Guidelines for creating man-machine multimodal interfaces

João Ranhel 1*, Cacilda Vilela2

1 Information Engineering, Federal University of ABC, Santo André, SP, Brazil
2 Faculty of Philosophy, Languages and Human Sciences, University of São Paulo, São Paulo, SP, Brazil

*E-mail: joao.ranhel@ufabc.edu.br (JR)

Abstract
Understanding details of human multimodal interaction can elucidate many aspects of the type of information processing machines must perform to interact with humans. This article gives an overview of recent findings from Linguistics regarding the organization of conversation in turns, adjacent pairs, (dis)preferred responses, (self)repairs, etc. Besides, we describe how multiple modalities of signs interfere with each other modifying meanings. Then, we propose an abstract algorithm that describes how a machine can implement a double-feedback system that can reproduce a human-like face-to-face interaction by processing various signs, such as verbal, prosodic, facial expressions, gestures, etc. Multimodal face-to-face interactions enrich the exchange of information between agents, mainly because these agents are active all the time by emitting and interpreting signs simultaneously. This article is not about an untested new computational model. Instead, it translates findings from Linguistics as guidelines for designs of multimodal man-machine interfaces. An algorithm is presented. Brought from Linguistics, it is a description pointing out how human face-to-face interactions work. The linguistic findings reported here are the first steps towards the integration of multimodal communication. Some developers involved on interface designs carry on working on isolated models for interpreting text, grammar, gestures and facial expressions, neglecting the interwoven between these signs. In contrast, for linguists working on the state-of-the-art multimodal integration, the interpretation of separated modalities leads to an incomplete interpretation, if not to a miscomprehension of information. The algorithm proposed herein intends to guide man-machine interface designers who want to integrate multimodal components on face-to-face interactions as close as possible to those performed between humans.

1. Introduction
Undoubtedly, the use of language is the most complex behavior known. Linguistics is a field of science that studies the attributes of languages, especially Natural Language(s) (NL), as human languages are known. NL serves as a tool that an agent (the emitter) uses to transmit multiples types of signs (sounds, gestures, facial expression, etc. – or multimodal signs), which are expected to be correctly interpreted by another agent (the recipient). In humans, we usually call “mind” such interpretative systems, thus, in this article we use the term mind instead of interpretative systems. In this sense, the purpose of NL is to function as a tool that emitters use to trigger particular "states" in an interpretative mind. We can also say that emitters are motivated by "intentions" for communicating something, what lead them to use multiple resources for producing specific mental states in the recipient's mind. We use mind for both, humans and machines, presuming that someday machines will have an information processing system (a kind of mind) that generates and interprets messages using multimodal signs; what is to say we expect that machines will use NL in the same way humans do.

Human multimodal communication has been the subject of studies in several areas, such as anthropology, psychology, linguistics, sociology, cognitive science, among others. This field of knowledge draws on sources derived from the studies as diverse as the origin of multimodal communication (1), or gestures and representation in animals compared to human multimodality (2), (3), until more recently, with the development of multimodal interfaces for man-machine
interactions with special attention to robotics (4), (5), (6). Knowledge on the area have advanced in the last decades from theoretical evaluation (as in (7), (8)) to the developing multimodal interfaces for humanoid robots – see review in Turk (9).

Our goal with this article is to describe progresses in Linguistics, especially in the field of multimodal face-to-face communication applied to human-machine interactions. We propose herein a kernel in the form of an algorithm that controls the multimodal fluidity in a face-to-face interaction, which uses verbal, prosodic, facial, gestural, and postural channels. This proposal results from an extensive work in which several videos of face-to-face interactions between humans were analyzed. The main goal of that work was to describe accurately how (and when) each multimodal sign interferes on fluency, on turn control, on repairs, as well as on the meaning of the messages. Within this algorithm we point out when certain “procedures” must be called to execute tasks such as interpreting linguistic meaning, semantic meaning, etc. We do not describe how to perform interpretation, which we understand are developed apart from this integrative algorithm.

We describe a core algorithm that performs a high-level doubled-loop feedback. This structure is necessary for managing the exchange of multimodal messages between two agents in real-time face-to-face interactions. Derived from human interaction observations, the approach presented here may serve as guideline for multimodal natural language processing developers in Artificial Intelligence (AI). We have tried to find a way to explain what kind of computational tasks must be performed to implement the human form of multimodal interaction in a human-machine interface (HMI). We neither have invented a new “model” for an HMI software nor have implemented this algorithm. Instead, the algorithm presented forward is a “description”, or a “translation” of the findings in Linguistics concerning multimodal interaction. We used the algorithmic form for easy translation to other scientific communities. The algorithm serves to illustrate a “set of requirements” that intelligent HMI designers will have to deal with.

Our research neither is based on data analysis of computational interactions nor brings results from a software interface implemented with the algorithm described below. Instead, what is described in this article results from hundreds of hours of face-to-face multimodal video analysis in the Linguistic field, a work that was already evaluated and published (10), (11). Our effort here is to translate that findings to another area of knowledge we suppose may be interested.

This article is organized as follows: in section II, the reader finds a brief description concerning theories and concepts that we borrow to compose this approach. In section III we describe materials and methods we have used in the analysis of human interactions. In section IV we show the results obtained in our observations and how we extracted the fundamentals to create the algorithm. In section V we describe the algorithm in the form of a pseudocode and analyze its operation. Section VI holds a critical analysis and discussions about some assumptions in our approach, as well as proposals for further investigation that can increase the knowledge about face-to-face multimodal interaction, especially research supported by neuroscience. In section VII we describe our conclusions.

2. Theories

Linguists try to comprehend and explain (the use of) language by various points of view, while scientists, engineers and the Artificial Intelligence (AI) community try to reproduce the use of natural languages in machines. Linguistics is composed of several theoretical fields, such as phonology, morphology, syntactics, semantics, among others related to social, historical and psychological aspects of language. A primary level of NL processing is to comprehend the content of a message, for instance, it would be to grasp an utterance such as "it's 7 o'clock". This level of interpretation is called "linguistic meaning". However, the one who emits this message may have a practical intention, which might be: "hurry up, we're late". Pragmatics is the area that studies this kind of linguistic interpretation, which is strongly dependent on context. Beyond that, there are higher levels of context-dependent interpretation. For instance, algorithms of "semantic interpretation" in AI usually assign a meaning to each word in an utterance, then words are combined within the sentences, and
then an overall meaning is extracted from that message. Semantic analysis is a complex area involving logic, which results in implications and assumptions. Beyond that, another area of linguistic interpretation is the "discourse analysis"; however, it is out of the scope of this article to go further on it.

In this article, we borrowed knowledge and concepts that are studied within the areas of Pragmatics, Conversation Analysis (CA) (12), Gesture Studies (13), and Nonverbal Communication (14). We have also borrowed concepts of BDI agents’ model, widely known in AI community (15), (16), (17). BDI stands for agents with “beliefs, desires and intentions” which are embedded on their ontology. Concepts used in BDI models are well suited to our approach. In BDI model, an agent has desires, e.g.: “to cause a mental state in another agent”. To realize such desires, the agent creates plans of actions. Plans that the agent undertakes to perform become intentions, hence the agent starts carrying them out. The causation of “mental states” in other agent’s mind is a subject studied on Theory of Mind (ToM) (18), an area of Cognitive Science from whom we have also borrowed concepts.

Concerning face-to-face interactions, a ‘common sense’ understanding is that there are well-defined roles when face-to-face interactions take place. For example, it is believed that actively, a speaker emits a message while a listener, passively, interprets the received signs. However, new findings show that this view is not correct, that both interactors are active during the time they interact, and that signs sent by both significantly affect the flow of communication, as well as the whole meaning of the message. For practical reasons, hereafter we will treat the "interactor/speaker" as the emitter and the "interactor/listener" as a recipient, although these roles are all the time executed by both, as shown later.

We use in this article the term sign borrowed from Semiotics (19), (20). It is common in AI community to define language as a discrete, categorical, and symbolic system, which deals with symbols, instead of signs. Semiotics studies the construction of meaning, the process of semiosis, as well as meaning in communication. One of the types of signs defined in semiotics is the symbol, but there are others, such as index and icon. We shall consider that, when planning and emitting a multimodal message, the agent uses signs that are not only symbols. For instance, when the agent’s finger points something or somewhere, or when the agent vocalizes ah, oh, hum etc. Thus, we understand that it is better to use the semiotic term sign for multimodal processes because sign is more comprehensive than the term symbol.

Other theories could be applied to this approach. It is well known that affects play an important role when people interact face to face. Several works describe the importance of affects not only in interactions but also in decision making (21). In this sense, the Affect Control Theory (ACT - see(22), (23), (24), (25)) can perfectly be included in this framework. In addition, other models that deal with semantic interpretation and generalized references, such as Database Semantics (DBS (26,27), (28)) can also be incorporated to this framework. These and other models can be part of a deeper structure when implementing our algorithm. Therefore, our approach deals with the explanation of how the operational flow of face-to-face interaction occurs.

As said, distinct theories from a set of disciplines have been used in the composition of this work. The main theories are briefly discussed in this section in order to make their concepts accessible to a wider audience.

2.1 The BDI model

The belief–desire–intention (BDI) is a software model developed for programming intelligent agents (see (15), (16) (17)), characterized by including beliefs, desires and intentions in agent’s model. The model uses these concepts to solve a problem in agent programming: to provide a mechanism for separating the activity of selecting a plan (from a library or an external planner application) from the execution of currently active plans. Consequently, BDI agents can balance the time spent on deliberating about plans (choosing what to do) and executing those plans (doing it). A third activity
foreseen in the style, the creation of plans (planning), is not within the scope of the model itself, and is left to the system designer and programmer.

The BDI model appears in our framework because we suppose two agents (people or machines) interacting face-to-face. In this scenario, each agent must analyze more than one type of signs generated by the other. To do that, firstly, the agents must share knowledge (see Grounding - section 2.5), which is related to the term beliefs in BDI agents. Moreover, in order to communicate an idea, an agent must carry out a plan of actions to generate a series of signs (sounds, gestures, prosody, etc.). The plan that the agent creates is equivalent to the term desire in the BDI model. Furthermore, to the extent the agent undertakes to execute such plan, this action equates to the intention in the BDI model. The algorithm we presented later uses BDI as inspiration, but it is not a direct implementation of BDI.

2.2 Theory of Mind
Theory of mind (often abbreviated ToM) is the ability of an “agent” to assign mental states—emotions, intentions, desires, knowledge, beliefs, etc.—to oneself, and to the others; as well as to understand that the others have emotions, beliefs, knowledge, desires, intentions, and perspectives that are different from one’s own (18), (29), (30). Theory of Mind is crucial to everyday social interactions and is used when analyzing, judging, and inferring the other’s behavior. ToM is a distinct subject and should not be confused with philosophy of mind.

Let us link ToM with BDI model in the context of face-to-face interaction. It has been said that a BDI agent must devise a sequence of signs to communicate. However, agents may want more than just communicate. In fact, most of the time agents want to be understood by the others unequivocally. Saying differently, agents may not want simply to broadcast an alarm. By using NL, agents enchain series of signs to pass complex ideas they want to be correctly interpreted. It would be part of the agent’s communication plan to predict the timing for emitting sounds, to choose prosody, gestures, expressions, among other signs. In order to choose which signs to place, and how to use them, the agent should have a notion of what is going to happen in the interpreter’s mind. Moreover, an agent often has the intention of being interpreted in a particular way, for example, being ironic, severe, sarcastic, etc. Hence, the agent plans which signs to use that may cause that interpretation in the recipient’s mind. Therefore, an agent use ToM for creating a communication plan, which is then executed within the scheme of the BDI model. How the agent performs such plan within an interactive situation is the subject of the next theory.

2.3 The Conversation Analysis Theory
To our knowledge, Conversation Analysis (CA) (12) (31), an area of NL studies, is the theory that best explains the face-to-face speech interactions. CA is an approach to the study of social interaction, embracing speech conduct, in situations of everyday life. As its name implies, CA began with a focus on casual conversation, but its methods were subsequently adapted to embrace more task- and institution-centered interactions, such as those occurring in doctors' offices, courts, law enforcement, helplines, educational settings, and the mass media. Consequently, the term ‘conversation analysis’ has become something of a misnomer, but it has continued as a term for a distinctive and successful approach to the analysis of social interactions.

CA has some important concepts revealed by microlevel analysis that explain how people interact in a conversation, the most relevant to our proposal are detailed below. It is important to remember that CA theorists do not work with the prospect of intentions or mental states. In the eyes of these theorists it may seem incompatible that we are joining such different theories. In a way, we are crossing the boundaries of one field of research by incorporating knowledge from other disciplines in order to go a step further.

CA was born inspired by ethnomethodology (32) which is the study of the methods people use to understand and produce the social order in which they live. Due to the ethnomethodological bias, it is not considered in CA the attribution of mental states to interacting agents. A CA researcher
considers only what is explicitly said / shown during a conversation. This is a viewpoint of a third person analyzing a video-recorded or audio-only interaction that should not guess the participant’s mental states. Although we used the concepts of CA, we cannot limit ourselves to just observing the interactors, since we must create agents that produces the conversation. In this sense, our point of view changes to an agent’s “mind”, which has to generate and interpret communicative action. This forces us to aggregate knowledge that seems incompatible to each other. In fact, we incorporate solid knowledge provided from CA and applied them in other area, preserving the main characteristics of what CA teaches us.

2.3.1 Adjacent Pairs
One of the concepts due to AC, left by H. Sacks, and later developed by G. Jefferson and E. Scheglof, concerns the fact that participants realize "paired" speeches (12), (33). In other words, what is said in an utterance by an interactor restrains the actions that will be produced by the other interactor’s next speech. In this sense, there are always shifts on the communicative actions, so called turns; thus, to make sense, the next turn for an interactor should be paired to the anterior utterances of the previous interactor. For instance, when one of the interactors makes an invitation to the other, the actions that would make sense on the next turn are "to accept", "to limit the acceptance", or "to refusal" such invitation.

2.3.2 The Organization of Preference/Dispreference
Closely linked to the notion of adjacent pairs is the organization of (dis)preferred responses (34). Let see once again the situation of an invitation. In general, when an interactor A invites an interactor B to do something, the interactor A expects that the interactor B accepts his invitation. When interactor B does positively answer the invitation, we say that B’s response is a preferred one, i.e. B’s response has fulfilled the expectation of interactor A. For CA, preferred actions are actions that are unnoticed, unmarked action, since it is the expected action; it conforms closely to the norm of a specific situation. The organization of preferred behavior is not only noticed in speech exchanges, it can be seen in any domain of social action. As Enfield (35) exemplifies: John is a plumber and by dressing in overalls while at work, he chooses the default, unmarked course of action. It is not only practical, but it is expected, the preferred action. He will not be held accountable or even commented on for doing it. However, if he works wearing a dress (dispreferred action), things could probably be different. From that understanding, we can say that the interactor’s response can be a preferred action whenever his responsive actions agree with the ongoing situation. Whenever his responsive action shows ambiguous alignment, we say that his response is neutral (e.g. whenever someone remains in silence when his interactor asks for an opinion about something; when he displays a blank face etc.). Notice that these behaviors can be interpreted as possible marks of disagreement, but since the interactor does not explicitly produced a speech showing his position about the topic, his behavior is considered a neutral one. Finally, when the interactor’s response shows disagreement with the ongoing situation, it is considered a dispreferred action, since it has not satisfied the other’s interactor expectation.

In sum, according to CA perspective, disagreement is closely related to the notion of (dis)preferred responses. Preferred responses are those that are expected, being restricted by context or by cultural rules. For instance, if agent A greets agent B, it is expected that agent B greets in reply. The agent B’s response, considering that he greets in reply, is considered a preferred response. In case the agent B does not greet in reply, he could be taken as a rude person, i.e. his behavior could be interpreted as a dispreferred action for this specific situation. Following this understanding, the agent’s responses of disagreement during a political dispute, for instance, will be considered preferred answers since they will fulfill the expectation of the moment. On the contrary, if the agent agrees with his opponent, his response will be considered dispreferred for the situation.
2.3.3 Turn-taking Organization

The turn-taking organization is a system that orients the social behavior people elicit in their context-situated usage of NL. Whenever people are requesting, inviting, questioning, answering, agreeing, disagreeing, complaining, excusing, insulting and so on, they do it in turns constructed and distributed through the turn-taking system. The notion of adjacent pairs can only be productive if we accept the turn-taking organization, if we understand that conversation is only possible when interactors change their role as emitter and recipient all the time.

According to Sacks, Schegloff, and Jefferson (31), a seminal text about the organization of turn-taking system, a turn consists of vocal/verbal production. The turn can be composed of a single word or many sentences. However, this description of a turn is being questioned by researchers from Gesture Studies and Nonverbal Communication. These researchers claim that gestural actions can also be interpreted as turns. For instance, a request can be done by the sequence organization: if agent B requests the knife using an imperative utterance (e.g. Pass me that knife), the interactor produces the responsive move of delivering the knife to the requester. In this case, some researchers (e.g. Enfield (35), Mondada, (36)) consider the gestural moves as turns as well.

2.3.4 The transition relevance place (TRP)

This notion was introduced in Sacks, Schegloff, and Jefferson (31) and describes a projectable moment during the flow of a speech in which the recipient can potentially take the floor and become the next emitter or the present emitter may continue with another utterance. For instance, let us see the dialogue:

A: How far is the next gas station?
B: It is about fifty-five miles away.
A: Thank you.

In situations like that, when agent A is saying his interrogative sentence (how far is the next gas station?), agent B projects the possible completion of the sentence at the moment agent A utters [sta] from gas station, taking under consideration not only the content of the sentence and the common knowledge shared by both interactors but also the raising pitch of the interrogative utterance. This moment is considered the transition relevance place (TRP), the moment in which agent B can take the floor and reply to agent A. Similarly, agent A is ready to reply his thanks as soon as agent B utters [fj] from fifty-five, since the information requested is already given, concerning this specific situation.

As Levinson (37) declares, empirical evidences have shown that it takes over 600ms (see also Levelt (38)) to a person to plan and execute the shortest turn, while on average the gaps between turns are around 200ms, depending on the language (see (39), (40)). This suggest that agent B must plan his turn before agent A’s turn is finished. It also implies that agent B’s turn must mostly be connected to agent A’s turn content, specially by the sequence organization: if agent A’s turn is a question, B’s turn is expectancy an answer; if it is an invitation, an acceptance or a refusal is expected and so forth.

2.3.5 Repairs

Repair is a notion from CA that describes how interactors deal with troubles that can arise in speaking (production), hearing, and understanding (e.g. (33), (41), (42)). According to this perspective, the repair can be initiated by the interactor/emitter (self-initiated repair) or by the interactor/recipient (other-initiated repair) and can be done by the emitter or by the interactor, resulting in four possibilities for the repair system. Although CA does not accept bodily moves as an action that can trigger repairs, empirical evidences show that this kind of behavior is widely used by interactors. For instance, a recipient’s frown can indicate to the emitter that his interactor has not
listen or understand well his utterance production. In this case, some researchers will describe the recipient’s frown as an action of other-initiated repair (e.g. (43)).

2.4 The Gestural Studies

2.4.1 Nonverbal Communication Studies
In this field, communication among people is investigated by the wordless perspective. Nonverbal Communication deals only with bodily actions without considering their relation to speech. Bodily actions are examined from two main points of view. In one perspective, they are considered to be indications of inner emotional states and processes (e.g. (44)). In this perspective, a smile is predominantly an indication of an individual’s positive emotions, such as joy and amusement, even though a smile can sometimes also index other emotions associated to negative feelings (e.g. (14), (45)). The other main perspective emphasizes the social dimension of bodily actions. They can be considered as communicative acts serving interpersonal functions. In this perspective, for instance, a smile signals an intention to affiliate, while a sad face signals a request for comfort (e.g. (46)).

2.4.2 Gesture Studies
Differently from Nonverbal Communication, Gesture Studies consider bodily movements as social and communicative resources, taking under consideration the relationship between bodily actions and speech. From this perspective, gestural behaviors can add, complete, substitute, negate, illustrate, emphasize, comment, among other things what is being uttered by the agent. Researchers from this area adopt the perspective of speech and gestural actions as a single, unified process, and a flexible interactional resource (e.g. (13), (47), (48), (50), (51), (52), (53), (54)). Moreover, the recipients’ bodily actions are also closely connected with what is being uttered by his interactor, showing how participants need to construct and sustain mutual alignment in regard of the momentary unfolding of face-to-face interaction.

2.5 Grounding or Common Knowledge
In broad accordance with Clark and Brennan (55) by common ground, we mean the set of assumptions that the interaction partners share about the ongoing interaction. These assumptions may concern objects and actions as well as the interaction partner’s understanding of the situation and the communicative goals. In certain sense, common ground is also a subject dealt by other theories; such as Affect Control Theory (ACT) (22), (23); but with a different name. See articles (25) and (24) for an implementation of ACT using a Bayesian model that deals with common interpretation or identities representation.

As said earlier, an agent not only must be able to use Theory of Mind but also must create a plan with a sequence of signs in order to communicate an idea. Thus, the base of signs and communication rules must be compatible with those used by the interacting agent. In other words, agents who communicate face-to-face must share a common knowledge as well as they must foresee what such signs may cause in the other’s agent mind. The concept of having common knowledge is called “grounding” in Linguistics.

For instance, suppose an agent only speaks Hungarian language and another agent only speaks the Kikamba language. It may be impossible direct communication between them since they do not share the basics for a verbal conversation. There is always the resource of gesture communication, but gestures may also generate signs that do not have the same meaning for both. Misunderstandings are easy to rise in such situations. Some gestures, such as asking for food or water, may be easy to understand, because humans share at least some bodily gestures that are directly related to physical actions we carry out to satisfy basic needs.

Let us suppose an agent A interacting with an agent B. The knowledge an agent holds about the other should grow as they interact in several situations. For example, A noted that agent B responds with a nervous smile when he hears a reprimand speech. A must realize throughout the progress of
interaction that this action of $B$ is not a sarcastic reaction, but only the way this agent reacts when reprimanded. Now, suppose $A$ noticed that an agent $C$ always wrinkles his forehead when he hears a reprimand speech. $A$ must realize during the interaction that reaction is not a threat, but only the way $C$ reacts when reprimanded. Probably, not always these agents will react the same way, the reactions are context dependent. This means that agents must create and maintain knowledge bases for each agent they interact with. A stable common knowledge allows agents to predict roughly what certain signs (or sequence of signs) may cause in other's agents mind. Thus, individualizing adaptive knowledge bases between agents may be important. It may help agents to avoid generating sequences of signs that will be mistakenly perceived, not understood, and not comprehended.

2.6 Components of multimodal communication

This is a summary of all the components that appear in a multimodal face-to-face communication. We will show the components as items, with a brief discussion of the modality and an example. It will not be possible to delve into the discussion of each modality or the intersection between modalities. Possibly with few exceptions, animals communicate a state of the world (or mental state) of the now, be it an imminent danger, an internal state like fear, anger, or even a warning of an action that is in execution. On the other hand, by using natural languages (NL) humans insert in the message temporal relations, description, of causal relations, as well as imaginative thoughts. NL is composed of segmental signs (information packets) that are recursively used (e.g. phonemes, letters, gestures, etc.). Because it is composed of recurrent signs, languages require a set of rules to organize what is expressed. One must separate and organize the subject, the verb, and the predicate. Firstly, natural languages require that speakers have a prior agreement on the meaning of the signs, and this requires that the speakers have a shared universe, that is, they know the same objects, affects, rules of language operation, etc. Moreover, NL also requires the possibility of temporal and spatial organization of signs (e.g. being able to speak of "where", "when", and sequences expressing "how"). In addition, NL requires the ability to express abstractions, thoughts, feelings and internal states, whether they are imagined or real. Finally, the NL requires a certain degree of self-awareness, that is, the agent must be aware that he is one of the participating elements of the interaction.

We use several distinct channels to create signs that transmit messages in NL. We can classify the modalities as "oral", "verbal", and "gestural".

2.6.1 Oral (or vocal) channel (or modality)

In the oral channel we use:
a) Sounds in the form of phonemes, interjections, markers, etc.;
b) Volume of sound (the intensity of sound generated);
c) Rhythm (pitch, or tempo, whether accelerated or slow);
d) Intonational curve (if ascending is a question, if descending is exclamation, or being stable);
e) Tonality (acute or severe - for example, the louder an emitter speaks the more shows that he is irritated or altered);
f) Sociocultural characteristics (such as gender, age, regionalism, educational level, accent, etc.);
g) Qualities of the voice (if you are tired, excited, sad, nervous etc.)
The set of some of these components forms also is called prosody or paralanguage.

2.6.2 Verbal channel (or modality)

In verbal mode we use:
a) Phoneme (minimum units of meaning, used to create more complex signs such as words, e.g.: P, B, U, etc. (e.g. English has 44 phonemes);
b) Morphology (joins the phonemes to form a more complex unit – e.g. a word). At this level we define what is prefix, infix, suffix, radicals, among others;
c) Syntax (set of rules of use and grammatical description of the language). Syntax rules the order of words for meaning determination. Here we define the role of words in sentences: who does the action (subject), what action is / was / will be made (verb), in addition to complements such as objects, adverbs, adjectives, etc. (The canon for English is SVO subject-verb-object);

d) Semantics (how the sentence is / should be interpreted by the recipient).

2.6.3 Gestural channel (or modality)

Here we consider the bodily actions that are visible and have communicative meaning, i.e., we are excluding the physiological bodily actions that do not contribute to the ongoing communicative purpose. For instance, if an agent blinks because there was a sudden change in light intensity, and such action do not contribute to the progress of the interaction, we consider it is only a physiological reaction of the agent’s body. We will separate the gestural mode in different classes of bodily actions derived from distinct parts of the body, which are face, limbs, and body as a whole. To generate the gestural mode, we use:

a) facial expressions and facial actions – facial expressions communicate our inner state in the means of basic emotions (e.g. fear, joy, sadness, surprise, anger, contempt and disgust). In addition to these, several facial actions (movement of eyebrows, nose, jaw and lips) are performed by interactors during conversational interaction in order to communicate something to others;

b) Eye gaze -- in a broad sense, the eye gaze is used to show a shared attention. For instance, when a person looks to his interactor, he is demonstrating that his attention is on the other person. In the sequence, when he turns his looks to an object, he is indicating to his interactor that both should share the attention towards that object;

c) Signature actions (manual gesture) - Hands are frequently used in communication and have a strong impact on what we communicate. Fingers point things, hands make gestures of protection, gestures that show nervousness, gestures that can complete an idea that is being uttered among other;

d) Body actions (body postures) - Also involved in human communication are body postures, head movements, trunk and shoulder movements, legs and feet movements, as well as distancing between the interlocutors.

It should be noted here that sign languages (SL) are natural languages that use the three modalities described above, using only image signs, not vocalized signs. In other words, sign languages do not use auditory resources, but construct complex meaning, using oral, verbal, and gestural channels throughout visible body resources. For example, in sign language a signal with a brow may represent a question; a duplicated sign (repeated sign in sequence) can act as an enhancer, a role like prosody; the intensity at which a manual gesture is performed may represent a feeling of anger. These are important information for those who intend to create interfaces for natural language communication with machines using SL.

3. Materials and methods

We have studied face-to-face conversations taking into account multimodal aspects of the interactions. We have video-recorded and (frame by frame) analyzed several interactions observing the theoretical aspects described in the previous section (11), (10). The way we have conducted our research is described below.

3.1 Participants

Twenty-one participants from different cities in Brazil took part in this research. All participants were native Brazilian-Portuguese speakers and have different gender, age, educational and professional background. Participants were videotaped in face-to-face conversation pairs. All of them signed the Consent Form before being recorded.
3.2 Materials and procedures
Thirty video recorded dialogues were analyzed for this research. Most of the dialogues were produced while the participants were seated facing each other at a comfortable conversational distance. The participants talked for approximately 30 min. In some recorded dialogues, the participants choose their own topic of conversation; in others, the participants were told to talk about a specific subject or to perform a joint task. Each dyad was recorded with five cameras: two for participants’ close-ups, two for participants’ frontal medium shots, and one for both participants’ lateral wide-angle shot.

3.3 Data analysis
We have used the software ELAN (https://tla.mpi.nl/tools/tla-tools/elan/) to transcribe the speech and to code the bodily action from the participants. We have transcribed the participants’ speech based on the notion of intonational unit (IU), which corresponds to a verbal content plus prosody (see (56), (57)), and some procedures described by Jefferson (58). Oriented by controlled vocabularies proposed by LLIC (Language, Interaction and Cognition Laboratory – USP – Brazil), schemes for head movement, facial actions (eyebrow, eyelid, and mouth), shoulder movement, body torque and manual gesture were continuously coded for both participants of the dialogue.

4. Results
By promoting a qualitative analysis on the video streams at a micro level, five video shots precisely synchronized for each conversation, we observe some regularities that are firstly noticed in situations of answers of disagreement (in terms of CA concepts).

4.1 The double-loop feedback
After thorough analysis of various human interactions, we can conclude that a face-to-face interaction between two agents will be complete only when a double feedback loop is established. The emitter is uttering not only the sentences for the recipient, but also is producing a series of multimodal signs, either by voice intonation, by facial expressions, by body posture, by manual gestures, etc. On the other hand, the recipient is also emitting speech/vocal signals, as well as a series of signs through facial expressions, gestures, posture, etc. Thus, we can certainly say that both are simultaneously emitters and recipients of signs during an interaction. Moreover, we can say that both are interpreters of the various multimodal signs emitted by the agent with whom they are interacting. Descriptions from a linguistic point of view are more detailed in (10).

4.2 Timing and Repair Situations
Paying attention to contexts of disagreement, we observe how recipient’s bodily action, and/or his vocal disagreement are firstly displayed during the emitter’s turn. We notice that this kind of recipient’s behavior promotes an opportunity for emitter to rephrase his speech attempting to obtain the preferred response. We also notice that the recipient’s disagreement display, being this display by vocal and/or gestural modalities, is produced at a specific moment (the moment of a possible turn competition - TRP) comprised into the emitter’s utterance that has elicited the recipient’s disagreement. Observing cases of disagreement situations (dispreferred answers), based on CA perspective (see topic 2.3.2) and of formulation of repairs (see topic 2.3.5), we can conclude that the time in which recipient’s feedback occurs is too important to be ignored. When the recipient’s feedback associated to negative behavior (e.g. a frown) is produced around the TRP of the emitter’s utterance, there is a tendency for the emitter to perform a ‘fix’ (a repair) or a ‘shift’ on his message. This is best understood by the following dialogue (Table 1). Let us consider an example about how recipient’s bodily action seems to interfere in the production of the emitter’s speech. This fragment of face-to-
face interaction shows two women who have just met and who were asked to perform the task of proposing a decoration for a living room.

At the beginning they try to agree about the type of room they might decorate (lines 1-8). Then, B proposes a hanger (line 9). A pause (line 10) occurs during which A produces some facial actions and questions about the word used by B (line 10), showing that a repair is needed. B provides the repair (lines 12 and 14), explaining what she meant by the word. During her verbal production, she pretends to be hanging a coat at a hook. B starts her bodily actions as soon as she finishes producing you know? (line 12). Interactor A starts proposing an understand (line 13) but holds herself until B finishes her repair. Then A shows her understanding of B’s proposition (ahn:: a rack – line 15). B initiates a self-repair, telling A that the rack can be the type of a hanger on the wall (lines 20 and 22-23). Taking only the verbal production, we understand that B’s produces a self-repair. However, the video shows that during the agreement between them (lines 16-19), A looks at the wall and points to it. As soon as she did that, B tries to rephrase her proposal, producing the apparently self-repair. She initiates her utterance with but (line 20), demonstrating that she probably had another type of hanger on her mind (a hat stand) and not a hanger on the wall, but accepts A’s understanding/idea about this piece of furniture. On the other hand, A also produces a self-repair, changing her initial understanding (a rack – line 15) to a comprehension about what B wants as a hanger (hat stand – line 21). All the adjustments of the emitters have been fired by the recipients’ bodily action during the emitter’s turn. One can argue that it is just a coincidence; however, in our analysis we note that this type of interactors’ behavior is recurrent. The change of emitters’ behavior shows that interactors search for preferred replies (agreement displays) all the time during the interaction. In this example, both interactors were trying to search the preferred reply for their plans: A seems to have in mind another type of decoration for the front hall; nevertheless, she tries to understand B’s

| Table 1: Transcription of an excerpt from an interaction between two participants |
|---------------------------------------------------------------|
| Brazillian-Portuguese | English |
|------------------------|---------|
| 1 A: bom aqui a gen-gen- a gente tem um hall:: né? | well, here it it’s a hall::, isn’t it? |
| 2 B: tá hã | yeah, hã |
| 3 de entrada, [né]? | front hall, [isn’t it]? |
| 4 A: [é ] | [yes ] |
| 5 aparentemente | seemingly |
| 6 um hall e us- a sala | a hall and a- the living room |
| 7 B: [t á ] | [yea h ] |
| 8 A: [n é ]? | [aren’t they]? |
| 9 B: acho que a gente precisa ter um cabide | I think we should have a hanger |
| 10 (.) | (.) |
| 11 A: cabide? | hanger? |
| 12 B: sabe [pra:: ] | you know? [to:: ] |
| 13 A: [‘cabid-’]? | [‘rack-’]? |
| 14 B: por casaco | hang coats |
| 15 A: ah:: um cabideiro | ahh:: a rack |
| 16 [tá ] | [yeah] |
| 17 B: [êh], sim | [yeah] yes |
| 18 A: tá | yeah |
| 19 (.) | (.) |
| 20 B: mas po[de ser ] | but it [ can be ] |
| 21 A: [cabideiro] | [hat stand] |
| 22 B: pode ser na parede | it can be on the wall |
| 23 não precisa ser no chão | it doesn’t need to be those that stand |
| 24 porque ocupa menos espaço do chão | because it can spare floor space |
| 25 A: tá:: êh | yeah:: yeah |
| 26 pratelei::ras::?: | shel::yes::? |
| 27 cab[deiros ou prateleiras e gavetei]ros= =com- | rack or shelves and draw[ers]= |
| 28 B: [ai::, nossa prateleira ] | [wow, yes shelf ] |
proposal and adjusts her plan (shelves with drawers on the wall - line 27) until B accepts her proposal (line 28). In her turn, B seems to have a plan of a hat stand, tries to adjust her plan to a hanger on the wall, and, finally gives it up to accept A’s proposal of a shelf. Taking this kind of regularities as parameters, we have investigated if recipient also displays the same behavior on contexts of agreement (preferred answers). We have observed the same regularities. Recipients also produce bodily actions and/or vocal agreement displays at the moment of a possible turn completion. From such observations, we have inferred a general practice and proposed the algorithm for face-to-face interaction.

4.3 Gesture and Repairs

Definitely, facial expressions and gestures interfere with the execution of a communication plan. Let us consider that an agent created a plan and committed to the execution of that plan; that is, the agent’s intention is to communicate something through a sequence of sentences, intonations, gestures and postures. According to the above postulate, as soon as a double feedback loop is established, to the extent that one agent is issuing an utterance it is also evaluating the reactions of the other agent. Without going into too many examples, we all have experienced to talk to someone who shows no interest in what we are saying. How do we perceive that? We also have seen someone to make a facial expression clearly showing not to like something we have just said. Or someone that has interrupted our speech by a manual gesture, or even turned his head and shown interested on something happening nearby. In any of these situations, the emitter may repair using one or more signs to achieve the goal of his communicational act: to create a mental state in the recipient that clearly demonstrates that the message has been understood. Note that this has nothing to do with the recipient “to agree” with or “to refuse” the idea contained in the message (see discussion in section 6.3). We are only pointing to the fact that the emitter repairs its communication plan with the sole objective of sending the message clearly and most of time completely. In sum, we mean that the agent repairs a communication plan with the intention of being fully and well understood.

5. Algorithm

In order to give us an intuition, let us firstly exemplify how the face-to-face algorithm occurs. Consider that a human being (agent H) is talking to a machine (agent M), let’s consider it a humanoid robot capable of performing human-like gestures. The robot first presents itself: “Hello, I am the agent M”. Following the theory of Conversation Analysis (CA), an adjacent pair must occur at this point; e.g., to an initial presentation, H should preferably respond with a greeting: “Nice to meet you, I’m H”; “Hi, I’m H”; “I’m H”; or any other form of greeting. Now, suppose that M must continue the conversation. According to CA theory, M will take the turn, that is, it will be M’s turn to execute a whole communication sequence that we call communicational action. In this sense, communicational action refers to all verbal and gestural content that will be communicated within one turn. Naturally, M must mentally devise a “plan” concerning which phrase, signs, emphasis, intonation, gestures, facial expressions, etc. will be carried out during the next communicational action. In such plan, M divides the full communicational action into Intonational Units (IU). According to CA theory, IU are the speech flow that occurs with a single prosodic contour. Distinctly, we consider IU associated to bodily actions (facial expressions, head movements, manual gestures, posture, etc.), so we use the term “Communicational Unit” (Cu) for each segment. In summary, M must quickly draw a plan on what and how to perform the next communicational action; then it must divide such plan into segments. Each segment (or unities) contains chunks of verbal communication and prosody associated to gestures and visible bodily actions, which will all be performed together. Each plan will transmit only one utterance, so to compose an idea or a speech with more than one utterance the agent must compose as many communicational action as the idea requires. Sometimes an idea requires multiples sentences within the same turn.
People have in common sense that, when one agent takes the turn, it becomes the "speaker" and the other becomes the "listener." Thus, in the case proposed above, M would be the emitter (active) while H would be the recipient (passive) during the period in which M has the turn and tries to perform its communicational action. However, as described in section 4.1, that is not what we have observed in our data. Throughout the time M performs its communicational action two feedback loops are established in the face-to-face interaction. Firstly, agent H hears and perceives the visible bodily actions performed by M. During this time, H interprets both, the verbal communication (what is said) associated to the prosody, as well as all kinds of gestural signs that M can perform (facial actions, head movements, hand gestures and body posture). Moreover, during the time H interprets M's communicational action, H generates feedback, which occurs not only in the form of vocal markers (such as hum, ah, oh, uh ...) or verbal content (such as well, what?) but also in the form of visible bodily actions (facial expressions, facial actions, hand gestures, head movement, etc.). Therefore, the "recipient" is an active agent, not a passive one. The second feedback loop is formed because upon perceiving such signals, M itself repairs its initial communication plane. Note that H can explicitly perform a repair on something expressed by M, but because M is interpreting H's feedback, M can perform self-repair on its initial plane.

It should be noticed that the process of conversation and face-to-face interaction is a dynamic system in which communication is hardly a simple and direct transmission of information. The process is not like a text-to-speech program reading an unchangeable ready-made phrase from a text. During face-to-face interaction, one agent can interrupt the other and perform a turn taking, for instance. Moreover, one may complement or disagree with the communicational action that the other interactor is performing even before the agent finishing his turn.

Let us consider that M was able to perform all its communicational action, so by the adjacent pair formed, H should continue with a compatible communicational action, taking the turn of the conversation. This implies H mentally should also create his communication action plan, should divide it into Cus, and then should perform each segment expressing himself verbally, with prosody, including gestures. For sure, H would expect that the machine (M) can interpret the signs he/she is emitting, as well as would expect that they both have the same shared knowledge and have contextualized the interaction.

5.1 Contextualization
After an interactive experience, agents should memorize the result, in the same sense humans memorized every interaction they perform. We filter most of the signs in our interactions, perhaps because we classify our interactors in stereotypes: “Tom makes excessive gestures while speaking”; “David is harsh in his responses”; “Anna has a nervous smile when she lies”. We may not know to explain how we do it, but we take all this information into account when we talk to other person. Such filters make our interpersonal communication more efficient. Because we use ToM, we include in our communication plan preferable signs for each one of our interactors. As a consequence, we also interpret certain signs emitted by the others as traces of their personalities – not a raw element. In this sense, we must observe that common knowledge is not just about sharing cultural or contextual knowledge. We create individual databases that help us to perform better personal communication; while we share more general contextual information that relates to a certain community in specific situations, as well as cultural information that aggregates us into a larger communities, most of the time sharing the same language, beliefs, habits, etc.

The Affect Control Theory (ACT) is a psychological social theory of interaction, which proposes that the perceptions, actions, and emotional experiences of human beings are governed by the need to minimize deviations between transient impressions and culturally established social feelings that comes from interactive situations. Such deviations can be quantified, so, this theory has great appeal for application in AI (22). ACT is a theory that explains the main idea described above. In few words: we create social dictionaries of meanings, affects and sentiments based on our interactions. The role one interactor plays in certain situation depends on an institutional context. For instance, your boss
or your teacher must play a hierarchical role in the office/classroom but he may play distinct role when playing soccer with his employees/students. These situations result quite different forms of interaction. Following the above reasoning, for each interaction, interactors not only contribute in constructing a common social dictionary of meanings, actions, sentiments, but also create a dataset of how a specific interactor normally responds, acts, reacts, interprets and reveals sentiments, etc. ACT is a good theory for creating subsystems for dealing with these components on interactions; however it is out of the scope of this article to go further on this subject (see (25)).

Considering what has been explained above, in order to have a face-to-face communication experience more alike to that of humans, we believe that machines also need to have a mechanism not only to interpret speech combined to bodily actions, but also to memorize individual interactions, which may generate an individualized common grounding knowledge.

Considering this, a context is a situation in which agents observe: (a) the institutional situation in which the interactors are immerse, (b) their roles and hierarchical positions in the interaction, (c) the appropriate cultural dictionary to be used in such situation, (d) the individual database about the other interactor. Thus, agents should restore how the other interactor acts, his/her typical gestures, intonations, lexical preferences, as well as several other details that form the other interactor’s personality. The overcome of the cultural aspects over the agent’s personality is strongly governed by the institutional situation. A formal or informal interaction is pre-established at the beginning of the interaction, as well as the role the interactors play.

5.2 The Formal Algorithm

The algorithm is formalized in an abstract way, so it will be possible to implement it partially or with more details than those described here. For simplicity, let us call $\xi$ the emitter agent and $\mathcal{R}$ the recipient agent. These roles will change frequently every time a turn change occurs. Still for simplicity, let us call mind a system composed by a set of complex modules that process the information the agents are receiving or emitting. Let us describe this system as modular as possible, so we can separate the computational tasks. Initially, the agents are in a context and must share some knowledge. This means that as soon as the agents are placed face-to-face, computational modules retrieve data and update the context. We may represent the data in the interactor’s minds as:

$$\zeta_\xi = \{ A_3, B_3 \}$$
$$\zeta_\mathcal{R} = \{ A_3, B_3 \}$$
$$K_\xi = \{ A_3, B_3 \}$$
$$K_\mathcal{R} = \{ A_3, B_3 \}$$

where $\zeta_\xi$ is a tuple with information about the initial context for the agent $\xi$, and $\zeta_\mathcal{R}$ is a similar tuple for the agent $\mathcal{R}$. Similarly, $K_\xi$ is a tuple with the shared knowledge the agent $\xi$ has concerning the agent $\mathcal{R}$, and $K_\mathcal{R}$ is a similar tuple with the knowledge that the agent $\mathcal{R}$ shares with respect to the agent $\xi$. Note that $\zeta_\xi$ and $K_\xi$ are processed on the mind of the agent $\xi$, and they are not equal to $\zeta_\mathcal{R}$ and $K_\mathcal{R}$, which are processed on the mind of the agent $\mathcal{R}$. Note also that it does not matter whether $\xi$ and $\mathcal{R}$ have ever met each other. The tuples $K_\xi$ and $K_\mathcal{R}$ are updated during (and after) the interaction to include or to delete data concerning the interactor. Thus, the agents can start an interaction only with the minimum grounding knowledge: the capability of exchange information using a shared language, as well as the cultural meaning of some gestures within that language.

5.2.1 The Agent’s Intention

Consider that an agent playing $\xi$ starts the interaction, so, it takes the first turn. Remember that, before starting, $\xi$ must create a communication plan that contains what it intends to communicate as well as how it intends to do that. The plan is composed by a sequence of verbal, intonational, facial, gestural signs. Let us write the plan formally as:

$$P_{\text{tx}}^\xi = \{ \{ \text{Cu}_1, \#_1 \}, \{ \text{Cu}_2, \#_2 \}, \{ \text{Cu}_3, \#_3 \}, ..., \{ \text{Cu}_n, \#_n \} \}$$
where $\mathcal{P}_{t_0}^\xi$ is a tuple representing the communication plan the agent $\xi$ creates for a turn $t_0$. Such tuple is a set of $z$ Communicational Unities ($Cu_1 \ldots Cu_z$). As said, $Cu$ is a cell in which $\xi$ associates phonemes/words with intonation and multimodal signs. Thus, a $Cu$ is a chunk of information the emitter performs expecting that the recipient interprets it as a meaningful unity. We consider that $\xi$ must compromise itself with the execution of each plan $\mathcal{P}$; hence, $\xi$ has intentions of fully executing all communication plans.

Human interactors take feedbacks into account to guarantee that the communication is correctly interpreted; thus, agents should associate each $Cu$ to certain stereotyped feedbacks ($\xi_1 \ldots \xi_z$). Hence, during its turn $\xi$ executes each $Cu_k$ expecting certain sound or visual feedbacks $\xi_k$. Each communicational unity is a tuple composed by several sequential elements, which we separate here into six main classes:

$$Cu_k = \{v_k, \mathcal{P}_k, e_k, b_k, \text{ln}_k, \text{lw}_k\}$$

where $Cu_k$ is the tuple $\xi$ has planned or executed during the turn “$t_0$” at the sub-unity $k$. As part of a communication plan, $Cu_k$ is compounded by: $v_k$ the verbal components made of phonemes, words, sentences grammatically suitable for the chosen language, etc. The component $\mathcal{P}_k$ is the prosody elements to be used during the verbalization, $e_k$ is the facial expression the agent intends to do, $b_k$ is the gesture the agent intend to perform with its hands and its superior members, $\text{ln}_k$ represents the head and shoulders movements, and $\text{lw}_k$ is the body posture the agent intend to apply.

Each expected feedback unity is a tuple that may be composed by several sequential components, which we separate here into six main classes:

$$\xi_k = \{s_k, \mathcal{P}_k, e_k, b_k, \text{ln}_k, \text{lw}_k\}$$

where $\xi_k$ is the tuple representing a set of feedback that $\xi$ may expect during the turn “$t_0$” at the sub-unity $k$. $\xi_k$ is part of an intentionality plan, thus, $s_k$ is the vocal and/or verbal feedback expected during this sub-unity, also called markers, which include sounds such as “hum, ok, oh, ah”, among others and/or words or expressions such “well, ok, no, yes, yeah, that’s it”, among other. The component $\mathcal{P}_k$ is related to the prosody associated to markers, for instance, a question marker such as “ahn?”. The $e_k$ represents facial expressions the agent may expect to interpret during the current $Cu_k$, as well as the feedbacks $b_k$, $\text{ln}_k$, and $\text{lw}_k$, which are related to gestural, head movements, and bodily postures respectively.

It is important to note that expected feedbacks are also context dependent, and agents must consider different types of feedback they may receive from distinct interactors. But why should an agent include in its communication plan all kinds of interactors’ feedback? Because we use communication protocols that we follow strictly for a variety of reasons: to communicate more efficiently, to be polite, to know if we have been correctly interpreted, among other reasons. Then, we use several channels to make sure that each fragment we want to communicate is correctly interpreted. If we detect any discrepancy, we make repairs to our plan. For this reason, $\xi$ expects feedback throughout the time. As $\xi$ executes the communication plan, it makes repairs to the communicational units and, sometimes, to the whole communication plan. Such reformulation can be trigged by the emitter $\xi$ itself or by the agent $\Re$’s feedback (see section 2.3.5).

So far, we considered that $\xi$ has created a plan $\mathcal{P}_{t_0}^\xi$ for the turn $t_0$, and it is going to execute it step-by-step. During the time in which $\xi$ executes its turn, the agent $\Re$ not only interprets all signs emitted by $\xi$ but also creates its own communication plan in real time to answer $\xi$ next. Moreover, $\Re$ also provides vocal/verbal and gestural feedbacks to ensure the interaction fluency. The agent’s $\Re$ plan has the same format as the emitter’s plan:

$$\mathcal{P}_{t+1}^\Re = \{\{Cu_1, \xi_1\}, \{Cu_2, \xi_2\}, \{Cu_3, \xi_3\}, \ldots, \{Cu_z, \xi_z\}\}$$

Suppose now that the agent $\xi$ starts executing the plan $\mathcal{P}_{t_0}^\xi$ for the first turn “$t_0$”. As said, the agent ought to vocalize sounds, to apply correct prosody to them, to perform facial expressions, gestures, head movements, and body posture. Concurrently, the $\xi$’s input modules must process feedbacks and interpret sounds, facial expressions, gestures, head movements, and body posture coming from the other interactor. Fortunately, the same algorithm can be used for the two roles. To change from
one state to another is only a matter of taking the turn. In the next section we describe a general
pseudocode in format of an algorithm showing how an agent may play both roles.
Note that we use tuples in the above reasoning to describe possible states of internal variables in the
agent’s mind. We do not prescribe a format for representing data or states of the world, neither a
theory to be applied for processing such information nor which technology can better perform such
processing. We suggested that ACT can be a theory to explain contexts and hierarchical interactions,
but other theories can be used. Likewise, each module in the algorithm can use different data type,
models or tools; for instance, Bayesian networks, Markov chains, artificial neural networks, fuzzy
logic, deep learning, or hybrid technologies.

5.2.2 The Algorithm Kernel
This algorithm describes the top level of an intricate information processing system. Every module
referred next is a complex computational system by itself. For example, we referred to a facial
expression module simply by triggering a module, however, image recognition itself is a quite
complex task of computer vision. Another example is a module that outputs sound with prosody,
which would be implemented as a text-to-speech engine with intonation and other vocal attributes,
which is not perfectly developed till the current days. In the pseudocode for the algorithm kernel,
consider that not necessarily the lines are executed in sequence as a normal computer program.
Instead, these lines describe more a loop that triggers several tasks for parallel execution.

In this pseudocode, consider: a=“agree”; d=“disagree”; t=“turn”; c=“change_turn”
V=or logical; ∧=and logical

01 while (interacting) // while the agents are interacting
02 φ = context[ ]; // phi holds parameters of the context
03 if (my.turn==1) // == my turn? (play the emitter)
04 z=length(P([]t)); // z tracks the size of the plan
05 for k=1 to z // for all items of the communication plan
06 out_Verbal(ʻ{k, φ}, p![{k, φ}]); // out planned ʻ{k} verbal + prosody
07 out_FacialExpression(ʻ{k, φ}); // out planned ʻ{k} facial expressions
08 out_Gesture(ʻ{k, φ}); // out planned ʻ{k} manual gesture
09 out_HeadMove(ʻ{k, φ}); // out planned ʻ{k} head/shoulde r movs
10 out_BodyPosture(ʻ{k, φ}); // out planned ʻ{k} body posture
// get feedbacks, interpret them, then change or not the planned Cu(k)
11 v = in_Verbal(P); // listen to verbal feedback
12 e’ = lang_meaning(k, φ, e’); // interpret them
13 p = in_Prosody(P); // listen to prosody feedback
14 p’ = prosody_meaning(k, φ, p’); // interpret them
15 e = in_Facial(P); // get facial images
16 e’ = facial_meaning(k, φ, e’); // interpret as facial expressions
17 g = in_Gesture(P); // get manual gesture images
18 g’ = gesture_meaning(k, φ, g’); // interpret them as gestured signs
19 η = in_HeadMovements(P); // get head and shoulders images
20 η’ = headMov_meaning(k, η, η’); // interpret them as gestured signs
21 d = in_BodyPosture(P); // get body images
22 d’ = posture_meaning(k, η, d’); // interpret them as body postures
23 (v, p, e’, g’, η, d’) = semantics(e, p, e’, g’, η, d’); // is feedback AP?
24 π = pref_dispreference(ʻv, p, e’, g’, η, d’); // is fdbk P/D?
25 // if some feedback is interpreted as a disagreement:
26 if (v=“disagree” v=“agree” v=“turn” v=“change_turn”)
27 call_my_Repair(k, φ); // self-repair the plan
28 z=length(P()); // change z if the plan was realized
// on the other hand, if all feedback indicate agreement:
29 else if (v=“agree” v=“agree” v=“agree” v=“agree”)
30 continue; // just continue the plan
// but if no feedback is given by the interactor
31 else
32 call_emphasize(v, p, e, g, d, k, φ); // emphasize some signals
33 z=length(P()); // it may have resized the plan
// these codes control which piece of the plan are executed
34 if (k=x) // if the index k <= size of the plan
35 k = K+1; // just increment the index
36 else (k>x) // after reaching the end of the plan
out_ChangeTurn(); // output a sign showing “turn changing”
my.turn=0; // change its own status to interpreter
// end of emitter loop, now the pseudocode for the recipient/interpreter
else
// ** not my.turn? (play the recipient)
while (my.turn==0) // while playing the interpreter...
k = 0; // start Cu(k) counter
φt = in_Verbal({}); // listen to verbal feedback
φt′ = lang_meaning(k,φ,φt′); // interpret them
ρt = in_Prosody({}); // listen to prosodic feedback
ρt′ = prosody_meaning(k,φ,ρt′); // interpret them
e′t = in_Facial({}); // get facial images
e′t = facial_meaning(k,φ,e′t); // interpret them as facial expressions
ρt = in_Gesture({}); // get manual gesture images
ρt′ = gesture_meaning(k,φ,ρt′); // interpret them as gestured signs
A′t = in_headMovement({}); // get head and shoulders images
A′t = headMov_meaning(k,φ,A′t); // interpret them as body postures
(π′t,π′t′,e′t′,g′t′,A′t′,b′t′,π′′t,π′′t′,φ′t) = semantics((πt,πt′,e′t,g′t,A′t,b′t,πt,π′t,φt));
ai = adjacent_pair((π′t,π′t′,e′t′,g′t′,A′t′,b′t′,π′′t,π′′t′,φ′t)); // is msg P/D?
// according to its interpretation, if the agent disagrees
if (π′t==dνπt′==dνe′t==dνg′t==dνA′t==dνb′t==dνai==dνπi)==d)
out_negativeFdbk(k,φ); // output a negative feedback (repair)
k = 0; // maybe reset Cu(k) index
// if the recipient agrees with the interpreted signs
else if (π′t′==a∧ρ′t′==a∧e′t′==a∧g′t′==a∧A′t′==a∧b′t′==a∧ai==a∧PI)==a)
out_positiveFdbk(k,φ); // output a positive feedback
k = k+1; // maybe increase Cu(k) index
// but, the agent may not give any feedback
else
out_neutralFeedback(k,φ); // increase Cu(k) index
// simultaneously, create a plan \( P(P) \), it must be ready at the turn change!
(\( P)_{\Pi}=\text{planCreator}(\pi,t,\rho,t′,e′,g′,A′,b′,\phi)\);
// as the receiver interpret any sign indicating turn change
if (πt′==c∨ρ′t′==c∨e′t′==c∨g′t′==c∨A′t′==c∨b′t′==c∨ai==c∨\pi′′t==c∨\pi′′t′==c∨\pi)==d)
my.turn=1; // change the role and play ‘emitter’
// end of loop

5.2.3 Explaining the Algorithm

When engaged in a face-to-face interaction, agents execute the algorithm from line 01 to 68 of this pseudocode. The line 02 describes that both agents update the context at the beginning of the interaction. This means that externally, they seek information about the scenario and situation they are in, they evaluate whether other agents can hear or participate in the interaction, and decide which relevant information may help to organize the form and content of the conversation. In addition, internal data are retrieved for information concerning the other interactors(s), the grounding knowledge they share, culture in which they are immerse, words, phrases, and gestures that are appropriate for use in such situation. Thus, \( \Phi \) (line 02) is a module that perform intense information processing. Each agent obtains its own \( \Phi \) depending on several initial conditions. Lines 03 and 39 describe an if-then-else structure that decides the role an agent is playing. If the variable \texttt{my.turn}=1, the agent plays the emitter (\( \xi \)) and executes lines 04 to 38; otherwise, if \texttt{my.turn}=0 the agent plays the recipient (\( \mathcal{R} \)) and executes the loop from lines 40 to 68. Let us first analyze the behavior of an agent playing an emitter.

5.2.4 The emitter’s role

By performing the emitter, an agent must first be aware of the size of the communication plan to be carried out, and it is what we described in line 04. From line 05 to 35 the algorithm describes a loop that controls the execution of all \( Cu(k) \) the agent has in its communication plan. The index \( k \) controls which \( Cu(k) \) is executed at certain time. Then, the agent generates a sequence of outputs. In lines 06,
07, 08, 09, and 10 the algorithm shows that agent triggers the modules for verbal output associated to prosody, facial expressions, gestures, head movements and body postures respectively. The verbal output module may be a simple text on a screen or a complex system that emits phonemes associated to intonation and other voice attributes. The facial expression module may be a physical face of a robot, the face of a 2D- or 3D-avatar on a screen, or a human face. Likewise, the gesture module controls gestures that may be produced by humans, maybe by robot arms or by a virtual character on a screen. The same applies to head movements and body postures, whenever these are applicable.

The order of activation of these modules is not fixed. In most of the situations all modalities will occur in parallel. It depends not only on the plan being executed, but also on the feedback received, as well as the intention of what the agent wants to provoke in the other interactor's mind. Simultaneously to what the agent is performing in lines 06 to 10, the agent executes the lines 11 to 22. The line 11 means that the algorithm receives data from a verbal input module, which can be a simple chat editor that receives text, but it can also be a complex voice recognition system with syntax analyzer and grammar parser. While on line 11 the algorithm obtains verbal information, in line 12 it triggers a module for interpreting data with semantic significance for the verbal signs received from the module in line 11. In parallel, in line 13 the algorithm receives data from prosody detection module, then it sends these data to an interpreter (line 14), a module that extracts meanings on prosody components. Verbal and prosody components usually are interpreted together.

Lines 15, 17, 19, and 21 show that the algorithm is linked to modules for visual information input. In line 15 the algorithm triggers a module that extracts facial information, in line 17 it triggers a module that captures information from manual gestures, in line 19 it triggers a module that capture head and shoulders movements, and in line 21 it triggers a module that captures bodily postures and movements. Equally to the previous verbal modules, meaning must be extracted from sequences of image signs. The lines 16, 18, 20, and 22 show the respective triggers for each interpreter module.

Single quotation mark applied to the name of a variable denoting a modal sign, e.g. \( \pi \), means that these signs are firstly interpreted in a linguistic interpretation module, derived from the raw sign perceived by the sensorial system. Then signs are interpreted in semantic meaning modules, whose results are represented with double quotation mark, as in \( \nu' \), which means they are derived from signs previously (linguistically) interpreted by earlier modules. These modules (lines 15, 17, 19, 21, 16, 18, 20, and 22) must perform a complex set of consecutive information processing and intense inter-process communication. For instance, an agent states "panda is a cool word. Do you know what it means?". To make sense, the adjacent pair (AP - section 2.3.1) should be an answer because a question was made. Besides, the other interactor should not answer "It is a pronoun", unless the interactor’s intention is to make the asker to burst into laughter. Therefore, agents must keep track and recall verbalizations (and gestures) occurring throughout the current interaction. It means that they must connect and contextualize the communication plan they are performing with what was said and expressed in previous turns and former Cus. In order to keep track of the AP as well as whether preferred or dispreferred responses (PD - section 2.3.2) have occurred, the algorithm has two modules: adjacent_pair() and pref_dismatchment (). The variable \( \alpha_k \) holds data for AP while \( \pi_k \) holds data for PD. \( \alpha_k \) holds answers for questions such as “was this response compatible with the expected AP?”,”how to fix a deviation from an expected response for the AP?” etc. Similarly, \( \pi_k \) holds answers for questions such as “was the last response a P or a D one?”, “is the last response relevant?”, “how to fix deviations from an expected response for the PD?” etc. Distinct implementations of the algorithm may require other questions for both AP and PD depending on what the designer intends to model.

In line 23, the algorithm calls a function that integrates the interpreted signs till the \( k^{th} \) iteration. This integrator compounds verbal, visual signs, AP and PD responses and contextual information to jointly obtain a meaning for such components. Three arguments are included in the function: \( \alpha_{k-1}, \pi_{k-1} \), and \( \Phi \). The argument \( \Phi \) defines the context. Note that the modules adjacent_pair() updates the value of
\( \alpha_k \) (for \( k \)th iteration) only in line 24, but its contents \( \alpha_{k+1} \) is used in line 23. The same occurs with \( \text{pref_dispreference}() \) that updates the value of \( \tau_k \) in line 25, but its contents \( \tau_{k+1} \), is also used as argument for the meaning integrator. Both modules (lines 24 and 25) update their respective data after integrating signs, performed in line 23.

Let us see an example: to smile and to wave hands may mean a greeting message, such as "hello, I'm here". A verbalization such as "humm" associated to a lateral head movement may mean a negation in certain cultures, but it may mean something else in other context, like a mockery. A formal situation in which a particular adjacent pair is expected, or a situation that some peculiar gesture is expected as preferred response can be determinant for the success or a fully failure when engaging in face-to-face interactive situations. That is why it is important to have the modules AP (line 24) and PD (line 25) always tracking the interaction steps, but it makes sense to update them only after interpreting the meaning of the chunk of information currently received. One can only determine if an adjacent pair is consistent or if a preferred response occurred after fully interpreting the meaning of the \( \text{Cu}_k \) currently received.

Back to the meaning integrator (line 23), we may say that the result of the signs integration is unique for each agent. The database used for integrating meanings may be something like a look-up-table (LUT) that the agent updates all the time. Such LUT depends on the experiences the agent had passed through its life. In addition, every interpreted element must update such LUTs whenever any sign association generates new meanings. In this sense, the function of the signs integration module is to query this database and associate the current \( \text{Cu}_k \) with information concerning the interactor, stereotyped behaviors, expected AP and PD responses, as well as with data describing the current context.

In lines 26, 29, and 31 the algorithm contains another decision-making structure. After the result from integrator module, the agent checks if feedback signs obtained \( \text{disagree} \) (line 26), \( \text{agree} \) (line 29) or are \( \text{neutral} \) (line 31) concerning some expected feedbacks. The correct understanding about how we use these terms will be discussed later (section 6.3). As said, the recipient (\( \Re \)) may realize verbal and/or visual feedbacks while the emitter is performing the current \( \text{Cu}_k \). If \( \xi \) interprets feedback signs of disagreement in any of the channels, verbal (\( \varphi^"k=d \)), prosodic (\( \varphi^"k=d \)), facial (\( \psi^"k=d \)), gestural (\( \psi^"k=d \)), head movements (\( \vartheta^"k=d \)), or postural (\( \delta^"k=d \)); then \( \xi \) probably triggers a routine to repair the situation (line 27). Linguists call this type of repair a self-repair (section 2.3.5). A self-repair can be characterized by correction in phonemes, intonations, facial expressions, gestures, as well as correction of words, of a whole utterance, of a sequence of facial expressions or gestures. Possibly, in situations that require critical repairs, a self-repair may imply changes on the entire communication plan \( \Pi_{k+1}^{\xi} \). A self-repair involves the modification of the \( \text{Cu}_k \); hence, the variable that controls the end of the turn may need update, as shown in line 28. In fact, agents may decide not to perform a self-repair even receiving feedbacks that show the communication is not causing the desired effect. Agents may ignore such signs and proceed with its initial communication plan due to many reasons, for instance, \( \xi \) may be confident that \( \Re \) will understand the message after fully completing the original communication plan.

On the other hand, if the current plan has generated the expected effect in the \( \Re \)'s mind, the perceived feedbacks should match with those expected by \( \xi \). In line 29, the algorithm tests if \( \Re \) has agreed with all \( \text{Cu}_k \) received till the \( k \)th iteration. It tests feedback signs of the channels verbal (\( \varphi^"k=a \)), prosodic (\( \varphi^"k=a \)), facial (\( \psi^"k=a \)), gestural (\( \psi^"k=a \)), head movements (\( \vartheta^"k=a \)), and postural signs (\( \delta^"k=a \)). It can be understood as test if \( \Re \) has understood the sequence of \( \text{Cu}_k \) till that point. If all feedbacks agree to the expected ones, \( \xi \) probably just continues the communication plan, as shown in line 30.

The algorithm may also execute lines 31, 32 and 33 when \( \xi \) has received neutral feedbacks or no feedback at all. In these cases, \( \xi \) may decide to emphasize aspects of its communication plan by modifying multimodal signals, or it may change sentences, etc. It means to modify the initial communication plan in the same way described earlier. However, due to uncountable reasons, \( \xi \) may also decide to keep the original plan and ignore the recipient's indifference. Modifications may imply
changes to the size of the communication plane, thus, in line 33 the algorithm revises the size of the plan.

To complete the explanation of how an agent performs the role of emitter, each time $\xi$ performs a $C_U$, the algorithm increments the index $k$. Then, the algorithm tests if $k=2$; which means to test if the communication plan reached the end (lines 34, 35, and 36). If it is the case, the algorithm calls a function in which $\xi$ forwards a hint to $\mathcal{R}$ showing that it is passing the turn to the other interactor (line 37), then $\xi$ changes its own status to recipient (line 38).

Let us consider that the protocol works normally, then, the agent $\mathcal{R}$, who performs now the recipient, will take the turn and will assume the role of emitter, vice versa. However, when it does not occur, agent $\xi$ may wait a while and tries to resume the conversation. Such situations may result in embarrassments that should be predicted in man-machine interactions. Such situations are not described in this algorithm to not extend the article.

As seen earlier in CA theory, humans seek for transition relevance places (TRP, section 2.3.4) for dynamically changing the turns in conversations. This component was not foreseen in the present algorithm. First, because it is an improvement in the performance of the model that can be implemented later. Second, because it is a subtlety that can be modeled in several ways in the algorithm, to the liking of the designer. Third, because this component may be learned to the extent that an agent uses the language interactively. In fact, it may be a perpetual learning process for each one the agent is interacting with. Thus, TRP is for future implementation for both, the emitter and the recipient role.

5.2.5 The recipient’s role

In the algorithm, the role of recipient ($\mathcal{R}$) is described from line 40 to 68. The perceptive and interpretative modules, triggered from lines 42 to 53, have been described previously for lines 11 to 22. These are the same modules that receive and interpret sounds and visual signs. In line 54, the algorithm triggers the meaning integrator module, which also perform similar function to the module described for line 23. As in the emitter role, the meaning integrator module uses $\alpha_{k,1}$, $\pi_{k,1}$, and $\phi$ arguments, and it works in the same way. Similarly, the recipient keeps track on adjacent pairs (line 55) and evaluates the preferred or dispreferred actions and responses (line 56).

There are minor operational differences between the receiver and the interpreter modules, as well as in the meaning integrator module for the recipient role. Their functions are to receive and interpret signs, with differences only in attention and in the information processing given to the chain of signs. A distinction can be that the recipient must combine all signs and then to recompose each $C_U$ sent by emitter for then to extract meaning; differently of the emitter that, while carrying out its communication plan, uses the receptor and interpreter modules focused on detecting feedbacks. Moreover, except for some well-defined and obvious adjacent pairs, the recipient has no a priori notion of what will be received, as well as it has no notion of the size of the emitter’s communication plan. The recipient’s task is to compose as many signs as possible, to interpret them, and then to extract a meaning from the entire message.

At the line 57, the agent checks if the received signs disagree with expectancies on a basal message comprehension. For now, let us consider that the logical test if $(\neg x_\xi=d \vee p''_k=d \vee e''_k=d \vee g''_k=d \vee h''_k=d \vee b''_k=d \vee a'_k=d \vee \pi_k=d)$ only detects a conformity (or not) with a basal level of interpretation. In case $\mathcal{R}$ disagrees with some interpreted sign, it may execute line 58 and triggers a negative multimodal feedback. It means that agent can perform some negative facial expression, gesture, sounds, or compound signs, which the agent may take from a learned LUT, possibly from a set of negative cultural signs. $\mathcal{R}$ can only expect that these signs would be interpreted by $\xi$ as a disagreement. Yet, $\mathcal{R}$ may also decide not to emit feedback at all, even disagreeing with what was received.

In our proposal, while receiving sequences of multimodal signs, the agent compares them to the contents of a database, which contains stereotyped sequences of signs for commonplace conversation. In such lower level, the agent generates feedbacks demonstrating whether the
interpreted signs match with sentences, gestures and facial expressions commonly used. If the sequence of received signs agrees with what \( \mathcal{R} \) expects from a well-performed message, the algorithm provides positive multimodal feedback. This is described in line 60 by the logical test if \( (v^i_k=a \land p^i_k=a \land e^i_k=a \land g^i_k=a \land h^i_k=a \land b^i_k=a \land c^i_k=a \land \pi_i=a) \) and the triggering of positive feedback is described in line 61. As said, \( \mathcal{R} \) may also decide not to emit feedbacks, as well as it may decide to generate neutral signs (line 64).

It is important to notice that, during the period in which \( \mathcal{R} \) interprets what the emitter communicates, it must prepare a communication plan to answer \( \xi \) as soon as it receives (or takes) the next turn. This is what the concept of adjacent pairs tells us, and it takes milliseconds in humans. In other words, while \( \mathcal{R} \) interprets a multimodal message, it prepares a plan with the contents of what will be communicated on the next turn.

It is necessary to clarify that not always \( \mathcal{R} \) is able to start elaborating a plan right after interpreting the \( k \)th Cu. In fact, \( \mathcal{R} \) must receive enough and relevant information in such a way that the conversation comes to a meaning. For instance, let’s consider that on the previous turn \( \mathcal{R} \), as an emitter, had asked: “how old are you?” At the current turn, \( \xi \) answers like this: “well, I was born in the previous millennium, in 1930! Yes, I am quite old man, I am 88 years old”. In this case, \( \mathcal{R} \) does not receive enough information until \( \xi \) says 1930 or 88 years. This is what was described as TRP (section 2.3.4). As said earlier, there may have several ways to implement TRP, because a designer may model such detection within the meaning integrator, another designer may implement it as a separate module, while another one may implement it within the adjacent pair module, once TRP is related to preserving the meaning of a message within an adjacent pair.

In summary, after receiving enough information, \( \mathcal{R} \) starts creating a plan to continue the logical structures of messages on the adjacent pair. \( \mathcal{R} \) does that even during the time it is still interpreting the message that \( \xi \) is emitting. Anticipation is a key point to make direct physical interactions more efficient. Our brains anticipate the creation of the next communication plan to improve the dynamics of face-to-face communication. For face-to-face interactions to look more natural and efficient, an AI agent should have a module that anticipates the creation of communication plans on the fly, mimicking the human brains. In line 66, the algorithm triggers a module that describes the plan creation for the next turn.

Finally, line 67 shows a structure if \( (v^i_k=c \lor p^i_k=c \lor e^i_k=c \lor g^i_k=c \lor h^i_k=c \lor b^i_k=c \lor c^i_k=c \lor \pi_i=c) \) that checks whether the agent \( \xi \) has sent any sign indicating a cue for turn change. If \( \mathcal{R} \) interprets any sound or visual sign as an indication that it should resume the conversation, then \( \mathcal{R} \) changes its own status to emitter (line 68) and takes on the execution of the next turn. As we have seen in the emitter’s role, when a cue for changing turn does not occur, agent \( \mathcal{R} \) may wait a while and tries to interrupt the emitter, what is called by Linguists as to take the turn (or to take the floor). The action of taking turns may also result in embarrassments and may be interpreted as a sign of \( \mathcal{R} \)’s rudeness or as an unpolished attitude. Although it is not described in this algorithm, processes for dealing with such situations should be predicted in man-machine interactions. Policies and subtleties of turn change can be learned to the extent that agents interact face-to-face.

### 6. Discussion

There are four fundamental discussions in this section: why we have used the notion of mental states on this algorithm, how to represent and process data represented by tuples on this proposal, how the ideas and thoughts are transmitted and interpreted within the framework, as well as how to deal with turn taking and disruptions on the normal flow of the algorithm.

#### 6.1 Why mental states.

A first discussion to be addressed refers to the way we used the Conversation Analysis perspective. We honestly will understand any criticism concerning how we associate intentions with this theory. CA uses ethnomethodology as the main methodological tool, a methodology that avoids attributing
mental states to the participants of the conversation. In fact, mental states are not attributed to the participants of the interaction in any way. When analyzing a video or audio of participants of a conversation the CA researcher never says that someone was mocking, or hesitant, embarrassed, shy or anything like that. In our proposal we are doing something that may receive severe criticism because we are associating CA with Theory of Mind. In ToM, agents do something to bring about a mental state in the interactor’s mind. In ethnomethodology, as observers, we cannot say what the participant’s intention was. Only the participant himself could reveal what was his real intention. Researcher’s intervention can be a source of problems because many humans do things in interactions without being aware of what they wanted to cause. Humans may have intentions that are hidden to themselves.

However, members of the AI community and computer programmers know that it is not possible to send a complex message without each step of such message (a sequence of signs) being pre-defined. Moreover, we intend to aggregate intonation and multimodal components to the message. Any message sender has to pre-organize such signs before to start transmitting them. Hence, it seems appropriate to make use of known concepts, such as plans and intentions, used in BDI agents, to explain how a machine could perform any natural communication with humans.

A plan is an anticipation of some steps that an agent must carry out pursuing a goal. Anticipatory systems are what AI most constructs and deals with. To some degree, acting intelligently means anticipating events and deciding which action(s), among the possible options, will be the most appropriate in each context. This brings the notion that intelligent agents choose certain responses because they have intentions and act aiming certain results.

In this sense, the agent’s goal when communicating is to cause states of comprehension of a message in the interpreter system of a recipient. Just to facilitate the writing of this article we have used the term mind for such interpreter system. If an agent intends to cause a certain state of understanding in a human mind, it is easier for us to use in the text mental states instead of other term. Mental state is the main matter of Theory of Mind, thus, if we want to speak natural language with machines, it will be natural that machines might be able to understand ironies, sarcasms, jokes, metaphors, facial expressions, gestures; besides the union of different multimodal signs for changing the overall meaning of what may seems a plain message. It means also that machines ought to know how to generate these language resources for human interpretation. Any set of signs planned to generate in humans certain mental states can be considered an intentional act. In a fabulous article, Nagel questioned "what is it like to be a bat?" (59). It would be the case to ask whether machines someday will have the answer for the question "what is it like to be a human?".

For our algorithm, we think that machines can deal with representations that give them abilities to provoke and understand certain human mental states. Therefore, we should consider they may have some kind of mind. Besides, they must have a planner capable of transforming the intentional act of provoking certain mental states into a human mind into a set of multimodal signs.

### 6.2 Improving the Methods

As mentioned earlier, our research method is based on qualitative analysis of videotaped semi-spontaneous interactions. We carry out micro-level analysis of conversational sequences having as foundation the perspective of CA. In spite of the quality and the number of cameras used on video capture, and in spite of the better the software tool used to perform the analysis, the identification of expressions and gestures is still subjective. When an interactor performs subtle manifestations, one analyzer may differ from another concerning certain manifestation of prosody, certain gesture, a relevant facial expression which modifies the meaning of a message, and so on.

In order to eliminate discrepancies in analyzes, a plausible proposal would be the inclusion of neuroscience in our method of capturing and analyzing facial interactions. There are today’s data acquisition systems that unequivocally show whether certain muscle activities really occur. As an example, we can mention motion capture systems for body postures, facial expressions and manual gestures. We can even make use of electromyograms to capture muscular activity on the faces.
Three types of valuable data may emerge from the introduction of neuroscience measurements in experiments: firstly, we can be sure of the activation of certain signs during communication; secondly, we obtain temporal measures of the occurrence of relevant events, and thirdly, we can have quantitative measures of the occurrence of events, which may favor the application of statistical tools on the data and thus allow to find correlations between events. A possible drawback by using neuroscientific method would be the loss of spontaneity in the interaction as we place the apparatus for neurological data capture. We suspect that, even if we do not use invasive equipment in the capture of neural signals, we may observe an interference of the devices in the spontaneity of the conversation, but it must be confirmed. Using automated analysis software should also improve the performance of the analysis. In addition, it would still be necessary to perform experiments with standardized method in various cultures to prove which components of the algorithm proposed herein are truly universal (if any), and which are culturally specific. Specifically, for human-machine interactions, we have also to investigate which kind of multimodal components may function and which ones may not work well. Several investigations are in course in this direction. Maybe certain multimodal components may work well for robots but not for mobile phones or for computers. Possibly, we will have to replace the task-oriented assessment method, often used in AI, by skill-oriented (60) or cognitive performance methods in order to better evaluate the performance of an artificial agent which uses multimodal interaction.

6.3 Resisting or Acquiescing to New Ideas
Another discussion concerns the use of the terms agreement, disagreement, and neutral in the algorithm. When we say that $\mathfrak{R}$ sends feedback indicating that it agrees with $\xi$, and vice versa, or when we say $\xi$ checks if the signs disagree with some expected feedback, we intend to say that $\mathfrak{R}$ and $\xi$ have informed that correctly interpreted the sequence of signs as a valid message. We do not mean that they have acquiesced to an idea issued by the other. Any effort for convincing the recipient is beyond the scope of our study. Attempting to convince the recipient of something leads an interaction to a level at which the emitter attempts to change the other agent’s beliefs, which is not the case explained in this algorithm. This is a subject that requires knowledge from Discourse Analysis, which is another area of Linguistic studies. We may say that a hierarchically superior layer is needed to deal with ideas, discourses and persuasion within the present algorithm. It seems to us that the agent must have mechanisms for creating a second order plan, not only for checking whether its basal communication plan is correctly interpreted but also for checking whether the other agent demonstrates that has captured the main idea, has convinced about key arguments, and so on. Nevertheless, after implementing such level of information processing, designers must be aware that agents send feedbacks about “acquiescence” or “resistance” to arguments and ideas within the same algorithm and by the same sign modalities we have described. In other words, because $\mathfrak{R}$ and $\xi$ may not have the same beliefs, their ideas may not match. Resisting or acquiescing to new ideas, concepts, and arguments is in other layer of information treatment, but the feedback channels are the same and when an agent agree/disagrees with the interactor (indicating it has not understood something) it sends the same multimodal feedback signs as when it resists or accept new information. An agent may use the same vocalization or facial expression for the whole discourse or for a single punctual misread sign. In this sense, it is up to some dedicated engine in the agents to interpret whether the feedback indicates accordance (or not) with a single chunk of information (a Cu), an utterance, an idea, or with the general discourse.

6.4 Interruptions and Turn Taking
Normally, after executing all Cus, the agent passes the turn to the other interactor. However, a communication plan in execution may be interrupted by other interactor(s) that may overlap or even forcibly take the turn before an emitter complete the execution of its turn. We may consider that
agents should learn how to deal with specific characteristics of turn taking (taking the floor - section 2.3.4) for each agents (or groups) in the same way they may learn traces of personalities of each interactor. It is known that, in certain cultures, people talk simultaneously, overlapping the conversation. In certain cultures, or in hot debates, it is usual the recipient to interrupt and to take the floor from a speaker, while in other cultures or in certain contexts it may mean rudeness. In other words, one agent should learn and memorize specific details about overlapping and taking the floor. Each agent must have some a database with descriptions of each interactor’s usual habits concerning this aspect; such as: “this agent always waits the end of my turn,” or “this agent frequently overlaps me when it is nervous,” or “this agent is rude because it usually takes my turn and does not let me finish my utterance,” and so on. Discussions concerning exceptions to the basic protocol presented in this text are beyond the scope of this article and can be an interesting subject for investigation.

7. Conclusion
Multimodal face-to-face interactions greatly enrich the exchange of information between two agents, mainly because both agents become active, that is, both agents are emitting and interpreting signs simultaneously. From the point of view of who is emitting an utterance, the advantage is to have more resources, more channels to add signs that add values to the information. From the standpoint of who is receiving an utterance, the advantage is to be able to judge relevant components that are not in the utterance itself. On the other hand, from the computational point of view, multimodal interactions simply mean exponential increase in computational demand, in addition to an increase in the complexity of the algorithms at levels still far from being reached. Linguists began to scratch the surface of multimodal communication. As pointed out by Vinciarelli and colleagues, there are many challenges and much to be done in modeling, analysis and synthesis of human behavior, in human-human and human-machine interactions (61). Recent advances in the area begin to show how the various modalities interact with each other (see (33), (36), (40), (43), (10)). These findings reveal not only how multimodal interactions modify meanings but also what kind of computation a mind must perform to maintain the flow of face-to-face interaction.

With data obtained from our research we could testify that several multimodal signs interact with each other when two agents are in direct face-to-face conversation. Based on these data we set up a framework that fits to linguistic theories we have based on, such as Conversation Analysis and Gesture Studies. Based on such findings we show in this paper an algorithm that elucidate what kind of information processing an artificial intelligence system must realize to mimic the human face-to-face interaction.

We conclude that it is possible to construct a double-feedback system that accounts for processing all types of signs (verbal, prosodic, visual signs, etc.) and extracting meaning from them. This loop can maintain the conversation flow and organization of turn-taking, adjacent pairs, preferences, repairs, self-repairs, as well as to implement transition relevance place (TRP) for increasing performance on direct interactions. Moreover, we conclude it is possible to create such loop aggregating independent and cooperative modules, each one dealing with one modality of human communication. Having separate modules that recognize syntax, semantic, prosody, facial expressions, bodily expressions, etc. independently allow the system to grow by incorporating modalities as desired.

A challenge is to integrate and keep track of the temporal evolution for each set of signs. Moreover, all these multimodal processing might be done in "real time", once humans react to face-to-face communication in a timescale of hundreds of milliseconds. Another big challenge is to construct higher hierarchical levels for interpreting ideas, concepts and thoughts.

In our view, a major contribution of this article is to present an overview on how the operational flow of multimodal face-to-face interactions happens. The algorithm describes how multimodal signs interfere on the interactive conversation, allowing the interactors to change turns, to perform repairs, or even changing “on the fly” their intentions of communication. We claim that a double
(positive or negative) feedback loop in face-to-face multimodal communication keeps interactions within culturally established patterns.

Lots must be done. Firstly, we deal with data coming from video analysis alone. In this direction, neuroscientific measures should validate and reinforce our findings. Secondly, we have analyzed only conversation of native Brazilian-Portuguese speakers. Despite this, based on cues from specialized literature, we are convinced that the algorithm should work in many (if not all) cultures. Moreover, the algorithm seems to work for several man-machine interactive situations, from chatterbots and virtual characters to robots, what makes it especially interesting for AI application. Further investigation may reveal what is common and what may be cultural artifacts in face-to-face interaction; but we are convinced that the double feedback loop remains valid. What may change are nuances of how specific cultures treat certain signs, especially when these signs overlap during the conversational flow.

**Acknowledgments**

We would like to thank the participants of all video-recorded face-to-face interaction sessions, which we analyzed in our research. We also would like to thank the colleagues at PPGINF / UFABC (Pós-Graduação em Engenharia de Informação) for invaluable discussions concerning the subjects of this research. Finally, we thank Leland McCleary and Evani Viotti for their constructive talks about the topic of face-to-face interactions and their permission to use the controlled vocabularies proposed by LLIC (Language, Interaction and Cognition Laboratory – USP– Brazil).

**Appendix A. Transcription conventions (symbols used on Table 1)**

- left square brackets indicate a point of overlap onset.
- right square brackets indicate a point at which two overlapping utterances both ends, or one utterance ends while the other continues, or simultaneous moments in overlaps which continue.
- equal signs indicate continuous utterance with no break or pause and/or latch.
- a brief pause, usually less than 0.2 seconds
- colons indicate prolongation of the immediately prior sound. The longer the colon row, the longer the prolongation.
- a dash indicates a cut-off of the preceding word or sound or an interruption in utterance.
- a degree symbol indicates whisper or increased volume speech
- question mark or rising pitch

**Reference**

1. Levinson SC, Holler J. The origin of human multi-modal communication. Philos Trans R Soc Lond B Biol Sci [Internet]. 2014 Sep 19 [cited 2018 Nov 18];369(1651):20130302. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25092670
2. Pollick AS, de Waal FBM. Ape gestures and language evolution. Proc Natl Acad Sci U S A [Internet]. 2007 May 8 [cited 2018 Nov 18];104(19):8184–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17470779
3. Krupenye C, Kano F, Hirata S, Call J, Tomasello M. Great apes anticipate that other individuals will act according to false beliefs. Science [Internet]. 2016 Oct 7 [cited 2018 Nov 18];354(6308):110–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27846501
4. Rajan K, Saffiotti A. Towards a science of integrated AI and Robotics. Artif Intell [Internet]. 2017 Jun 1 [cited 2017 Nov 4];247:1–9. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0004370217300310

5. Lemaignan S, Warnier M, Sisbot EA, Clodic A, Alami R. Artificial cognition for social human–robot interaction: An implementation. Artif Intell [Internet]. 2017 Jun 1 [cited 2017 Nov 4];247:45–69. Available from: https://www.sciencedirect.com/science/article/pii/S0004370216300790

6. Thomessen T, Niitsuma M, Suzuki K, Hatano T, Hashimoto H. Towards Virtual Presence Based on Multimodal Man-Machine Communication: A Remote Operation Support System for Industrial Robots. IFAC-PapersOnLine [Internet]. 2015 Jan 1 [cited 2018 Nov 30];48(19):172–7. Available from: https://www.sciencedirect.com/science/article/pii/S2405896315026531!

7. Oviatt S, Sharon. Ten myths of multimodal interaction. Commun ACM [Internet]. 1999 Nov 1 [cited 2018 Nov 18];42(11):74–81. Available from: http://portal.acm.org/citation.cfm?doid=319382.319398

8. Sebe N, Cohen I, Gevers T, Huang TS. Multimodal approaches for emotion recognition: a survey. In: Santini S, Schettini R, Gevers T, editors. Proc SPIE 5670, Internet Imaging VI, 56 [Internet]. International Society for Optics and Photonics; 2005 [cited 2017 Jun 10]. p. 56–67. Available from: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=857894

9. Turk M. Multimodal interaction: A review. Pattern Recognit Lett [Internet]. 2014 Jan 15 [cited 2017 Jun 10];36:189–95. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0167865513002584

10. de Lima CV. A multimodalidade na conversa face a face em episódios de desacordo [Internet]. [São Paulo]: Biblioteca Digital de Teses e Dissertações da Universidade de São Paulo; 2018. Available from: http://www.teses.usp.br/teses/disponiveis/8/8147/tde-04062018-182512/

11. Vilela C, Ranhel J. A framework for the multimodal joint work of turn construction in face-to-face interaction. Cogn Syst Res [Internet]. 2017 Mar 1 [cited 2017 Nov 23];41:99–115. Available from: http://www.sciencedirect.com/science/article/pii/S1389041716300675

12. Sacks H. Lectures on conversation (2 vols). Malden: Blackwell Publishing; 1992.

13. Kendon A. Gesture: visible action as utterance. Cambridge: Cambridge University Press; 2004.

14. Ekman P. Telling lies: Clues to deceit in the marketplace, politics, and marriage. New York and London: W.W. Norton & Company; 1992.

15. Rao AS, Georgeff MP. BDI Agents: From Theory to Practice [Internet]. 1995 [cited 2018 Nov 9]. Available from: www.aaai.org

16. Rao AS, Georgeff MP. Modeling rational agents within a BDI-architecture. In: Huhns MN, Singh MP, editors. Readings in agents. San Francisco, CA: Morgan Kaufmann Pub. Inc; 1998. p. 317–28.

17. Wooldridge MJ. An Introduction to MultiAgent Systems [Internet]. Chichester - England: John Wiley & Sons; 2002 [cited 2018 Nov 18]. 461 p. Available from: https://books.google.com.br/books?hl=pt-BR&lr=&id=X3ZQ7yeDn2IC&oi=fnd&pg=PR13&dq=multiagent+systems+wooldridge&ots=WGlgs50sS5&sig=Zg44wBXARPm2itndoiCF346GuWA#v=onepage&q=multiagent systems wooldridge&f=false

18. Goldman AI. Theory of Mind. In: Margolis E, Samuels R, Stich S, editors. The Oxford Handbook of Philosophy of Cognitive Science [Internet]. Oxford Handbooks; 2012 [cited 2018 Nov 18]. p. 576. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.405.75&rep=rep1&type=pdf

19. Short TL. Peirce's Theory of Signs. New York, NY: Cambridge University Press; 2009.

20. Nöth W. Handbook of Semiotics: Advances in Semiotics. Bloomington,IN: Indiana University Press; 1990.

21. Loewenstein G, Lerner JS. The role of affect in decision making. In: Davidson R, Scherer K, Goldsmith H, editors. Handbook of Affective Sciences. New York, NY: Oxford University Press; 2003. p. 619–42.
22. Heise DR. Expressive order: confirming sentiments in social actions. Springer; 2007. 163 p.
23. Robinson DT, Smith-Lovin L, Wisecup AK. Affect Control Theory. In: Handbook of the Sociology of Emotions [Internet]. Boston, MA: Springer US; 2006 [cited 2018 Nov 17]. p. 179–202. Available from: http://link.springer.com/10.1007/978-0-387-30715-2_9
24. Schröder T, Hoey J, Rogers KB. Modeling Dynamic Identities and Uncertainty in Social Interactions. Am Sociol Rev [Internet]. 2016 Aug 8 [cited 2018 Nov 18];81(4):828–55. Available from: http://journals.sagepub.com/doi/10.1177/0003122416650963
25. Hoey J, Schröder T, Alhothali A. Affect control processes: Intelligent affective interaction using a partially observable Markov decision process. Artif Intell [Internet]. 2016 Jan 1 [cited 2018 Nov 18];230:134–72. Available from: https://www.sciencedirect.com/science/article/abs/pii/S000437021500140X
26. Hausser RR. Computational linguistics and talking robots: processing content in database semantics [Internet]. Springer; 2011 [cited 2018 Nov 18]. 286 p. Available from: https://books.google.com.br/books?hl=pt-BR&lr=&id=nGMIHogQIQp8C&oi=fn&pg=PR3&dq=computational+linguistics+and+talking+robots&ots=3mLl9Nnnri&sig=ZIQtQqMzYwT-1Woz30qtIvCkrTw#v=onepage&q=computational linguistics and talking robots&f=false
27. Hausser R. Generalized Reference. In Springer, Cham; 2017 [cited 2018 Nov 18]. p. 427–46. Available from: http://link.springer.com/10.1007/978-3-319-58071-5_32
28. Hausser R. A computational treatment of generalized reference. Complex Adapt Syst Model [Internet]. 2017 Dec 21 [cited 2018 Nov 18];5(1):2. Available from: http://casmodeling.springeropen.com/articles/10.1186/s40294-016-0042-7
29. Povinelli DJ, Preuss TM. Theory of mind: evolutionary history of a cognitive specialization. Trends Neurosci. 1995;18(9):418–24.
30. Heyes CM. Theory of mind in nonhuman primates. Behav Brain Sci. 1998;21(01):101–14.
31. Sacks H, Schegloff E, Jefferson G. A simplest systematics for the organization of turn-taking for conversation. Language (Baltim). 1974;50(4):696–735.
32. Garfinkel H. Studies in Ethnomethodology. Englewood Cliffs: Prentice-Hall; 1967.
33. Schegloff E. Sequence organization in interaction: a primer in Conversation Analysis I. Cambridge: Cambridge University Press; 2007.
34. Sacks H. On the preferences for agreement and contiguity in sequences in conversation. In: Button G, Lee JRE, editors. Talk and social organization. Clevedon: Multilingual Matters; 1987. p. 54–69.
35. Enfield NJ. Relationship thinking: agency, enchrony, and human sociality. Oxford: Oxford University Press; 2013.
36. Mondada L. Multimodal resources for turn-taking: pointing and the emergence of possible next speakers. Discourse Stud. 2007;9(2):194–225.
37. Levinson S. Action formation and ascription. In: Sidnell J, Stivers T, editors. The Handbook of Conversation Analysis. Blackwell Publishing Ltd; 2013. p. 103–30.
38. Levelt WJM. Speaking: from intention to articulation. Cambridge, Mass: MIT Press; 1989.
39. de Ruiter JP, Mitterer H, Enfield NJ. Projecting the end of a speaker’s turn: a cognitive cornerstone of conversation. Language (Baltim). 2006;82(3):515–35.
40. Stivers T, Enfield NJ, Brown P, Englert C, Hayashi M, Heinemann T, et al. Universals and cultural variation in turn-taking in conversation. PNAS. 2009;106(26):10587–92.
41. Schegloff E. The relevance of repair for syntax-for-conversation. In: Givón T, editor. Discourse and syntax. New York: Academic Press; 1979. p. 261–86.
42. Schegloff E, Jefferson G, Sacks H. The preference for self-correction in the organization of repair in conversation. Language (Baltim). 1977;53(2):361–82.
43. Enfield NJ, Dingemanse M, Baranova J, Blythe J, Brown P, Dirksmeyer T, et al. Huh? What? – A first survey in 21 languages. In: Hayashi M, Raymond G, Sidnell J, editors. Conversational repair and human understanding. Cambridge: Cambridge University Press; 2017. p. 343–80.
44. Ekman P. A language das emoções. Sao Paulo: Lua de Papel.; 2011.
Ekman P, Friesen W. Felt, false, and miserable smiles. J Nonverbal Behav. 1982;6(4):238–52.

Niedenthal P, Krauth-Gruber S, Ric F. Psychology of emotion: interpersonal, experiential, and cognitive approach. New York: Psychology Press; 2006.

Broth M, Mondada L. Walking away: The embodied achievement of activity closings in mobile interaction. J Pragmat [Internet]. 2013 Feb 1 [cited 2018 Nov 1];47(1):41–58. Available from: https://www.sciencedirect.com/science/article/pii/S0378216612003050

Goodwin C. The co-operative, transformative organization of human action and knowledge. J Pragmat. 2013;46(1):8–23.

Kendon A. Semiotic diversity in utterance production and the concept of “language.” Philos Trans R Soc B Biol Sci. 2014;369(1651).

Mondada L. The local constitution of multimodal resources for social interaction. J Pragmat. 2014;65:137–56.

Peräkylä A, Ruusuvuori J. Facial expression in assessment. In: Knoublauch H, Schnettler B, Raab J, Soeffner H-G, editors. Video analysis: methodology and methods - qualitative audiovisual data analysis in Sociology. Frankfurt am: Peter Lang International Academic Publishers; 2006. p. 127–42.

Peräkylä A, Ruusuvuori J. Facial expression and interactional regulation of emotion. In: Peräkylä A, Sorjonen M-L, editors. Emotion in interaction. Oxford: Oxford University Press; 2012. p. 64–91.

Streeck J. Gesturecraft: the manu-facture of meaning. Amsterdam/Philadelphia: John Benjamins.; 2009.

Streeck J. Interaction and the living body. J Pragmat. 2013;46(1):69–90.

Clark HH, Brennan SE. Grounding in communication. In: Resnick LB, Levine JM, Teasley SD, editors. Perspectives on socially shared cognition. Washington, DC: American Psychological Association; 1991. p. 127–49.

Chafe W. Intonational units. In: Chafe W, editor. Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing. Chicago and London: The University of Chicago Press; 1994. p. 53–70.

DuBois J, Schuetze-Coburn S, Paolino D, Cumming S. Discourse transcription. Santa Barbara: University of California; 1990.

Jefferson G. Glossary of transcript symbols with an introduction. In: Lerner G, editor. Conversation Analysis: studies from the first generation. Amsterdam/Philadelphia: John Benjamins Publishing Company; 2004. p. 13–31.

Nagel T. What Is It Like to Be a Bat? Philos Rev [Internet]. 1974 Oct [cited 2018 Nov 18];83(4):435. Available from: https://www.jstor.org/stable/2183914?origin=crossref

Hernández-Orallo J. Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. Artif Intell Rev [Internet]. 2017 Oct 19 [cited 2018 Nov 18];48(3):397–447. Available from: http://link.springer.com/10.1007/s10462-016-9505-7

Vinciarelli A, Esposito A, André E, Bonin F, Chetouani M, Cohn JF, et al. Open Challenges in Modelling, Analysis and Synthesis of Human Behaviour in Human–Human and Human–Machine Interactions. Cognit Comput [Internet]. 2015 Aug 12 [cited 2017 Nov 5];7(4):397–413. Available from: http://link.springer.com/10.1007/s12559-015-9326-z