different *Su*, and each matrix will be of size $N^3$ with N the number of label encountered.

### B. Fast Mapping

As we have seen earlier, the goal of the system is to retrieve the fourth property of an interaction, where the robot knows three properties. It can be used for the example when the robot sees Maxime moving (*Maxime = Ag)*, and want to explain the situation (*iCub = Sp)* to Peter (*Ad = Peter*). What *Su* should he use in this context (*iCub = Sp, Ad = Peter*, Maxime = Ag)? Another utilization could be when the robot hears Maxime speaking (*Maxime = Sp)* while looking at the robot (*Ad = iCub*), using the pronoun "You" (*Su = "You"*), and the iCub wants to find who is the agent to know if he is concerned (context: *Maxime = Sp, Ad = iCub, Su = "You"*).

To find a missing label (*Sp*, *Ad*, *Ag*) or a respectively a pronoun (*Su*), we list all the labels (resp. pronouns) known, and for each, we calculate a Chi Square associated to the corresponding situation (see Table 1). The p-value relative to the Chi-Square will give us some information about the distribution of event (context or no-context) given the label. If this p-value is strong, the two distribution are different and there is an effect of the context on the use or not of the pronoun. The score of the Chi Square will determine the likelihood to use (or not) a specific label or pronoun in a specific context. For a specific context, and a specific label (or pronoun) the Chi Square will be calculated with the data shown in the Table 1, where A is the number of sentences heard with this label (or pronoun) in this context, B is the number of sentences with a different label (or pronoun) is this context. C is the number of sentence with this label (or pronoun) in a different context and D the use of a different label (or pronoun) in a different context.

**Table 1: Table of fast mapping for a specific label/pronoun and a specific context**

| | LABEL/PRONOUN | ~LABEL/PRONOUN |
|---|---|---|
| CONTEXT | A | B |
| ~CONTEXT | C | D |

The Table 1 can correspond to the following situation: *"Can I use this label (resp. pronoun) in this context?"*. If the p-value associated to the Chi Square is above a threshold, the property is rejected. If the p-value is under this threshold, we add the distribution of the property to the score of the label (resp. pronoun) as shown in the pseudo code of the Figure 4.

```
GET THE SPECIFIC CONTEXT.
FOR EACH KNOWN LAB./PRON.:
{
IF: P-VALUE(X²) < THRESHOLD
  - DON'T CHANGE THE LAB./PRON. SCORE
ELSE:
  - ADD TO THE SCORE OF THE LAB./PRON., THE DISTRIBUTION OF
      THE PROPERTY: (A/C – B/D)
}
IF: ONE OR MORE LAB./PRON. HAS A SCORE > 0
- RETURN THE LAB./PRON. WITH HIGHER SCORE
ELSE:
- REMOVE LAB./PRON. WITH SCORE < 0
```

**Figure 4: Pseudo code corresponding to the searching part**

### C. *Data Collection*

In this section we will explain how we manage the collection of learning data. During an interaction with one or several agents around the Reactable, the different modalities will be retrieved as follows:

- ***Sp***: We identify the speaker by using the binaural microphones placed on the iCub.
- ***Ad***: As described in the section II-A, we use a Kinect to detect the orientation of the head of each speaker.
- ***Ag***: We use the Kinect placed above the iCub to detect who is moving, or when the agent is the iCub, he will use proprioception (i.e. check if motors are moving).
- ***Su***: The speechRecognizer is used to extract the subject of the sentence, and to return it.

All this contextual information is collected from the OPC, and stored in the ABM. Once all the information is in the ABM, the reasoning module can create the matrices. In the case of simulated data, each sentence is repeated 5 times in order to simulate the redundancy present in language

## IV. EXPERIMENTS AND RESULTS

In this section, we will present the 7 conditions tested and the results. Each subsection will detail one specific condition and the results obtained. We consider that the robot understands a subject when he can retrieve the *Ag* of the action with a *Su, Sp* and *Ad*, and that he has a correct use of a subject when given *Sp, Ad* and *Ag* he retrieves the correct *Su*.

### A. *Triadic spectator*

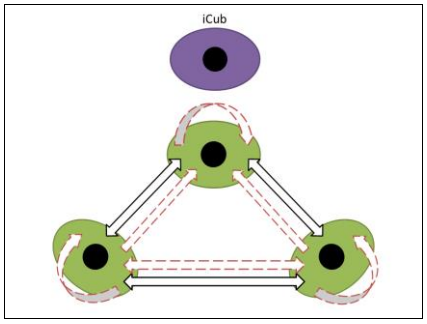


**Figure 5: triadic spectator condition. Three agents talk to each other, about each other.**

This first and "complete" situation is the one we recorded with real data and is the only condition where the sentences are not repeated (see Figure 5). We have three agents: Peter, Maxime and Grégoire. Each agent talks to the other two, about his own action, the action of the addressee, and the action of the third agent. We thus have a total of 18 different sentences. This is a case of "perfect" data without repetition. With this set of 18 sentences, the system is able to understand and to correctly use the pronouns *"I"* and *"You"* and also the use and the understanding of a known name (Peter, Maxime or Grégoire) but not of an unknown name.

| | |
|---|---|
| UNDERSTANDING *"I"* | ✓ |
| CORRECT USE OF *"I"* | ✓ |
| UNDERSTANDING *"YOU"* | ✓ |
| CORRECT USE OF *"YOU"* | ✓ |
| UNDERSTANDING A KNOWN NAME | ✓ |
| CORRECT USE OF KNOWN NAME | ✓ |

The learning data for this condition are summed up here:

| Interaction | *Sp* | *Ad* | *Ag* | *Su* |
|---|---|---|---|---|
| **1** | Greg | Maxime | Greg | "*I*" |
| **2** | Greg | Maxime | Maxime | "*You*" |
| **3** | Greg | Maxime | Peter | "Peter |
| **4** | Greg | Peter | Greg | "*I*" |
| **5** | Greg | Peter | Maxime | "Maxime" |
| **6** | Greg | Peter | Peter | "*You*" |
| **7** | Maxime | Peter | Greg | "Greg" |
| **8** | Maxime | Peter | Maxime | "*I*" |
| **9** | Maxime | Peter | Peter | "*You*" |
| **10** | Maxime | Greg | Greg | "*You*" |
| **11** | Maxime | Greg | Maxime | "*I*" |
| **12** | Maxime | Greg | Peter | "Peter" |
| **13** | Peter | Maxime | Greg | "Greg" |
| **14** | Peter | Maxime | Maxime | "*You*" |
| **15** | Peter | Maxime | Peter | "*I*" |
| **16** | Peter | Greg | Greg | "*You*" |
| **17** | Peter | Greg | Maxime | "Maxime" |
| **18** | Peter | Greg | Peter | "*I*" |

### B. *Dyadic spectator*

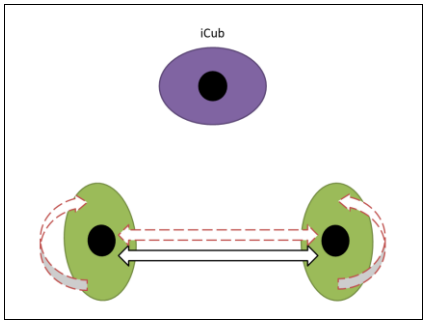


**Figure 6: dyadic spectator condition. Two agents talk to each other, about each other.**

This condition refers to the iCub watching two persons talking to each other about each other (see Figure 6). The iCub is only spectator and in no way involved in the sentences. The agent uses only *"I"* and *"You"* and never an

agent's name. We used for this condition 4 different sentences repeated 5 times, for 20 training sentences. The results show that the robot can understand correctly the use of *"I"* and *"You"*. That means that in the case where the robot wants to describe (*Sp=iCub*) what he or his addressee does (*Ag=iCub or Ag=Peter*) to someone (*Ad=Peter*) he will correctly use the pronoun *"I"* or *"You"*. The robot is also able to understand *"I"* and *"You"* in a sentence (ie: when we give the robot *Su = "I"*, resp. *Su = "You"*, *Sp* and *Ad*, the robot assumes that the *Ag* is the *Sp* for *"I"* resp. *Ag = Ad* for *"You"*.

| | |
|---|---|
| UNDERSTANDING *"I"* | ✓ |
| CORRECT USE OF *"I"* | ✓ |
| UNDERSTANDING *"YOU"* | ✓ |
| CORRECT USE OF *"YOU"* | ✓ |
| UNDERSTANDING A NAME | × |
| CORRECT USE OF A NAME | × |

### C. *Triadic agent*

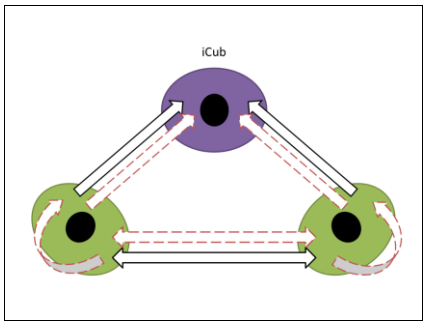


**Figure 7: triadic agent condition. Two agents talk to each other and to the iCub about the three of them.**

This condition is similar to the previous one (*Triadic Spectator)* but this time, one of the agent is the iCub, and does not speak (see Figure 7). We have thus not 18 but 12 different sentences that we repeated 5 times each for a total of 60 learning sentences. The results are similar to those for the triadic spectator, except that this time the robot is unable to use correctly the pronoun *"I"*. When the robot is talking about him doing an action, he will prefer using *"iCub"* rather than *"I"* while for the other agent, he can use it correctly. But an interesting fact is that if we ask the robot who would be the agent in the case of a sentence said by the iCub using *"I"*, the robot correctly understands that *"I"* refers to him.

| | |
|---|---|
| UNDERSTANDING *"I"* | ✓ |
| CORRECT USE OF *"I"* | × |
| UNDERSTANDING *"YOU"* | ✓ |
| CORRECT USE OF *"YOU"* | ✓ |
| UNDERSTANDING A KNOWN NAME | ✓ |
| CORRECT USE OF KNOWN NAME | ✓ |

### D. *«Blind» three agents*

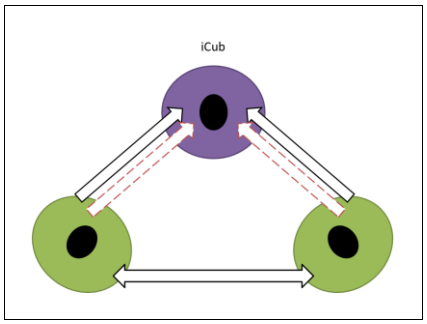


**Figure 8: *«blind»* three agents condition. Two agents talk to each other and to the iCub, about the iCub's actions.**

This condition is a triadic condition with the robot and two agents (see Figure 8). The two agents can only talk about the action of the robot (*Ag* = iCub). In this condition one agent can talk to the other or to the iCub, about the iCub (third person). We have a set of 4 different sentences repeated 5 times for a total of 20 learning sentences. The results are that the robot understands and uses correctly *"You"* but not *"I"* (because he has never heard it). Also, the robot is confused with understanding and using the name *"iCub"*. It could be the same as *"he"* or could just refer to "someone else".

| | |
|---|---|
| UNDERSTANDING *"I"* | × |
| CORRECT USE OF *"I"* | × |
| UNDERSTANDING *"YOU"* | ✓ |
| CORRECT USE OF *"YOU"* | ✓ |
| UNDERSTANDING A NAME | × |
| CORRECT USE OF NAME | × |

### E. *«Blind» two agents*

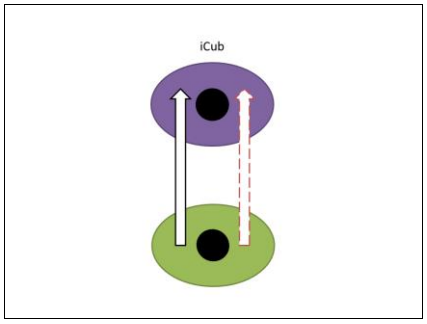


**Figure 9: *«blind»* two agents condition. An agent talks to the iCub about the iCub's actions.**

In this condition, we have only two agents: the iCub and a human agent (see Figure 9). The human only talks to the robot, about the robot. We have thus only one sentence possible (*"You do ..."*), repeated 5 times. With this learning data, as expected, the robot is unable to understand or use *"I"* or *"You"*. Also, the robot doesn't acquire any knowledge about the use of any name.

| | |
|---|---|
| UNDERSTANDING *"I"* | × |
| CORRECT USE OF *"I"* | × |
| UNDERSTANDING *"YOU"* | × |
| CORRECT USE OF *"YOU"* | × |
| UNDERSTANDING A NAME | × |
| CORRECT USE OF NAME | × |

### F. «Addressee» three agents

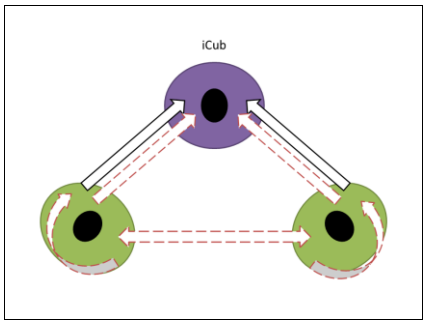


**Figure 10: *«addressee»* three agents condition. Two agents talk to the iCub about the action of everyone.**

In this condition, the robot is in presence of two agents (see Figure 10). The robot only perceives the sentences directed to him (*Ad* = iCub). But an agent can talk of the action of a third person, while he is talking to the robot. We have a set of 6 different sentences (2 *Sp*, talking to 1 *Ad*, about 3 different *Ag*), repeated 5 times, for a total of 30 sentences. The results are that the robot correctly understands and uses *"I"* and *"You"* and understand a known name as pronoun (ie: *"Peter does …"*) but not an unknown name. In this condition, the robot never hears his own name.

| | |
|---|---|
| UNDERSTANDING *"I"* | ✓ |
| CORRECT USE OF *"I"* | ✓ |
| UNDERSTANDING *"YOU"* | ✓ |
| CORRECT USE OF *"YOU"* | ✓ |
| UNDERSTANDING A KNOWN NAME | ✓ |
| CORRECT USE OF KNOWN NAME | ✓ |

### G. «Addressee» two agents

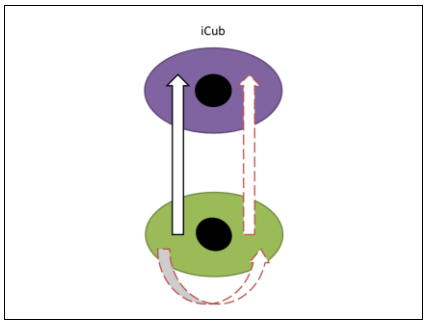


**Figure 11: *«addressee»* two agents condition. A agent talks to the iCub about the actions of both the human and the robot.**

In this condition, the robot is in presence of one other agent (see Figure 11). This human agent talks to the robot about the actions of both human and robot. We have a set of 2 different sentences (*"I do …" "You do …"*) repeated 5 times for a learning set of 10 sentences. One notable thing in this condition is that the robot fully understands and uses correctly *"I"* and *"You"*, even if he did not hear them in at least two different situations, where we could expect the robot to understand that, as a child does sometimes, his name is *"You"* and the name of the human is *"I"*. We will discuss it in the following section.

| | |
|---|---|
| UNDERSTANDING *"I"* | ✓ |
| CORRECT USE OF *"I"* | ✓ |
| UNDERSTANDING *"YOU"* | ✓ |
| CORRECT USE OF *"YOU"* | ✓ |
| UNDERSTANDING A NAME | ✓ |
| CORRECT USE OF NAME | ✓ |

## V. DISCUSSION

The results we obtain with our cross validation system are primarily those that we expected based on child development. Indeed, we have seen that the case of a full triadic interaction observed by the robot provides the most information about the use of a personal pronoun (or proper noun). A fast mapping system allows the detection of the situation where a certain pronoun should be use. The learning phase (filling the matrices) and the working phase (finding the good lacking property) work in real time. The system is not greedy in computation or memory, but allows the understanding of an interaction with several agents, and to use correctly different subjects for a sentence according to the situation. However we can see a few limitations with our system.

The first limitation is the non-generalization of plural pronouns like *"We"* or *"They"*. This is in current development, and will need a processing of several *Sp*, *Ad*, *Ag* and not a one-to-one system as we have currently.

The second limitation is in the gender or social relation. One of our future research axis is to work on the discrimination of *"he"* and *"she"* but also more social relation has: parent/child professor/student. To do so, we will have to extend our matrix system for a more dynamical system extensible to more properties than the 4 that we have now (*Su, Sp, Ad, Ag*).

The third limitation is observed on the result that we obtained in the condition *"«addressee» two agent"*. The robot only witnesses a Human saying *"I"* when he refers to himself, and *"You"* when he talks to and about the iCub. We thus expected the robot to be confused between the use of *"I"* for *Sp=Ag* and when it refers to the Human, and vice versa for *"You"* and the robot. Why the robot does not think he is *"You"* and the Human is *"I"*? The answer, is because we put the same weight to each property. The simple properties like: *"Ag=Human"* have as much weight as a property "double" like *"Sp=Ag"*, or a property "triple" like *"Sp fixed, Ad fixed, Su fixed"*. A "triple" property corresponds to an exact known situation, and a "simple" property to a simple fact. A "double" property is less intuitive: the robot searches a more complex correspondence between different contextual information. In our case, when the robot has to use *"I"*, he checks for example the case where he has to talk about his own actions: *"Sp=Ag"; "Ad!=Ag"; "Ag=iCub"*. These three properties are true. The first two will be in favor or using *"I"* and the third one is in favor of *"You"*. Because each property has the same weight, the robot will choose to use *"I"*. With a bigger weight to the simplest properties, and for the exact known situation, we

could solve this kind of situation and recreate the ambiguity seen in children as shown by Gold and Scassellati [8].

The fourth limitation is the fact that the robot cannot generalize to an unknown name. Indeed, if one hears a sentence of the type: "Mark put something somewhere", we know that *Mark* will be the agent of the action. We do so, because we know that *Mark* is not another unknown pronoun until now, but is a name. The robot does not know if what he hears is a name or a pronoun. The first time he hears *Mark*, he could try to analyze it as he would do for *"You"*. This is another limitation also experienced, until we reach the knowledge of all existing pronouns, and we categorize every other subject as a name, especially if we know that the word is a name.

With this system as a proof of concept, we can now in the future generalize the learning of the nomination of any agent according to the context. The future step in our work will lead to the good use or the good understanding of the appropriate word for a more complex interaction. For example, a child will address to his father with the word *"dad"*, while a friend will call him by his name. With this concept we will now be able to extract more advanced relationship between people, and also we will be able to use this knowledge for a better interaction. However, the first step that we showed in this study, is needed in order to develop more complex reasoning and knowledge.

## VI. Conclusion

We provide here a simple system to learn correctly different personal or impersonal pronouns, through fast mapping. A small amount of data is required in the memory to have the expected result and a good comprehension of different pronouns. Indeed 18 sentences are enough. This system is easily embodied and allows the robot to be more efficient in the case of a complex interaction (several agents present). We can free ourselves from the classical HRI with one robot, one human, to go to more realistic interaction, and a better understanding of the robot of the world in front of him. Even if the system is not yet totally accomplished, this first version is a good step forward for the robot in term of understanding other, and situating himself in a complex world made of several agents.

## Acknowledgment

This work was supported by WISIWYD (FP7-ICT-612139) and the ANR SWoOZ (PDOC01901) projects

.

## References

[1] D. K. Roy and A. P. Pentland, "Learning words from sights and sounds: a computational model," *Cogn. Sci.*, vol. 26, no. 1, pp. 113–146, Jan. 2002.

[2] L. Steels and F. Kaplan, "AIBO's first words: The social learning of language and meaning," *Evol. Commun.*, vol. 4, no. 1, pp. 3–32, 2001.

[3] X. Hinaut, M. Petit, G. Pointeau, and P. F. Dominey, "Exploring the Acquisition and Production of Grammatical Constructions Through Human-Robot Interaction," *Front. Neurorobot.*, pp. 1–34.

[4] P. Dominey and J. Boucher, "Learning to talk about events from narrated video in a construction grammar framework," *Artif. Intell.*, vol. 167, no. 1–2, pp. 31–61, Sep. 2005.

[5] K. Gold and B. Scassellati, "Grounded pronoun learning and pronoun reversal," in *Proceedings of the 5th International Conference on Development and Learning.*, 2006.

[6] Y. Oshima-Takane, "The learning of first and second person pronouns in English: network models and analysis," *J. Child Dev.*, vol. 26, pp. 545–575, 1999.

[7] T. Shultz, D. Buckingham, and Y. Oshima-Takane, *A Connectionist Model of the Learning of Personal Pronouns in English*. 1994.

[8] K. Gold and B. Scassellati, "Using context and sensory data to learn first and second person pronouns," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, 2006, pp. 110–117.

[9] K. Gold, M. Doniec, and B. Scassellati, "Learning grounded semantics with word trees: Prepositions and pronouns," *2007 IEEE 6th Int. Conf. Dev. Learn.*, pp. 25–30, Jul. 2007.

[10] K. Gold and B. Scassellati, "Learning acceptable windows of contingency," *Conn. Sci.*, vol. 18, no. 2, pp. 217–228, Jun. 2006.

[11] A. Imbens-bailey and A. Pan, "The Pragmatics of Self- and Other-Reference in," 1998.

[12] V. Corkum and C. Moore, "The origins of joint visual attention in infants.," *Dev. Psychol.*, vol. 34, no. 1, pp. 28–38, Jan. 1998.

[13] M. Tomasello, *The cultural origins of human cognition*. 2009.

[14] G. Sandini, G. Metta, and D. Vernon, "The iCub Cognitive Humanoid Robot : An Open-System Research Platform for Enactive Cognition Enactive Cognition : Why Create a Cognitive Humanoid," *50 years Artif. Intell.*, pp. 359–370, 2007.

[15] G. Geiger, N. Alber, S. Jordà, and M. Alonso, "The reactable: A collaborative musical instrument for playing and understanding music," *Her&Mus. Herit. ...*, vol. 4, pp. 36–43, 2010.

[16] G. Pointeau, M. Petit, and P. Dominey, "Embodied simulation based on autobiographical memory," in *Living Machines*, 2013, pp. 240–250.

[17] G. Pointeau, M. Petit, and P. Dominey, "Successive Developmental Levels of AutobiographicalMemory for Learning Through Social Interaction," *IEEE Trans. Auton. Ment. Dev.*, vol. 0, p. in Press, 2014.

[18] G. Fanelli, T. Weise, J. Gall, and L. Van Gool, "Real time head pose estimation from consumer depth cameras," *Pattern Recognit.*, 2011.

[19] K. Gold and B. Scassellati, "Using probabilistic reasoning over time to self-recognize," *Rob. Auton. Syst.*, vol. 57, no. 4, pp. 384–392, Apr. 2009.

 