Sequential annotations for naturally-occurring HRI: first insights

Lucien Tisserand
UMR 5191 ICAR, CNRS, Univ Lyon, ENS Lyon
Lyon, France
lucien.tisserand@ens-lyon.fr

Frédéric Armetta
Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622
Villeurbanne, France
frederic.armetta@liris.cnrs.fr

Heike Baldauf-Quilliatre
UMR 5191 ICAR, CNRS, Univ Lyon, ENS Lyon
Lyon, France
heike.baldaufquilliatre@ens-lyon.fr

Antoine Bouquin
Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622
Villeurbanne, France
antoine.bouquin@liris.cnrs.fr

Salima Hassas
Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622
Villeurbanne, France
salima.hassas@liris.cnrs.fr

Mathieu Lefort
Univ Lyon, UCBL, CNRS, INSA Lyon, LIRIS, UMR5205, F-69622
Villeurbanne, France
mathieu.lefort@liris.cnrs.fr

ABSTRACT

We explain the methodology we developed for improving the interactions accomplished by an embedded conversational agent, drawing from Conversation Analytic sequential and multimodal analysis. The use case is a Pepper robot that is expected to inform and orient users in a library. In order to propose and learn better interactive schema, we are creating a corpus of naturally-occurring interactions that will be made available to the community. To do so, we propose an annotation practice based on some theoretical underpinnings about the use of language and multimodal resources in human-robot interaction.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; • **Human-centered computing** → *Text input*; *HCI theory, concepts and models*; **Field studies**.

KEYWORDS

social robotics, methodology, dataset, tagging, sequence organization, multimodality, in the wild

ACM Reference Format:
Lucien Tisserand, Frédéric Armetta, Heike Baldauf-Quilliatre, Antoine Bouquin, Salima Hassas, and Mathieu Lefort. 2023. Sequential annotations for naturally-occurring HRI: first insights. In *Proceedings of Workshop on Human-Robot Conversational Interaction (HRCI Workshop '23).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

HRCI Workshop '23, March 13th, 2023, Stockholm, SE

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION: PROJECT GOALS

This methodological paper draws from an ongoing project called Peppermint[1]. As part of this project, *Conversation Analysis* researchers and *Artificial Intelligence* researchers team up in a collaborative effort to improve interactions in the wild with an autonomous Pepper[2] robot as regard with turn-taking and action recognition with the perspective of sequential organization. The robot's use case is offering information and orientation services in a university library in France.

We follow an inductive, step-by-step approach that rely on the production and analysis of naturally-occurring data. In short, we first created an *ad hoc* autonomous conversational system as a state machine. This first software allowed us to video-record naturally-occurring data (see 3.3) of *human-robot interactions* (*HRI*) for empirical, inductive findings (of which are *HRI* specifics). We are now structuring this corpus and annotating it with regard to a core principle of human interaction: *sequence and sequential organization.* This dataset will be used to improve conversational *HRI* by using machine learning / NLU methods.

In this paper, we first explain what it means to consider sequential organization as a temporal, continuous achievement of mutual understanding, and its relevance for having a conversational system to respond appropriately and timely (2). We then explain how a heterogeneous dataset of naturally-occurring HRI can be systematically managed through a labelling scheme (3). We finally sketch an annotation syntax addressing sequential organization (4) before discussing its potential (5).

2 THEORETICAL UNDERPINNINGS

Within this section, we explain how the *Conversation Analytic* (*CA*) approach to human interaction provides new insights on the analysis and annotation of a dataset that account for the sequential organization of multimodal HRI. Especially, we focus on the dynamics of normative expectations (vs. predictions) and interpretative feedbacks that allow a completely unique and unpredictable interaction to be controlled on a turn-by-turn basis.

From a *CA* perspective, the analysis of talk and gestures in interaction is above all the analysis of how talk (e.g. lexical choice,

[1]Full title "Interacting with Pepper: Mutual Learning of Turn-taking Practices in HRI" (2021-2024). Project website: https://peppermint.projet.liris.cnrs.fr/
[2]Manufactured by Aldebaran. Please visit https://www.aldebaran.com/fr/pepper

intonation...) and other resources (body position, gaze, gestures...) are in fact designed to be used in interaction [28]. This vision is opposed to intrinsically meaningful conducts that would simply be adapted to an interaction setting. Growing on *ethnomethodological* roots, *CA* shows how the mutual understanding is "an operation rather than a common intersection of overlapping sets" as Garfinkel puts it [13, p. 30]. As we will see, such process is achieved by the mean of sequential organization through *turn-taking* and it implies that participants formulate turns for accomplishing contextually relevant actions.

2.1 Mutual understanding in human-human interaction: bottom-up and top-down

CA has been partly founded on the investigation of noticeable regularities with regard to the accomplishment of turn-taking practices such as the transitions without gaps and overlaps despite the fact that turns have various durations [32]. The management of these regularities led to the idea that turns are composed of units (*Turn-Constructional Units*) processed by a commonly shared *turn-taking system* (*TTS*) which rules have been described. This system has sometimes only been seen as a finely tuned *bottom-up* mechanics [36], and some streams of research focused on such units [11, 12, 35], how these were implemented by an analyzable signal and some explored their computational processing [37].

But while it may appear relevant to study how turn-ending could be identified in order to handle turn-taking (such as [37]), we consider the other route, a more *top-down* approach: interactants do not take turns for the sake of taking turns, they do so in order to create a delimited and purposeful context for future interpretations of actions and intents that will be implemented by talk and bodily conducts. By doing so, they accomplish collaborative activities while continuously ensuring that what had to be interpreted was actually interpreted as such. Speakers make use of a *sequential* approach to interpretation: a next speaker's conduct is always interpreted within a slot temporally projected by a prior action even if the next action is ultimately interpreted as a completely unexpected next move. That functioning leads to the fact that next speakers do display such departures from projected next turns. For example, in the imaginary case below, it is indicated with the turn-beginning `"well,hum"` followed by an extended account:

```
A: "Hello, can I help you ?"
B: "Well, hum, I'm just waiting for my friend."
```

We might think of a commonly shared inventory of such contextual practices with the notion of *adjacency pairs* that draws on the idea that sequences of actions are culturally typified as normative pairs (greeting-greeting, offer-acceptance/reject...) [34]. Thanks to the turn-taking organization, different speakers participate alternatively to the first pair part or second pair part. They do not follow a set of rules or instructions that will determine their conducts, they refer to this norm in order to ease the action ascription of turns [4, 5, 24]. Although this "repertoire" vision is limited when it comes to grasp the complexity of human-human interaction [8], it appears adapted to the human "simplistic" approach to service encounter HRI (see 5).

If some verbal and multimodal resources participate to the *bottom-up* recognition of such actions (e.g. a Wh-questions projecting types of responses at turn-beginning [7]), sequence organization and adjacency pairs are crucial *top-down* resources for the recognition of actions in a delimited context [24] and thus the recognition of turn completions [25]. From that perspective, that also means that the "right" interpretation of a turn is the understanding of what can be produced next for all practical purpose (*versus* the semantic management and selection of all interpretable actions and meanings of a verbal turn).

2.2 Temporal, turn-by-turn increments

As explained above, humans make use of sequentiality to incrementally secure their interpretation of what they are expected to do next, and arguments point toward universals with regard to such infrastructure [21]. Just to give a glimpse of all intricate sequences that implies, participants may recognize and accomplish the answers that are expected [38], but they may also initiate *repairs* [18] projecting *reformulations* by the previous speaker, they may reformulate themselves what they understood [6] or produce feedback during the turn [16]. Moreover, these methods may be used at different places as regard with the *adjacency pair* organization: before (e.g. "what I wonder is") , in-between (e.g. "what do you wanna know?") or after ("okay great"). Some previous turns or whole sequences may be reformulated, expanded, but also normative expectations may just be abandoned.

Thus, the approach to the modelization of the temporal trajectory of an interaction might imply the suspension of the immediacy of second pair parts or following sequences. This turn-by-turn temporally incremented display of successive interpretations is at the heart of the mutual understanding process in interaction [26]. As Levinson [23, p. 47] recalls us, the representation of successive turns of a human-human interaction is then less a linear representation like `[A1->B1->A2->B2]` than some kinds of stacking structures like `[A1->B2->A2->B1]`, where letters are the interactants, and numbers distinguishing sequence types. These are the structures we aim at investigating and annotating.

2.3 What about artificial conversational agents ?

If we apply this perspective to the design of an autonomous conversational agent, a quite reluctant implication is that no word-based treatment of the human input is sufficient to ascribe the human turn to actions (*bottom-up dead* end). Moreover, the sequence organization being not a set of instructions but a set of conventions, every next move is virtually possible, and no pre-drawn scheme of action can be hard-coded (*top-down* dead end).

A more attractive perspective is to consider the fact that humans make use of turns and norms as a way to produce more flexible and negotiable interpretations and that speakers leave cues that make these practices recognizable like the "well hum" above. The criterion for a successful next turn is the formulation of a possible next that projects further sequences, which means that there is more than only one "good answer" produced by the machine. What can then be investigated is the set of procedures that humans rely on in order to make sense of every next turn.

When it comes to the design or analysis of conversational agents, several researchers explore the benefits of the incremental dimension of interaction for securing a face-to-face encounter with an

autonomous and responsive system, like Fischer and Sikveland [10] in the case of what Stivers and Robinson have identified as *progressivity* [38], or Julian Hough for self-repair practices [19]. Housley and colleagues [20] advocated for collaborative attempts to apply CA to AI in the case of big interactional data with a sequence-of-action oriented approach (*vs.* linguistic features or emotional cues). The attempts we present in this methodological and reflexive paper can be read as one way to go in that direction.

In fact, considering the sequential infrastructure (as a set of possible next, preferred next, insertions, expansions, projections of series of sequences...) for computing, as CA already established it, is a proposition that goes back more than thirty years ago, when Gilbert, Wooffitt and Fraser[14] addressed the fact that the sequence analysis drawing on adjacency pairs could be subject to formalisation, although these would not explain all the contextual cues that participate to mutual understanding agreement. This initiative was duly demotivated [4, 5] by arguments that we mentioned and that we also agree with: the contingent and contextual character of interpretations, the non-scriptable character of interactions, the conventional (vs. instructional) character of sequence organization.

However, back in the days, what was discussed was the possibility to hard-code such grammar rules for the management of turns in interaction and for sentence/action recognition, as a deductive approach. Given the progress that have been made into the automatic discovery of statistical/probabilistic rules from annotated datasets (both in Natural Language Processing and Understanding), even with a small number of tokens (in the case of few shots learning), we advocate that it worth trying to rely on complex sequential annotations and adequate algorithms in order to provide a conversational agent with a statistically-oriented sequence management for all practical purpose. Moreover as conversational user interfaces are now ubiquitous in various societies, people display an alignment with such functioning (see [31] and 5).

3 CORPUS CONSTRUCTION

Within this section, we present how we acquired the data in order to make sure we would obtain naturally-occurring data that account for the complex sequential and multi-party dimensions of interactions. By essence, such a corpus is heterogeneous, that is why we explain how we manage this through a labelling scheme (3.5).

3.1 Use case

The dialog proposed by Pepper was based on a state machine (see appendix B for the details) where the transitions rely on the detection of some specific words. We used the manufacturer's built-in APIs[3] for word detection and prompt-to-answer rules. In order to anticipate user questions, we asked what were the most simple and recurrent requests that the library users were asking to the reception desk agents (location of toilets, how to connect to the wi-fi...). We asked them what they would like to see accomplished by a robot, so that it could be seen as a alternate service provider for these minimal and repetitive tasks. We also added some "Pepper-centered" answers to questions about the robot's age, name, purpose, feeling, capabilities...

3.2 Data acquisition

We placed the robot in the same area as where the reception desk was situated, at the entrance of the university library. Two large angle cameras were strategically placed in order to grasp the whole scene and especially to understand how users approached the robot before the opening of the interaction. We recorded the audio and video streams from Pepper's tablet. As the robot was not programmed to move, it was easier to define a record area. As regard with personal data protection, posters were placed near the various entrances of the library. After each interaction, a team member obtained signed consent, otherwise the data was deleted. Eleven recording sessions took place: seven days in March 2022 (17 hours of recording in total) and four days in September 2022 (12 hours of recording in total) as we expected more newcomers at this period of the year in a university library.

3.3 Naturally-occurring interactions

If we consider the sequential organisation of talk as a mean to secure the appropriate interpretation *for all practical purpose* between two humans, we assume that this minimal understanding procedure between the user and the robot depends on interactional emergent goals brought by the human in front of a seemingly speaking-and-hearing humanoid robot. Hence, a laboratory setting biases these procedures. For example, users migh pursue the goal of accomplishing a given script, or officially leave the face-to-face configuration having produced reasonably enough turns, overcoming the robot's failures, in order to not disappoint the experimenter....

When we talk about "naturally-occurring interactions" ([27] for in depth definition and reflection), we point at the fact that interactions were not orchestrated by the researchers. No instructions were given, users were free to interact with the robot and leave whenever they wanted. One can argue that we intervened in the routines of the library users, which could contradict the "natural" and "unorchestrated" character of the interactions in the ethnographic sense. However, we disrupted the users habits by the mean of recognizable and acceptable practices: we used a commercial-looking robot that institutions and enterprises use in order to provide basic services while accounting for some technological modernity. During the recordings, a vast majority of library users thought that it was the library initiative to showcase this robot (despite the posters) until we explained the purpose of its presence.

3.4 Data format

The large-angle views and the robot's view recordings were manually synchronized in a video editor. We exported long-format edited multi-scope videos corresponding to each recording sessions (2-3 hours long) that become a temporal reference for time-aligned annotations (4). The ELAN[4] software is used for all the annotation tasks. The identification and time alignment of the original robot's

[3]Mainly the QiChatbot API using QiChat script language. Please visit https://qisdk.softbankrobotics.com/sdk/doc/pepper-sdk/ch4_api/conversation/qichat/qichat_index.html

[4]ELAN (Version 6.4) [Computer software]. (2022). Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. Retrieved from https://archive.mpi.nl/tla/elan

view recordings have been indicated in a dedicated tier in the annotation software: this way, a script can extract clips corresponding to annotated segments created on other tiers and create different training datasets.

Hence, the data consists of original Pepper's view video files, long-format reference video files, ELAN annotation files, and scripts that extract text (transcriptions and annotations) altogether with original video clips from ELAN annotations.

3.5 A sequence-oriented labelling scheme for heterogeneous data

Because we decided to go with naturally-occurring interactions, the corpus is heterogeneous by design. No instructions were given to the users, they did not follow a script with recognizable stages, and the state machine allowed for a large set of combinations (see state machine representation in Appendix B). Before doing time-aligned transcriptions and annotations of human-robot interactions (4), we needed to manually define time-aligned shortclips from the long-format video references and characterize them in order to deal with this heterogeneity.

Thus, we created a labelling scheme and syntax that refer to approximately identified actions that can be part of a sequence. For example, the "greeting1" and "greeting2" tags account for the two parts of a greeting exchange. These tags are entered following their order of appearance in the clips. These are prefixed by letters indicating if the transmitter is the human (h) or the robot (p), and suffixed by the same letters indicating the recipient. This way, we may characterize a whole clip with a string like:

`hgreeting1p, hquestionp, silence, pgreeting2h`

where we can account for the fact that a human question and a noticeable silence preceded the answered greeting produced by the robot. This is always the human interpretation that supersedes the interpretation of the robot's action. For example, a same turn produced by the robot "I can provide you information about the library" might be understood as an account for not having responded to a previous request but also as a proposal. Other sequential and turn-taking features are placed on the same string, like repairs, repeats, or overlaps. Finally, remarkable and specific resources and phenomena that have been identified in the first analyses of data (internal data sessions) have been added (the fact that the robot might gaze away, when a new eye-contact is established, laughter...).

This way we could identify some first regularities in order to deploy research strategies (with appropriate search strings) and select the more relevant data to segment and transcribe in ELAN, as it is a time-consuming work (around 1 hour per minute of interaction). For a person who is experienced with *CA* and the relevant actions and phenomena identified for the project, such a labelling practice takes 7 minutes on average per minute. A link to the documentation of this labelling scheme is provided in the appendix B. It may be used and adapted for any large corpus of heterogeneous interactional data.

4 ANNOTATING INTERACTIONS

If the labelling system presented above has to deal with heterogeneity, within this section, the annotation system we present has to deal with the temporality and complexity of naturally-occurring interactions. We want to show what it's like to segment transcribed speech segments and ascribe annotated action against a turn-by-turn sequential analysis. These annotated actions must have their own inter-segment syntax as a mean to account for their temporal and sequential dynamics (normative expectation, abandonment, delaying, repair...). In the next subsections we show two samples from our corpus. The first sample will allow us to explain the mechanics of such an annotation practice. The second sample will show that other resources than talk may receive annotations, especially for the management of turn-taking and byplay participation framework.

4.1 Sample 1: a multi-threaded sequential infrastructure

When *CA* researchers perform a sequential analysis of a transcribed interaction, they proceed systematically on a turn-by-turn basis [39, pp.120-124]: they aim at reproducing the online analysis performed by the participants involved. The idea behind our time-aligned annotation practice is to formalize this analytical process as a mean of standardized annotations temporally embedded.

We will analyze the piece of data below, extracted from our corpus (see conventions in Appendix A). It is the very start of an interaction between two humans (Hum1 and Hum2) and the robot Pepper (Pep) in the university library. Pepper and the humans are in a face-to-face configuration, an eye-contact between Hum1 and Pep just happened. As a matter of readability the turns at talk were directly translated from French:

```
1 Pep : hi (.) can I help you?
2         (1.0)
3 Hum1: hi
4         ((hum1 and hum2 laugh))
5 Hum1: you alright? yes you can help me
6         (1.5)
7 Hum1: if you do not respond
8         (2.0)
9 Pep : how can I help you?
```

Figure 1 displays a graphical representation of how the annotations a rendered into the time-aligned annotation software ELAN. We will refer both to the simplified transcript and to the figure. The sequential annotations are results from sequential analysis (*vs.* behavior descriptions or speech transcription only). The vertical axle is temporal and its segmentation is homothetic. In the "Speech segments" stream, time segments correspond to utterances that can be isolated as actions (one line per participant). In the "Sequential threads" stream, sequential labels ("offer", "wait()"...) are annotated with segments aligned with speech segments. The "byplay" sequential threads use the same syntax, but it simply indicates that these actions are not addressed to the robot. The "threads" are populated depending on free space in the A-B-C order.

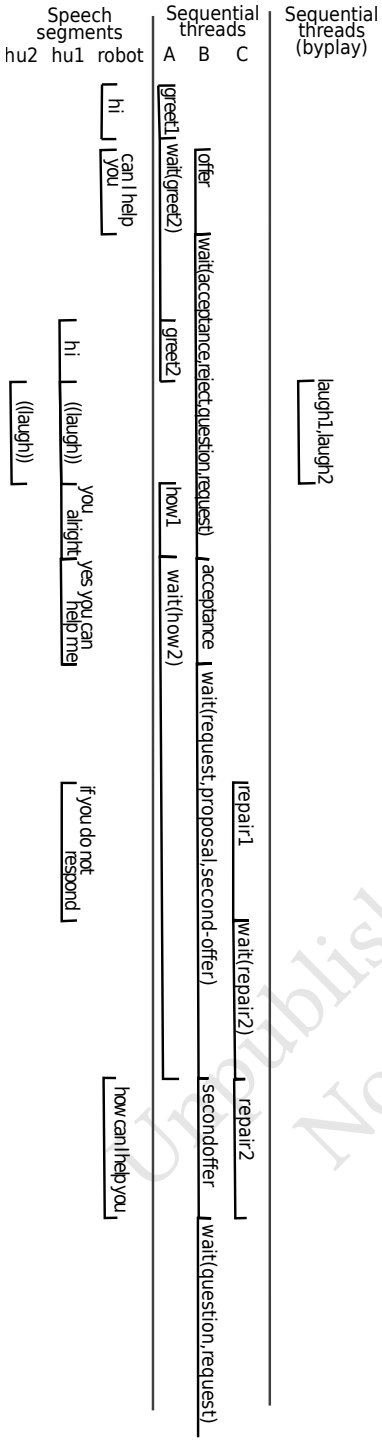

Figure 1: Multi-linear representation of concurrent sequential threads.

Here follows the turn-by-turn sequential analysis:

Line 1, Pepper produces two recognizable actions packaged into a single turn: a greeting ("hi") and a generic offer "can I help you?". These actions, if recognized as such, project two slots for responses: a second greeting, but also an offer acceptance OR rejection, OR some request/question [22, p.101]. These are the constraints/resources for the human to produce some next turn. These projections are annotated as "wait(relevant1,relevant2,...)" in the sequential thread A. That means that we give a context for the interpretation of the following silence (in order to discriminate this silence with one produced after a sequence completion).

Line 2, a silence is first noticeable: these long silences are recognized as a HRI [30] specificity and its metrics must be part of the data (in order to discriminate silences interpreted as the absence of response like e.g. line 6).

Line 3, the human produces the projected second greeting. The projection in thread A is stopped/complete. The other projection (in thread B) is maintained as still relevant.

Line 4, the two humans produce laughter in overlap (there are laughter sequences [15], like when a first makes a second relevant), in a byplay participation framework[17]. These are not addressed to the robot (separated in an other thread) but this activity (vs. a silence) account for delaying the projected actions in thread B.

Line 5, the human produces a non-projected new action (annotated in thread A) in a first segment of her turn ("you alright ?") which projects a new response (projection also annotated adjacently in thread A). Within the same turn, the projected offer acceptance is finally produced (annotated in thread B). Relevantly for a service encounter, an other sequence is now projected either from the robot initiation (a proposal or second offer), or the human may now produce a request.

Line 6, a silence has to be interpreted in the context of two projected types of responses (with two active threads). Line 7, the previous silence is designated as a failure. Thus it can be interpreted as a repair initiation (annotated in thread C), the repair being completed depending on the completion of the actions projected on thread A or B OR an account of the abandonment of the repair (like "sorry I didn't understand what you said"). Also, the human withdraws the projected possibility for her to produce a request as the turn-allocation to Pepper is reinforced. Line 8, an other rather long silence has to be interpreted in the context of three projected types of responses within three threads. Line 9, finally, Pepper produces the awaited second offer (designed as a question). The repair is completed, the relevance for a response to "how are you" is abandoned and new actions are projected...

What did we do here ? A lot of the complex dynamics that we analyzed do not appear on the annotated data, nor can be inferred from the sequential threads alone. For example, the dynamics between sequences themselves, informed by the study of service encounters, or silence categorization, nor did we address the fact that the howareyou-sequence had lower relevancy as it was situated in the first part of the human turn line 5.

We used the sequential annotations as a mean to reify actions and projections at the adjacency pair level only. The stacking structures mentioned earlier (2.2) may then be approximated, thanks to this multi-threaded approach, by the mean of probabilistic relations between segments and threads. We think that the human "simplistic" inferences about conversational agents (see 5) account for this sequence level of reification, whereas the larger sequential dynamics offer more space for negotiations.

4.2 Sample 2: more than text

As we mentioned earlier (2.2), as the interaction is a continuous process unfolded in time and physical co-presence, other bodily resources gain some relevancy for the meaning making at sequentially relevant slots. One exemplary case will show how multimodal resources can be analyzed for their contribution to turn-taking management, action ascription, and phenomenons that are specific to HRI such as suspended participation. In the transcription below, the human verbal response to the robot (a request) starts with "hum" line 12, which is 6,6 seconds after the robot's offer (line 1). Another human is behind:

```
01 Pep: Hi (.) can I help you ?
02       (0.4)
03 Hum: (1.0) ((starts torquing away))
04 Hum: ((laugh while orienting back towards pepper))
05      (0.9)
06 Hum: ((inhale demonstrably))
07      (0.3)
08       (0.2) ((starts torquing away))
09 Hum: but hum what do I ask?
10 Hum: ((orients back towards pepper))
11      (1.6)
12 Hum: hu::::m
13      (0.5)
14 Hum: I'm looking for a biology book.
```

If we consider this transcribed data with a verbocentric approach, we have to wait for the (rather recurrent) byplay question used as a delaying device as a cue (after 3,6 seconds of silence) in order to account for the fact that the human will try to respond to the robot. However, this turn (and the laughter l.4 as in 4.1) that the human addresses to another participant is recorder with a lower voice intensity, which could lead to the recognition of speech only line 12, after 6,6 seconds of silence which is rather long.

We may now consider all the bodily conducts and vocalizations produced in this interaction. We can see that it is only 0.4 seconds after Pepper's offer that the human turn around towards her fellow. A relevant bodily cue here is the fact that she accomplishes a *body torque* with the head directed towards the human behind while her legs are still oriented towards Pepper. This resource has been identified [33] as indicating an instability that project a short end: the head is oriented towards a temporary interaction goal (first a laugh l.4 then a question l.9) while the lower body part indicate the main interactional focus: interacting with Pepper. It results that the whole body is in a recognizable torque position with the shoulders and torso oriented towards nothing in particular: they appear sideways from Pepper's view. Moreover, the displayed inhalation in front of the robot (l.6) is also a cue of turn pre-beginning [29]. In other words, we can rely on these cues in order to recognize the fact that he human is actually preparing a response l.3, which is only 0,4secs after the robot's offer, and then have an additional cue l.6, after 2,8 seconds (vs. 3,6 or even 6,6 seconds of silence that could be interpreted as a disengagement).

This sample shows that a *torque*, if recognized, can contextually provide cues about what comes next (response relevance maintained), turn management (delay), participation (byplay). Being relevant for byplay sequences, its recognition could also inform us about practices where humans assess the robot's behaviour/response after sequence completion, as it is frequent in our corpus.

5 DISCUSSION: HUMAN'S PERSPECTIVE AND STATISTICAL PERSPECTIVE

If we consider the purposefulness of the use case, interactions might appear quite specific. But as humans appear to draw on generic resources for making sense of their first encounter with a robot, data show that they invoke basic sequences of action (offers, requests, proposals, questions, instructions, greetings, closings) as a way to secure their participation. This "basicness" was also a feature when the software was designed.

Our corpus suggests that humans already infer basic features, probably from the use of other conversational systems and devices [31]). They use intonation emphasis on what appears to be the most relevant keyword to recognize for them (see also [2]). They allow longer silences between turns (4.1 but also in [30]). They may even suspend the participation framework with Pepper at every moment, by the mean of torques such as in 4.2. They also perform less actions per turn, giving back the turn-at-talk to the robot, as in 4.1 where the offer acceptance is not immediately followed by the request as compared with human-human interaction. In other words, natural HRI show that acting in a simulacrum of conversation [4] raise recognized and established practices such as those exemplified above, which contributes to a better ecology between the human and the conversational agent.

By reifying sequences of actions, we do not aim at replicating an interactional competence, especially because humans do not use or learn statistically such sequential features (for e.g. see [9] for child acquisition). Our goal is to accompany the rational practical work accomplished by the human that is aware of being talking with a machine (see [1] for this *ethnomethodological* perspective).

One of the limits of the annotation system we proposed is the quantity of data that can be annotated, as qualified CA researchers must perform it. Once this qualitative-oriented annotation system is stabilized, we will also assess inter-rater reliability.

Actual natural language understanding models are able to learn predictive word models and to recognize intents, even with little data (thanks to few-shot learning) [3, 40]. As a perspective our work may improve these AI conversational systems by coupling the intentions detected by the system (learned thanks to our annotated data) with the turn taking sequences we began to identify, to make the conversation more natural. While some errors of the system can be tolerated and corrected by humans that adapt their behavior to an artificial entity (as observed in our data), we may even study how the robot can improve in the wild by interacting with human users and progressively refine the detected intention and sequences.

ACKNOWLEDGMENTS

The authors are grateful to the ASLAN project (ANR-10-LABX-0081) of the Université de Lyon, for its financial support within the French program "Investments for the Future" operated by the National Research Agency (ANR).

REFERENCES

[1] Morana Alac, Javier Movellan, and Fumihide Tanaka. 2011. *When a robot is social: Spatial arrangements and multimodal semiotic engagement in the practice of social robotics.* Vol. 41. https://doi.org/10.1177/0306312711420565 Issue: 6 Pages: 926 Publication Title: Social Studies of Science.

[2] Iuliia Avgustis, Aleksandr Shirokov, and Netta Iivari. 2021. "Please Connect Me to a Specialist": Scrutinising 'Recipient Design' in Interaction with an Artificial Conversational Agent. In *INTERACT 2021*. Springer Nature, 155–176.
[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *CoRR* abs/2005.14165 (2020). arXiv:2005.14165 https://arxiv.org/abs/2005.14165
[4] Graham Button. 1990. Going Up a Blind Alley. In *Computers and Conversation*. Elsevier, 67–90. https://doi.org/10.1016/B978-0-08-050264-9.50009-9
[5] Graham Button and Wes Sharrock. 1995. On simulacrums of conversation: Toward a clarification of the relevance of conversation analysis for human-computer interaction. In *The social and interactional dimensions of human-computer interfaces*. Cambridge University Press, New York, NY, US, 107–125.
[6] Arnulf Deppermann. 2011. The Study of Formulations as a Key to an Interactional Semantics. *Human Studies* 34, 2 (2011), 115–128. https://doi.org/10.1007/s10746-011-9187-8
[7] Arnulf Deppermann. 2013. Turn-design at turn-beginnings: Multimodal resources to deal with tasks of turn-construction in German. *Journal of Pragmatics* 46, 1 (2013), 91–121. https://doi.org/10.1016/j.pragma.2012.07.010 Conversation Analytic Studies of Multimodal Interaction.
[8] N. J. Enfield and Jack Sidnell. 2022. Action and Accountability in Interaction. In *Action Ascription in Interaction*, Arnulf Deppermann and Michael Haugh (Eds.). Cambridge University Press, Cambridge, 279–296. https://doi.org/10.1017/9781108673419.015
[9] Anna Filipi. 2018. Making Knowing Visible: Tracking the Development of the Response Token Yes in Second Turn Position. In *Longitudinal Studies on the Organization of Social Interaction*. Palgrave Macmillan, 39–66.
[10] Joel E. Fischer, Stuart Reeves, Martin Porcheron, and Rein Ove Sikveland. 2019. Progressivity for voice interface design. In *Proceedings of the 1st International Conference on Conversational User Interfaces - CUI '19*. ACM Press, Dublin, Ireland, 1–8. https://doi.org/10.1145/3342775.3342788
[11] Cecilia Ford and Sandra Thompson. 1996. Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In *Interaction and Grammar*. Cambridge University Press, 134–184.
[12] Cecilia E. Ford, Barbara A. Fox, and Sandra A. Thompson. 1996. Practices in the construction of turns. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)* 6, 3 (1996), 427–454. https://doi.org/10.1075/prag.6.3.07for
[13] Harold Garfinkel. 1967. *Studies in Ethnomethodology*. Prentice-Hall, Englewood Cliffs, New Jersey. https://doi.org/10.2307/2092244
[14] Nigel Gilbert, Robin Wooffitt, and Norman Fraser. 1990. Organising Computer Talk. In *Computers and Conversation*. Elsevier, 235–257. https://doi.org/10.1016/B978-0-08-050264-9.50016-6
[15] Philip J. Glenn. 2003. *Laughter in interaction*. Cambridge University Press, New York, NY, US.
[16] Charles Goodwin. 1986. Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies* 9, 2 (1986), 205–217. https://doi.org/10.1007/BF00148127
[17] Marjorie Harness Goodwin. 1990. Byplay: Participant structure and framing of collaborative collusion. *Réseaux* 8, 2 (1990), 155–180. https://doi.org/10.3406/reso.1990.3555
[18] Makoto Hayashi, Geoffrey Raymond, and Jack Sidnell. 2013. *Conversational Repair and Human Understanding* (cambridge university press ed.).
[19] Julian Hough. 2015. *Modelling Incremental Self-Repair Processing in Dialogue*. Philosophy. Queen Mary University of London.
[20] William Housley, Saul Albert, and Elizabeth Stokoe. 2019. Natural Action Processing. In *Proceedings of the Halfway to the Future Symposium 2019*. ACM, Nottingham United Kingdom, 1–4. https://doi.org/10.1145/3363384.3363478
[21] Kobin H. Kendrick, Penelope Brown, Mark Dingemanse, Simeon Floyd, Sonja Gipper, Kaoru Hayano, Elliott Hoey, Gertie Hoymann, Elizabeth Manrique, Giovanni Rossi, and Stephen C. Levinson. 2020. Sequence organization: A universal infrastructure for social action. *Journal of Pragmatics* 168 (2020), 119–138. https://doi.org/10.1016/j.pragma.2020.06.009
[22] Kobin H. Kendrick and Paul Drew. 2014. The putative preference for offers over requests. In *Requesting in Social Interaction*, Paul Drew and Elizabeth Couper-Kuhlen (Eds.). John Benjamins Publishing Company, 87–114. https://doi.org/10.1075/slsi.26.04ken
[23] Stephen C. Levinson. 2006. On the human" interaction engine". In *Roots of Human Sociality*. Routledge, 39–69. https://doi.org/10.1023/a:1018829907604
[24] Stephen C. Levinson. 2013. Action formation and ascription. In *The Handbook of Conversation Analysis*, Jack Sidnell and Tanya Stivers (Eds.). Blackwell Publishing Ltd, Chichester, UK. https://doi.org/10.1002/9781118325001.ch6
[25] Stephen C. Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology* 6, JUN (2015), 1–17. https://doi.org/10.3389/fpsyg.2015.00731
[26] Lorenza Mondada. 2011. Understanding as an embodied, situated and sequential achievement in interaction. *Journal of Pragmatics* 43, 2 (2011), 542–552. https://doi.org/10.1016/j.pragma.2010.08.019
[27] Lorenza Mondada. 2013. The conversation analytic approach to data collection. In *The Handbook of Conversation Analysis*, Jack Sidnell and Tanya Stivers (Eds.). Blackwell Publishing Ltd, Chichester, UK, 32–56. http://onlinelibrary.wiley.com/doi/10.1002/9781118325001.ch3/summary
[28] Lorenza Mondada. 2014. The local constitution of multimodal resources for social interaction Lorenza. *Journal of Pragmatics* 65, 2014 (2014), 137–156.
[29] Lorenza Mondada. 2016. L'énonciation comme phénomène émergent dans l'interaction : le cas des pre-beginnings. In *L'Enonciation aujourd'hui: un concept clé des sciences du langage*, Marion Colas-Blaise, Laurent Perrin, and Gian Maria Tore (Eds.). Lambert Lucas, Limoges, 317–340. http://edoc.unibas.ch/54978/
[30] Hannah R.M. Pelikan and Mathias Broth. 2016. Why that nao? How humans adapt to a conventional humanoid robot in taking turns-at-talk. *Conference on Human Factors in Computing Systems - Proceedings* (2016), 4921–4932. https://doi.org/10.1145/2858036.2858478 ISBN: 9781450333627.
[31] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (2018). https://doi.org/10.1145/3173574.3174214 ISBN: 9781450356206.
[32] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 4 (1974), 696–735. http://www.jstor.org/stable/412243
[33] Emanuel A. Schegloff. 1998. Body torque. *Social Research* 65, 3 (1998), 535–596. http://www.jstor.org/stable/40971262
[34] Emanuel A. Schegloff. 2007. *Sequence organization in interaction: Volume 1: A primer in conversation analysis*. Cambridge University Press, New York.
[35] Margret Selting. 2000. The construction of units in conversational talk. *Language* 29 (2000), 477–517.
[36] Wes Sharrock. 2000. Where the simplest systematics fits: A response to Michael Lynch's 'the ethnomethodological foundations of conversation analysis'. *Text* 20, 4 (2000), 533–539. https://doi.org/10.1515/text.1.2000.20.4.533
[37] Gabriel Skantze. 2020. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech and Language* 67 (2020), 101–178. https://doi.org/10.1016/j.csl.2020.101178 Publisher: Elsevier Ltd.
[38] Tanya Stivers and Jeffrey D. Robinson. 2006. A preference for progressivity in interaction. *Language in Society* 35, 03 (2006), 367–392. https://doi.org/10.1017/S0047404506060179
[39] Paul Ten Have. 2007. *Doing conversation analysis : A practical guide* (second ed.). SAGE Publications, London.
[40] Congying Xia. 2022. *Natural Language Understanding for Conversational Agents*. Ph. D. Dissertation. University of Illinois at Chicago.

A CONVENTIONS

Transcript conventions, drastically simplified from ICOR/Jefferson:

- `(.)` perceptible silence <200ms
- `(1.0)` mesured length of a silence >200ms in seconds
- : prolongation of the immediately prior sound (impressionistic representation with additional colons)
- ? a raising intonation (not a question mark per se, as raising intonations might appear at the end of other types of utterances)
- `((event))` events or conducts that could not be transcribed

B PROJECT DOCUMENTATION

- The state machine graphical representation of the dialogue system ad hoc version can be found here: https://page.hn/shhg8j
- The documentation of the labelling system for annotating shortclips can be found here: https://page.hn/0fkplh

Received 12 February 2023; revised 3rd March 2023; accepted 1st March 2023