philosophies

MDPI

Article

Enactivism and Robotic Language Acquisition: A Report from the Frontier

Frank Förster †

Adaptive Systems Research Group, School of Computer Science, University of Hertfordshire, Hertfordshire AL10 9AB, UK; frank.foerster@gmx.net

† Current Address: College Lane, University of Hertfordshire, Hatfield AL10 9AB, UK.

Received: 17 January 2019; Accepted: 1 March 2019; Published: 7 March 2019

Abstract: In this article, I assess an existing language acquisition architecture, which was deployed in linguistically unconstrained human–robot interaction, together with experimental design decisions with regard to their enactivist credentials. Despite initial scepticism with respect to enactivism's applicability to the social domain, the introduction of the notion of participatory sense-making in the more recent enactive literature extends the framework's reach to encompass this domain. With some exceptions, both our architecture and form of experimentation appear to be largely compatible with enactivist tenets. I analyse the architecture and design decisions along the five enactivist core themes of autonomy, embodiment, emergence, sense-making, and experience, and discuss the role of affect due to its central role within our acquisition experiments. In conclusion, I join some enactivists in demanding that interaction is taken seriously as an irreducible and independent subject of scientific investigation, and go further by hypothesising its potential value to machine learning.

Keywords: AI robotics; enactivism; language acquisition; human–robot interaction; developmental robotics

1. Introduction

In the last two decades or so, research into robotic language acquisition and symbol grounding has made some progress in providing machines with non-linguistic meanings for linguistic entities in certain limited domains, thereby solving the so-called symbol grounding problem in those domains. The vast majority of research has been restricted to what will be referred to as *referential words*. Referential words are those whose main function might be regarded as referring to classes of objects ('apple', 'box', 'square', 'dog'), object properties such as colour or size ('red', 'blue', 'small', 'heavy'), or temporally extended events, processes, and actions ('push', 'pull'), in the perceptible world outside of the agent's own embodiment. Our[1] own contribution in this area was to investigate possibilities to extend the domain of groundable words beyond the referential by focusing on the acquisition and grounding of negation words such as a simple 'no'. Given that language and meaning are central concepts in many branches of philosophy and cognitive science, the purport of this article is to elaborate on the philosophical underpinnings of that research rather than to report the technical details and results which has been previously reported elsewhere [1,2]. A more general account of an enactivist perspective in humanoid robotics has already been provided in [3], which is applicable to a number of architectures and approaches. This article seeks to provide a more detailed analysis of one particular architecture and its experimental deployment in human-robot interaction (HRI) in terms of the aforementioned core notions of enactivism. In doing so, I hope that such detailed analysis

[1] The plural pronouns 'our' and 'we' will be used to refer to the authors of [1,2]

can yield important insights regarding the suitability of cognitive theories in terms of informing experimental research on social phenomena such as language, and also that, vice versa, insights and detailed observations of cognitively oriented human–robot interaction may inform efforts to advance theories of social cognition. Section 2, following a short introduction into the research field, provides a short description of the language acquisition system under consideration. The theoretical background that motivated our experimental choices such as the addition of an affect-generating module, or the evaluation of participants speech in terms of their communicative function, consists of a mix of theories from various disciplines including speech act theory (SAT), pragmatics, developmental psychology, and conversation analysis. Despite initial scepticism as to whether the chosen approach would qualify as a genuinely enactivist one, there are many commonalities shared by both our architecture and experimental methodology, and central tenets of enactivism. Section 3 discusses whether this approach is compatible with the autopoietic branch of enactivism in particular, and analyses some central enactivist tenets from the perspective of a practising HRI and interaction researcher. Section 4 concludes with a discussion of the findings and hints towards future directions.

2. Theoretical Underpinnings and Methodological Choices

2.1. Robots Acquiring Language: Why and How?

There appear to be two main motivations for driving a renewed interest in language-related technologies in general, and robotics in particular. Advances in speech recognition brought about by deep learning, in combination with the advent of smart phones, speech-based assistants, and other speech-based interfaces, appear to have created the expectation that machines that understand human language are within our reach. In the case of robotics, there is the additional expectation that robots will soon be deployed in a variety of places that are inhabited by humans. These expectations combined with the fact that speech is one of humans' primary means of communication appears to render speech the 'coordination tool' of choice in mixed human–robot settings. Given that machines are still limited in their understanding of spoken human language[2] this is a strong motivation to focus language-related research.

Central within language-related research in robotics is the notion of *symbol grounding* [7]. The symbol grounding problem refers to the circumstance that a cognitive system cannot make sense of symbols or words based solely on manipulations of other symbols or words. The assumption is that at least a basic set of symbols or words needs to be grounded in something other than symbols in order for the system to "get off the symbol/symbol merry-go-round" [7], in order to use Stevan Harnad's apt expression. Robotics, in contrast to disembodied natural language processing (NLP) research, has the advantage of involving physical machines that *do* have access to "something other than symbols", namely sensorimotor data. Hence, many roboticists' answer to the symbol grounding problem has been to ground referential words in (derivatives of) sensorimotor data, with the most frequently used type of sensorimotor data being visual data.

Grounding in this context refers to the agent establishing a link between symbolic or linguistic entities, such as words, or multi-word utterances on the one hand, and data or data abstractions originating from extra-linguistic sources such as the agent's sensorimotor apparatus, on the other. A variety of AI techniques have been used to create grounding links between symbol and sensor data ranging from logical frameworks in symbolic AI (e.g., [8]) over sub-symbolic neural networks (e.g., [9]), to data-driven and, arguably, content-less forms of machine learning (e.g., [10–12]). Once such a

[2] There are still many occasions where spoken dialogue systems (SDS) fail dramatically. In robotics, the on-board speech recognition is often times sub-optimal due to the presence of noise or due to a limited speech corpus (cf. [4]). However, even assuming perfect speech recognition, a dialogue system may have too narrow a focus in terms of the application area that the designers envisioned (cf. Ward et al. for a wish-list for future dialogue systems [5]). Moreover, our understanding and modelling of conversational capabilities seem to lag far behind developments in speech recognition (cf. [6] for some revealing transcripts of 'conversations' between humans and voice interfaces).

link is established, we can say that a word has been *grounded*. One can conceptualise the grounding relationship as a technical form of association: the agent, when being presented with a grounded linguistic unit, can retrieve the unit's sensorimotor 'meaning', and vice versa: certain sensorimotor contexts may 'evoke' or lead to the retrieval of linguistic units to which these contexts are attached[3] (cf. [13] for a short discussion distinguishing symbol grounding from the related but distinct notions of conversational and societal grounding).

The vast majority of work in robotic symbol grounding focuses on the techniques of how to accomplish the world-to-word and word-to-world mappings for referential words and their referents. I am unaware of any other work which focuses on the grounding of non-referential words such as socio-pragmatic words and expressions such as 'hi', 'no', or 'bye', nor am I aware of any discussions as to whether these words should or could be grounded in a manner similar to the grounding of referential words.

A tacit yet important assumption that is typically not spelled out in the pertinent technical literature is that the referent and referring symbol are taken to temporally co-occur. Socio-pragmatic theories of language acquisition, especially the version supported by Michael Tomasello [14], provide grounds to believe that this assumption is reasonable for referential words and in the context of child-directed speech (CDS): word learning within these theories is thought to occur in joint-attentional frames between caretaker and child. If these theories are correct, then the temporal simultaneity between word and referent can probably be assumed. Whether or not this is the case for the acquisition of socio-pragmatic words, assuming we can construe something akin to a referent, and where triadic joint-attentional frames make less sense as explanatory framework, is an open question.

2.2. Robots Acquiring Linguistic Negation: Why and How?

The technical aspects and detailed results of our work on the acquisition of linguistic negation is described in [1,2] formed as part of the ITALK project [15]. Originally, the investigation was to focus on grammatical questions such as the lexical scope of negative 'operator' *not*. However, following the realisation that none of the contemporary psycho-linguistic theories have much to say about the acquisition of socio-pragmatic words, *no* being one of them, the scope was changed towards the lexical acquisition of negation words rather than grammatical issues. The fact that these type of words are rarely mentioned in the literature is surprising given the fact that socio-pragmatic words dominate toddlers' first ten words [16]. This indicates that these words are likely to play a crucial role in the earliest stages of language acquisition. As outlined in the previous section, it is genuinely unclear whether socio-pragmatic words should be grounded. It is conceivable that some socio-pragmatic words and expressions, such as 'hello', 'please', or 'bye', could be acquired via rote learning or imitation in which case they may not need any form of sensorimotor grounding. For negation words where even the earliest productions already fulfil a variety of communicative functions, most of which are bound up with affect, the case is more complicated. For the purposes of our studies, we assumed that negation words *should* be grounded, which leads to the question of how this may be accomplished.

Based on the observation that the earliest types of linguistic negation employed by toddlers are strongly related to affect or volition [17], we decided to integrate affect as a 'new ingredient' into an existing language acquisition architecture. Saunders and colleagues' architecture had at that time already been used in conjunction with the iCub humanoid robot [18] to acquire nouns, property names, and simple multi-word constructions based on linguistically unrestricted HRI [10–12]. The results from those experiments were used as a comparative baseline. Applying Occam's Razor to inform

---

[3] While some symbol grounding architectures ground linguistic entities directly in sensorimotor data, others may link them to abstractions thereof ('concepts'). In particular, logical-type architectures similar to the one developed by Siskind [8] involve multiple layers of explicit 'data abstraction'. In these cases, the established link is not a direct one. The derived concepts, however, are typically causally linked to the agent's sensorimotor context, such that a link between 'word' and 'world' can be postulated and just requires the analyst's willingness to go through several abstraction layers.

our design decisions, affect was introduced into the architecture so that the resulting changes would be as small as possible. The architecture and experiment are described in [1,2], and philosophical background and technical details in [19]. In the following, I summarise important elements of these past experiments in so far as they are relevant for the analysis in Section 3. A newly created minimal affect module was integrated into the architecture which modulates the robot's bodily behaviour, including the generation of facial expressions. Additionally, the architecture streams the new affective values to other components of the cognitive architecture, including those involved in symbol grounding. In comparison, in the previous architecture, the robot behaved somewhat randomly but believably in terms of its body movements without displaying any 'strong feelings' towards any of the present objects. In the modified architecture, the robot did bodily express clear preferences with respect to the present objects: affect values were object-bound or -triggered, and the object-to-value assignment was randomly changed between experimental sessions. The design rationale behind these decisions was that, for a start, it was more important *that* the robot displayed preferences rather than investigating why it would have those preferences in the first place. As the experiments' focus was on linguistic negation, we paid particular attention to any potential changes in participants' speech as triggered by the robot's affective behaviour. In terms of the experimental setup, an already established format was employed, in which participants would repeatedly interact with the robot as language tutors across several sessions. The robot would acquire language in-between sessions and based on participants' speech in the previous sessions.

Two aligned experiments were devised, the *rejection* [1] and the *prohibition experiment* [2], each of which tested separate but mutually non-exclusive hypotheses on the origin of linguistic negation. The *rejection experiment* investigated whether so-called *intent interpretations* could be a major source of the earliest negation words—chiefly 'no' in English. We encountered this notion in an early publication of Roy D. Pea [17], citing Joanna Ryan [20]. Ryan proposed that parental intent interpretations might be the source of children's first emotion words, and Pea extended this suggestion to include negation words. The *prohibition experiment* was designed as an extension of the *rejection experiment* and tested Spitz' hypothesis [21], which sees parental prohibition at the root of children's first active productions of negation. Both experiments differed from the baseline in that the session length was increased, and participants were told that the robot would have preferences with respect to certain objects. By keeping the experimental format from Saunders et al., the negation experiments were effectively rendered single-blind, meaning that participants were unaware of the true topic under investigation: the overt task of teaching the robot object labels was a mere side-task.

In terms of the results, the magnitude of the difference in frequency of negative utterances and words produced surpassed our initial expectations considerably. Participants in both negation experiments produced a large amount of negative utterances, whereas participants in the baseline had produced very few to none. However, none of the two hypotheses on the origin of linguistic negation could be excluded, as both intent interpretations and prohibitions abounded in negation words. As a consequence, the robot picked up on these words and started to produce them in subsequent sessions with felicity rates, as judged by external coders, of circa 66% in the case of the rejection experiment. Another important observation was that approximately two-thirds of intent interpretations were produced while the robot was in a negative affective state. Whilst not in perfect alignment, correlations of this size indicate a relationship far from random; therefore, intent interpretations may be an example of relative coordination (cf. [22]).

3. Enactivism

Enactivism is typically attributed to Varela, Thompson, and Rosch's book *The Embodied Mind* [23]. It can be construed as an attempt to introduce phenomenological notions such as 'experience' into cognitive science [24]. Already in *The Embodied Mind*, Varela et al. pitch enactivism against alternative theories of cognition such as cognitivism and connectionism, and characterise cognition as "(a) history of structural coupling that brings forth a world" [25]. This characterisation may appear overly opaque,

so it may help to look at the approach via some central notions which can be found in many enactivist publications: autonomy, sense-making, embodiment, emergence, and experience [22,26]. The meaning of these within enactivism will be sketched in the respective sections below.

Since its inception, at least three related, but distinct branches of enactivism have developed, *autopoietic* (e.g., [26]), *sensorimotor* (e.g., [27]), and *radical enactivism* (e.g., [28]), each of which puts different levels of emphasis on these different notions. Within the following discussion, our approach will be compared mainly against literature from the *autopoietic* branch as authors from this branch have recently started to extend their theories to encompass social cognition [22].

In order to evaluate our experimental approach with respect to its enactivist credentials, I will start by analysing it along these five topics. The section will close with a short analysis of the notion of *affect* within enactivism due to its relevance in language acquisition.

3.1. Autonomy

The meaning of *autonomy* in enactivism is slightly more specific than in robotics or artificial intelligence. Most roboticists would probably agree to label a robot autonomous when it is not remote controlled and performs its tasks, or exhibits otherwise functional behaviour, for extended periods of time without human control. Within the enactive framework, an organism only qualifies as autonomous if "it is composed of several processes that actively generate and sustain an identity under precarious conditions" [22]. Linked to enactive autonomy is the notion of *operational closure*, where a system is seen to be operationally closed if, "among the enabling conditions for any constituent process in the system, one would always find one or more other system processes" (ibid.). This definition relies on the notion of *identity*, which is somewhat opaque. In this context, it is important to note that social systems in particular can give rise to more than one identity on more than one level. As a singular system, the iCub robot, while qualifying as autonomous within our experiments in the robotics sense of the word, would not qualify as autonomous in the enactive sense. The reason for this failure is the lack of 'constituent processes' that actively try to sustain its identity *as a robot*. In other words, the robot cannot count as an enactive autonomous system because it lacks substantial self-repair processes. It also does not actively work against human attempts to switch it off, where the switch-off can be seen as an act of obliteration of its systemic identity as a robot. The threshold for being enactively autonomous is set so high that I am unable to name any existing artificial physical machine that would qualify as enactively autonomous. The same holds true for the majority of simulated robots as these typically do not exert control over their life conditions in terms of maintaining the simulation within which their identity exists. Neither do we typically want them to do so. This lack of enactive autonomy also extends to Di Paolo's own creations, as also in evolutionary robotics, it is the designer who controls the simulations within which the created creatures exist and not the creatures themselves. The only artificial systems that might qualify as enactively autonomous are certain types of computer viruses whose main purpose is self-replication.

On the other hand, going with De Jaegher and Di Paolo's idea that social systems can, and typically have multiple identities, we can analyse social robots under their identity as social interactors. In addition, the interaction they contribute to bring into existence may qualify as separate identity. Under the interactor identity, the prospect of being autonomous in the enactive sense of the word looks somewhat more realistic. Our language acquisition system, however, even under the guise of interactor, would still not count as autonomous as it does not actively work to maintain an ongoing interaction. In its current form, it lacks the necessary sensing capabilities and feedback loops required for establishing interactional autonomy. However, the planned future directions which we will sketch in Section 4 have the potential to render it interactionally autonomous despite its lack of a self-preservational identity.

3.2. Sense-Making

The notion of *sense-making* in autopoietic enactivism differs in important ways from more established meanings of *meaning* or *sense* in the philosophy of language, or computer science:

Sense-making in autopoietic enactivism is biologically grounded, and refers to a system's exchanges with the world that contribute to the "continuity of the self-generated identity of identities that initiate the regulation" [26]. In that sense an organism's metabolism appears to be one of the most fundamental forms of *sense-making*. It is unclear how this notion links up to more widely established meanings of 'sense' in linguistics and cognitive science. Indeed, Mylin and Hutto, as proponents of the radical enactivist branch, are somewhat critical of the use of this notion. In their view, the use of the word may lead to confusion given their assertion that low-level cognitive activities have no need for content or meaning in the representational sense of the word [28].

In the context of language acquisition, the notion of sense-making is mainly relevant in its recent extension into the social domain: *participatory sense-making* [22]. De Jaegher and Di Paolo conceive of participatory sense-making as "the coordination of intentional activity in interaction, whereby individual sense-making processes are affected and new domains of social sense-making can be generated that were not available to each individual on her own" (ibid.). I take it that the meaning of *sense* in these new domains may resemble more its standard meaning in linguistics and philosophy of language. This interpretation is supported by the fact that the authors later cite a conversation analytical example of talk-in-interaction where the presence of a long pause significantly impacts on both interlocutors' interpretation, or sense-making, of what is being meant. The authors go on to discuss the example as illustrative of the coordination processes at work, where the participation in talk is less a matter of second-guessing the other interlocutor's intentions, and more a matter of participating in the conversation in a timely manner, where relevant intentions will manifest themselves eventually. This links up well with this author's observations regarding the meaning of negation words, especially 'no', where conversational timing appears to be a major factor in determining its meaning. For example, the question of whether a 'no' constitutes an instance of truth-functional negation, or an instance of rejection, can hinge on the timing of its production—timing relative to the production of the preceding human turn[4]. Importantly, the notion of meaning in our, but presumably also De Jaegher's and Di Paolo's account, is not necessarily bound up with truth-functional content. In instances where rejective 'no's are used, for example, we don't see a need to postulate the presence of propositional content in the interlocutor's mind when producing such an utterance. This is not to say that content is not necessary or present in other forms of talk.

3.3. Embodiment

Embodiment can be a difficult criterion to evaluate for two reasons. First, some may bend its meaning such that it loses most of its significance, for example by calling virtual agents embodied with respect to some virtual world. The second relates to cases of so called trivial embodiment, where the importance of having a body and its function with respect to the processes under investigation is neglectable. Language acquisition approaches that employ designer–robot interaction as the ultimate test-bed is problematic for the second reason. By choosing a system designer or a trained person as human interlocutor, the researcher decides to give away one of the greatest resources of an HRI setup: genuine interaction itself. Both designers as well as trained persons will modify their behaviour in a way that either fits with the acquisition algorithms, or with the perceived or actual expectations of the experimenter. To use Dumas' and colleagues' term: the interactor thereby ceases to function as a *human clamp* [29]—at least to some degree. This does not categorically exclude the possibility of genuine interaction taking place, but it endangers its occurrence severely. The onus for the human actor will be on "doing the right thing" to make the interaction work from what is perceived to be the experimenter's vantage point, as opposed to doing "what feels right", or allowing to "be pulled" into an interaction. This is an important reason why we decided in favour of both naïve participants, and a

[4] A detailed analysis of potential confusions of coders with respect to negation words that occurred in our studies and the reasons behind these confusions are provided in [19].

blind experimental format in our experiments. In this way, it is effectively impossible for participants to guess our expectations and to try to fulfil them.

There are two important aspects of embodiment that we utilise. Firstly, volition or affect are expressed by modulating both the robot's facial features and its bodily behaviour: its affective state determines whether it will engage in grasping, disinterested, or rejective behaviours respectively. Secondly, and only within the prohibition experiment, we encouraged participants to touch the robot by pushing its arm away when it approached a forbidden object. Both bodily aspects appear to have triggered considerable changes in participants' talk: they uttered a multitude of negation words which had hardly occurred in the comparative baseline. Therefore, the two aforementioned aspects of embodiment were crucial features of the experiment which in all likelihood caused the observed changes in participants' speech.

3.4. Emergence

According to Di Paolo et al., emergence is "the formation of a new property *or* process out of the interaction of different existing processes or event" [26] (emphasis added). In order to distinguish emergent processes from simple aggregates, the authors state that the former need to fulfil the following additional two properties: (1) they need to have an autonomous identity, and (2) they need to "sustain (...) this identity, and the interaction between the (...) process and its context must lead to constraints and modulation in the operation of the underlying levels" (ibid., brackets indicate the present author's omissions). The last property, emergent processes imposing constraints and impacting the constitutive processes, are also emphasised by De Jaegher et al. [22]. While I conceive of talk of new 'identities' as rather opaque, my interpretation of this definition is the following: properties of an emergent process can only be observed on the level of that process and not by analysing its constituent lower-level processes in separation from each other. The "downward constraint and modulation" property is no magical power on the part of the emergent process, but can be explained quite simply with: it takes two to tango. In order for an emergent process to occur, say a dance, the constituent processes, the dancers, need to follow certain rules and behave in certain ways. Thus, we only see a dance happening when the dancers restrict their action repertoire to the one required by, or compatible with, the respective dance. Apart from the deliberate choice of the dancers to, say, not read a book, while dancing, after having agreed to engage in a dance, there are other, less deliberate restrictions. Restrictions pertaining to the single actor's "freedom of movement", for example, may be imposed by the dynamics of dance, caused by the physical forces at play. The latter will make certain moves either physically demanding or impossible, which, in the absence of a dance, might have been viable actions to take by each single dancer.

Social interactions such as conversations impose similar constraints on the interlocutors. Some of these are culture-specific, for example saying "thank you" when being handed a gift, whereas others seem to be more implicit or subconscious, and appear to have a more universal character. Examples of the latter are the conversational repair mechanisms that have been documented by conversation analysts such as Sacks and colleagues [30], or the one-second time constraint for conversational moves hypothesised by Jefferson [31].

In the interactions within our language acquisition experiments, we observed at least two emergent processes and properties. The first process is the one of the interaction itself. While we don't have a name for this particular process, it consists of the interplay of utterances and bodily displays, moves and expressions on both sides, human as robot. We can tell that what we observed was indeed an emergent process by pointing to the downward constraints that it imposed on the interlocutors. When classifying the meaning of utterances of 'no', as produced by the robot, it became apparent that the coders' interpretation of their meaning often hinged on the precise timing of the utterance. As timing is a property of the conversation rather than of any singular constituent utterance, the observed conversational interaction is an emergent process. Importantly, the difficulty in classifying such utterances experienced by the coders is reflected in the interactors' behaviour themselves. As we

know from many instances of conversation analyses, interactors orient their interpretations of the interlocutors' utterances with respect to constraints such as timing thresholds with demonstrable consequences for the ensuing interaction sequence. In other words, utterance timing can in certain instances be crucial in determining the future direction of the conversation because it can influence how a certain utterance is interpreted by the interlocutors.

The second set of emergent properties or processes are the aforementioned intent interpretations, which themselves involve a temporally constrained interplay between the bodily expressions and utterances of both interactors. This is an exemplary case where the addition of a new kind of bodily behaviour on the lower level, affective displays produced by one interlocutor, gives rise to a new type of inter-subjective interaction on the higher level with consequences for both interactors. The sequences involving intent interpretations appear to exhibit relatively tight temporal constraints: the affective displays of the child-like robot trigger participants' production of these interpretations within a relatively small time window. The fact that the two moves appear to be temporally coupled in this manner indicates that intent interpretations are indeed a case of social coordination.

3.5. Experience

Experience in the phenomenological sense may be the enactivist theme that some of us roboticists are least familiar with. During the execution of the experiments, the experimenters strived to stay as emotionally neutral as possible in order not to influence our participants. Once the experiments had completed, the analysis was performed in terms of countable artefacts such as word types, utterance lengths, prosodically marked utterance locations, etc. in order to be able to perform statistical analyses. Despite these attempts to 'objectify' our findings, there are at least two occasions where something akin to first-person experience became relevant. The first occasion is linked to the employed assessment methodology. In order to judge a negative utterance on both its felicity or adequacy in the respective situation, and when judging the same utterance for its functional type, say rejection vs. truth-functional denial, we rely on coders' judgements. Coders can issue these judgements due to them being fluent English speakers, but also, more implicitly, due to them being humans and thereby 'fluent interactors'. In this sense, human coders, like the human participants themselves, may act as a human dynamic clamp [29]. The latter function cannot be qualified precisely because we, as scientists, still only have a limited understanding of the dynamics of conversation. For this reason, I would argue that, by relying on human coders in letting them judge, at times intuitively, on the felicity or adequacy of certain utterances, we effectively rely on their experience as interactors—experience in the phenomenological sense.

The second pointer to the relevance of experience, which is more anecdotal, is the observation that some interactions *feel* like 'working' interactions, whereas others feel awkward and stilted. When shown recordings of the fluent interactions, some observers approached us and pointed out the quality of the interaction. The fact that people by and large appear to agree that a given interaction is a 'working' interaction, as opposed to a dysfunctional one, indicates that these judgements might be linked to the phenomenological notion of experience, as opposed to just being made on a purely random basis.

3.6. Affect or Volition

Given that affect was the 'crucial ingredient' in our attempt at modelling language acquisition, I shall discuss its role in our system vis-à-vis to its postulated role in enactive theories of cognition. Generally speaking, the importance of affect in enactivism does not seem to be en par with the central five notions discussed above. The radical enactivists Hutto and Myin, in their attempt to purge cognitive science from its dependency on propositional content, do not discuss the notion at all [28]. The autopoietic enactivists, which we have used as our main point of reference, discuss it mainly in the context of value systems. For the purpose of this paper, I will therefore equate 'value' with 'affect'. Di Paolo et al. [26] are very critical of *static* value systems that assign values to certain worldly

artefacts or processes independently of the sense-making process. The affect module in our architecture would certainly fall under this category as it not only assigns values to such objects independently of the ensuing interaction, but even randomly. Di Paolo et al. define value as "the extent to which a situation affects the viability of a self-sustaining and precarious network of processes that generates an identity" (ibid.). Given that our employed affect module was a first attempt at determining the effect of emotional expressions onto participants' speech, and was never thought to be anything else than a stub, I do not feel strongly about this kind of criticism. On the contrary, in Section 4, I will outline future plans that appear to be more compatible with Di Paolo and colleagues' view on the role of value in cognitive systems.

At the same time, however, I am critical of their view in declaring, that value could never originate from anything but the sense-making process. Taking a developmental perspective, I argue in favour of a pluralistic view which acknowledges that something has to get the interaction off the ground in the first place. Adopting this view, I deem it likely that initial, possibly even random values, assigned to objects or processes are conceivable, and that these may eventually be superseded or replaced by values and affect originating from aspects of the interaction process, once the interaction is established. Without assigning the process of social interaction itself an initial value that logically and temporally precedes this process, I cannot conceive of how the interaction would get off the ground in the first place.

4. Discussion and Future Work

As can be gleaned from the last section, there are several commonalities between our approach to robotic language acquisition and the tenets of enactivism. Enactivists regard interaction as an emergent system that is irreducible to its constituent parts—the interlocutors, and potentially sub-personal processes, in the case of a conversation. I agree that interaction should indeed be conceived of as a rich and dynamic resource for a language learner where certain interpersonal phenomena hinge on the temporal flow of the dynamics, intent interpretations being one such phenomenon. In this sense, phenomena such as intent interpretations are irreducible due to their inherently dynamic character but also because they involve more than one component from the sub-conversational level. This inherent dynamism, however, does not prevent us from analysing recordings of such phenomena iteratively. When engaged in such an analysis, I would not hesitate to refer to, and name the constituent parts of the process, and the role they play in the temporal unfolding of the conversation. In this sense, I am considerably less critical than Di Paolo, De Jaegher and colleagues with respect to explanations of dynamic conversational phenomena in terms of their constituent processes—"boxology" in their words [22,26].

In terms of sense-making, our approach seems to be largely in agreement with enactivism. Agents-in-interaction do remain independent autonomous systems in the sense that they do not lose their capacity to terminate the interaction, the interaction being the emergent process. On the other hand, interactions do impose constraints on the agents in the sense that conversational rules restrict the types of utterances that can be produced at any given time in the conversational sequence. This phenomenon has been well documented by conversation analysis in the previous half century. Within our own studies, we were able to document occasions where participants lost their composure when faced with certain negative utterances on the part of the robot, despite their awareness of talking to a machine, and despite them trying to maintain severe self-imposed restrictions in terms of their own word choices. Observations such as these document the *pull* that a conversation can exert on its participants, an example of what some enactivists refer to as 'downward causation' [26].

In terms of the role of affect or 'value', we are in some disagreement with the autopoietic enactivist branch in that we allow affect to derive from systemic components that are not directly linked to sense-making. Our future plans, however, might lead us into a position of stronger agreement (see below).

A severe drawback of using a content-free memory-based learner as a central learning mechanism is that the meaning of words, currently consisting of a multitude of grounded exemplars held within the memory of the learner, never 'congeal' into concepts or prototypes. While the lack of conceptual representation renders the architecture content-free, this 'laziness' [32] comes at a price. We observed that, once the robot is perceived to have reached a certain mastery of words, many participants switch from using content-related words to more encouraging expressions such as "well done". Such words and expressions are meaningless to the current system because all extracted words are assumed to be groundable in sensorimotor-affective data. Grounding these type of words in such data is most likely nonsensical and certainly leads to a deterioration of the system performance.

Future Work

Purely associative learning, as we have employed it, is probably too minimalist a choice for any advanced forms of language learning. I therefore intend to supplement the system with some form of reinforcement learning in the widest sense, with three potential sources of reinforcement. Firstly, reinforcement could be generated when extra-linguistic goals are met as a consequence of successful language use. An example for such a goal would be the closing of a window in response to the utterance "could you close the window, please". Secondly, reinforcement could possibly be derived intra-conversationally through prosodic features corresponding to a lexical "well done". Thirdly, positive reinforcement could be attributed based on the efficacy of conversational moves depending on whether they contribute to the perpetuation of the conversation itself. The last suggestion seems to be particularly well-aligned with Di Paolo's and De Jaegher's view that value ought to be derived from appraisals of processes that are involved in maintaining the emergent process.

5. Conclusions

I have analysed some enactivist core tenets vis-à-vis the methodological decisions that we made with respect to a language acquisition system and the experiments in which it was deployed. Summarily, our approach appears to be mostly compatible with major tenets in enactivism. Disagreement mainly originates from rather radical assertions made by authors of the autopoietic branch with respect to the origin and role of affect. Despite initial scepticism regarding the suitability of enactivist theorising in the realm of social interaction, the introduction of *participatory sense-making* into the enactivist canon by De Jaegher and colleagues appears to extend the reach of the framework toward higher-level and interpersonal cognitive phenomena such as language. The enactive framework can serve as a useful tool for both motivating and scrutinising methodological and architectural decisions in AI research involving social interaction. I support De Jaegher's and Di Paolo's demand to take interaction more seriously than is typically the case in AI robotics, and would like to go even further by suggesting that regularities of the dynamic interaction process may serve as substrates for machine learning purposes.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

NLP	Natural Language Processing
AI	Artificial Intelligence
CDS	Child-Directed Speech
HRI	Human–Robot Interaction
SDS	Spoken Dialogue Systems

References

1. Förster, F.; Saunders, J.; Nehaniv, C.L. Robots that Say 'No'. Affective Symbol Grounding and the Case of Intent Interpretations. *IEEE Trans. Cogn. Dev. Syst.* **2017**, *10*, 530–544. [CrossRef]
2. Förster, F.; Saunders, J.; Lehmann, H.; Nehaniv, C.L. Robots Learning to Say 'No': Prohibition and Rejective Mechanisms in Acquisition of Linguistic Negation. *arXiv* **2018**, arXiv:1810.11804.
3. Nehaniv, C.L.; Förster, F.; Saunders, J.; Broz, F.; Antonova, E.; Köse, H.; Lyon, C.; Lehmann, H.; Sato, Y.; Dautenhahn, K. Interaction and experience in enactive intelligence and humanoid robotics. In Proceedings of the 2013 IEEE Symposium on Artificial Life (ALIFE), Singapore, 16–19 April 2013; pp. 148–155.
4. de Jong, M.; Zhang, K.; Roth, A.M.; Rhodes, T.; Schmucker, R.; Zhou, C.; Ferreira, S.; Cartucho, J.; Veloso, M. Towards a Robust Interactive and Learning Social Robot. In Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, Stockholm, Sweden, 10–15 July 2018; pp. 883–891.
5. Ward, N.G.; DeVault, D. *Ten Challenges in Highly-Interactive Dialog System*; Turn-Taking and Coordination in Human-Machine Interaction: Papers from the 2015 AAAI Spring Symposium; The AAAI Press: Palo Alto, CA, UAS, 2015.
6. Porcheron, M.; Fischer, J.E.; Reeves, S.; Sharples, S. Voice Interfaces in Everyday Life. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018.
7. Harnad, S. The Symbol Grounding Problem. *Physica D* **1990**, *42*, 335–346. [CrossRef]
8. Siskind, J.M. Grounding the Lexical Semantics of Verbs in Visual Perception using Force Dynamics and Event Logic. *J. Artif. Intell. Res.* **2001**, *15*, 31–90. [CrossRef]
9. Sugita, Y.; Tani, J. Learning Semantic Combinatoriality from the Interaction between Linguistic and Behavioral Processes. *Adapt. Behav.* **2005**, *13*, 33–52. [CrossRef]
10. Saunders, J.; Nehaniv, C.L.; Lyon, C. The acquisition of word semantics by a humanoid robot via interaction with a human tutor. In *New Frontiers in Human-Robot Interaction*; Dautenhahn, K., Saunders, J., Eds.; John Benjamins Publishing Company: Amsterdam, The Netherlands, 2011; pp. 211–234.
11. Saunders, J.; Lehmann, H.; Sato, Y.; Nehaniv, C.L. Towards Using Prosody to Scaffold Lexical Meaning in Robots. In Proceedings of the 2011 IEEE International Conference on Development and Learning (ICDL), Frankfurt am Main, Germany, 24–27 August 2011.
12. Saunders, J.; Lehmann, H.; Förster, F.; Nehaniv, C.L. Robot Acquisition of Lexical Meaning—Moving Towards the Two-word Stage. In Proceedings of the 2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL), San Diego, CA, USA, 7–9 November 2012.
13. Kennington, C.; Plane, S. Symbol, Conversational, and Societal Grounding with a Toy Robot. *arXiv* **2017**, arXiv:1709.10486.
14. Tomasello, M. *Constructing a Language*; Harvard University Press: Cambridge, MA, USA, 2009.
15. Broz, F.; Nehaniv, C.L.; Belpaeme, T.; Bisio, A.; Dautenhahn, K.; Fadiga, L.; Ferrauto, T.; Fischer, K.; Förster, F.; Gigliotta, O.; et al. The ITALK Project: A Developmental Robotics Approach to the Study of Individual, Social, and Linguistic Learning. *Top. Cogn. Sci.* **2014**, *6*, 534–544. [CrossRef] [PubMed]
16. Fenson, L.; Dale, P.S.; Reznick, J.S.; Bates, E.; Thal, D.J.; Pethick, S.J. Variability in Early Communicative Development. *Monogr. Soc. Res. Child Dev.* **1994**, *59*, i-185. [CrossRef]
17. Pea, R. The Development Of Negation In Early Child Language. In *The Social Foundations of Language and Thought: Essays in Honor of Jerome S. Bruner*; Olson, D., Ed.; W.W. Norton: New York, NY, USA, 1980; pp. 156–186.
18. Metta, G.; Sandini, G.; Vernon, D.; Natale, L.; Nori, F. The iCub humanoid robot: An open platform for research in embodied cognition. In Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems, Gaithersburg, MD, USA, 19–21 August 2008; pp. 50–56.
19. Förster, F. Robots That Say 'No': Acquisition of Linguistic Behaviour in Interaction Games with Humans. Ph.D. Thesis, University of Hertfordshire, Hertfordshire, UK, 2013.
20. Ryan, J. Early language development: Towards a communicational analysis. In *The Integration of a Child into a Social World*; Richards, M.P.M., Ed.; Cambridge University Press: Cambridge, UK, 1974.
21. Spitz, R.A. *No and Yes: On the Genesis of Human Communication*; International Universities Press: New York, NY, USA, 1957.
22. De Jaegher, H.; Di Paolo, E.A. Participatory sense-making. *Phenomenol. Cogn. Sci.* **2007**, *6*, 485–507. [CrossRef]

23. Varela, F.J.; Thompson, E.; Rosch, E. *The Embodied Mind: Cognitive Science and Human Experience*; MIT Press: Cambridge, MA, USA, 1991.
24. Wilson, R.A.; Foglia, L. Embodied Cognition. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Spring 2017 ed.; Metaphysics Research Lab, Stanford University: Stanford, CA, USA, 2017.
25. Varela, F.J.; Thompson, E.; Rosch, E. *The Embodied Mind: Cognitive Science and Human Experience—Revised Edition*; MIT Press: Cambridge, MA, USA, 2016.
26. Di Paolo, E.A.; Rohde, M.; De Jaegher, H. Horizons for the Enactive Mind: Values, Social Interaction, and Play. In *Enaction: Toward a New Paradigm for Cognitive Science*; Steward, J., Gapenne, O., Di Paolo, E.A., Eds.; MIT Press: Cambridge, MA, USA, 2010.
27. O'Regan, J.K. *Why Red Doesn't Sound Like a Bell: Understanding the Feel of Consciousness*; Oxford University Press: Oxford, UK, 2011.
28. Hutto, D.D.; Myin, E. *Radicalizing Enactivism: Basic Minds without Content*; MIT Press: Cambridge, MA, USA, 2012.
29. Dumas, G.; Lefebvre, A.; Zhang, M.; Tognoli, E.; Kelso, J.S. The Human Dynamic Clamp: A Probe for Coordination Across Neural, Behavioral, and Social Scales. In *Complexity and Synergetics*; Müller, S., Plath, P., Radons, G., Fuchs, A., Eds.; Springer: Cham, Switzerland, 2017.
30. Sacks, H.; Schegloff, E.A.; Jefferson, G. A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language* **1974**, *50*, 696–735. [CrossRef]
31. Jefferson, G. Preliminary notes on a possible metric which provides for a "standard maximum" silence of approximately one second in conversation. In *Conversation: An Interdisciplinary Perspective*; Roger, D., Bull, P., Eds.; Multilingual Matters: Clevedon, UK, 1989; Chapter 8, pp. 166–196.
32. Aha, D.W. (Ed.) *Lazy Learning*; Springer-Science+Business Media: Dordrecht, The Netherlands, 1997.