Action coordination and learning in dialogue*

Arash Eshghi, Christine Howes, and Eleni Gregoromichelaki

1 Introduction: The challenge of conversational AI

Conversational Artificial Intelligence (AI) systems (such as Amazon Alexa, Microsoft Cortana, Apple Siri) have recently become ubiquitous and are now an integral part of our everyday lives. There have been huge advancements recently in the achievement of conversational AI with many claims regarding the closeness of attaining the goal of artificial general intelligence (AGI) based on these successes (see e.g. Bommasani et al., 2021). Nevertheless, in practice, the scope of this success has been limited. End-users of such systems often treat them in the same way they would another human in that they have expectations of naturalness, intelligence, flexibility, and smooth interaction, leading regularly to disappointment and frustration (Moore, 2017; Clark et al., 2019; Chaves and Gerosa, 2021; Park et al., 2017; Luger and Sellen, 2016; Fischer et al., 2019) because these systems do not offer this expected potential.

The reason for this is that natural language (NL) use in general, but especially in conversation, presents numerous challenges that have been traditionally distinguished and isolated from each other as pertaining to various encapsulated modules. For example, autonomous domains of competence such as syntax, semantics, and pragmatics are distinguished while non-verbal aspects of NL processing like facial expressions, eye gaze, manual gestures are ignored, as are the effects of the physical and cultural environment. This standard strategy of separating phenomena and treating them as encapsulated modules with idiosyncratic vocabularies led rule-based approaches to an inability to integrate seamlessly the various assumptions that are required for the resolution of various challenging aspects of processing in dialogue. As a result, open-ended and multi-domain artificial conversational systems, in particular, have been found to be unmanageably complex, brittle, and unreliable.

The advent of end-to-end neural architectures suggested that the challenge of successful meshing of all aspects of multimodal processing in dialogue could be overcome (see Vinyals and Le, 2015; Serban et al., 2016; Li et al., 2017; Lowe et al., 2017; Wolf et al., 2019, a.o.). End-to-end dialogue systems are trained directly on large amounts of conversational data, learning a mapping from dialogue history to a system response, either in a supervised or unsupervised fashion, without modularisation of conversational knowledge. Such systems are robust and general with respect to the domains they are designed to deal with. Nevertheless, it seems that progress has stagnated and that the provision of even larger amounts of data will not improve the situation (see e.g. Lowe et al., 2017; Zadrozny, 2021), even when using large-scale, state of the art, Transformer-based models (Vaswani et al., 2017; Devlin et al., 2019) pretrained on dialogue data (see e.g. Bao et al., 2020; Noble and Maraev, 2021; Caldarini et al., 2022).

Recent large-scale end-to-end neural systems (e.g. Wolf et al., 2019), while displaying impressive capacities with regard to producing fluent surface structures, do not adequately capture human capacities in learning *appropriately adaptive* conversational behaviours. Often the responses of such

*We would like to thank Stergios Chatzikyriakidis and Jean-Philip Bernardy for constructive comments on an earlier draft of this chapter, infinite patience, and ensuing discussion. The content of this contribution has emerged from our collaborations over many years with our friends: thank you Ruth Kempson, Greg Mills, Pat Healey, Julian Hough, Matthew Purver, Oliver Lemon, Robin Cooper, Staffan Larsson, Mehrnoosh Sadrzadeh, Simon Dobnik, Ellen Breitholtz, Asad Sayeed, Graham White, Jonathan Ginzburg among others.

systems are generic, uninformative, and neglectful of the overall coherence of a dialogue in that they take into account only the immediately previous turn(s) thus lacking consistency with respect to the longer history of the dialogue and its future prospects with respect to achieving some goal (see e.g. Li et al., 2020; Vinyals and Le, 2015; Shang et al., 2015; Sordoni et al., 2015). As a result, they can also be unreliable with respect to trustworthy responses because, as they predict single utterances at a time, they ignore the purposeful nature of action in dialogue in the service of achievement of local and global goals. On the whole, today's conversational AI systems are *static* in that they are unable to adjust to the dynamic environment of the dialogue history and evolving goals and do not come equipped with strategic skills to enable them to negotiate the ambiguity, vagueness, and nuances of human-to-human conversation, and thus adapt to new people, tasks, and situations.

We argue below that what is needed is a radical reconceptualisation of linguistic models away from conceptions of NL as a shared code (Sec. 2). Instead, we suggest viewing NLs as sets of skills for goal-driven (inter)action and coordination in order to exploit *affordances* in the socio-material environment of an agent (Sec. 3). Efforts should thus focus on linguistically implemented feedback mechanisms such as *repair* that allow interacting agents to coordinate their actions (Sec. 4). In terms of models, neural or otherwise, such a move should enable designs that are able to *actively adapt* to previously unseen situations. In terms of artificial agents, it should enable their using the immediate local feedback from their environment including their interlocutor. Thus this approach goes well towards the direction of freeing conversational interfaces from the shackles of the data on which they were trained. Crucially, this move should involve dynamic, predictive models that are able to actively engage with the world, observe the effects of what they do/say and not just in how the conversation moves forward. In addition, they need to be checking the effects brought about in their *multimodal*, physical situation of utterance. In the face of prediction error, model architectures would, on this view, allow real-time, focused and local updates to their parameters. Doing this using just underlying large-scale neural language models seems very promising but remains an open problem in terms of the requisite neural architectures, attention mechanisms, training objectives as well as the nature of the data. In Sec. 5, we present a low-level, dynamic model of NL interaction and coordination. The model, DS-TTR, is a combination of Dynamic Syntax (Kempson et al., 2016, 2001) and Type Theory with Records (Cooper, 2005; Cooper and Ginzburg, 2015), that we argue satisfies the desiderata listed above. In Sec. 6 we present a couple of case studies showing how such a model can be implemented and used in practice to bootstrap interaction.

2 The inadequacy of code models

On the one hand, what artificial conversational architectures show is that the complexity of NL behaviour is underappreciated due to the apparent ease with which people handle their everyday interactions. As a result, human communication is often modelled under the 'code model', namely, as one agent coding and transmitting a message (the 'sender') with reception and decoding at another agent (the 'receiver'). This approach has failed spectacularly to account for the complexity and subtlety of sense-making in human interaction (see, e.g., Fowler and Hodges, 2016). Models of communication which assume idealised perfect speakers and listeners sharing mental representations of interpretations of meaning – dating back to Shannon and Weaver (1949), but still underlying much research today, can only ever be an abstraction, and one that we argue is detrimental to understanding successful communication (see also Rączaszek-Leonardi et al., 2014).

On the other hand, the backlash against this simple-minded approach led to models of high modularity, domain specificity, and complexity. This was due to fact that it was deemed necessary to enhance the code model with individualistic recursive reasoning about others' mental states, as in Gricean, Neo-Gricean, and Post-Gricean accounts of NL, and accompanying plan-based and belief desires and intention (BDI) dialogue models (e.g. Grosz and Sidner, 1986; Matheson et al., 2000). Recent responses to the ineffectiveness of end-to-end conversational AI include pleas to return to

such highly complex intentionalist approaches (e.g. Kopp and Krämer, 2021). But these approaches were in fact the reason of failure of rule-based/symbolic dialogue systems with explicit hand-crafted, but in the end intractable, 'mind-reading' components (see e.g. Gregoromichelaki et al., 2011; Mirski and Bickhard, 2021).

We believe that progress in dialogue modelling is impeded due to such standard assumptions that still underlie much research in linguistics, cognitive science, and AI. These assumptions are shown to be unsustainable, when we consider dialogue and interaction both for traditional rule-based approaches and modern neural architectures. Standard theories of communication rely on a separation between speaker and hearer, with the speaker encoding and transmitting a message, and the hearer decoding it. Even in intention-based accounts, speaker and hearer share the linguistic 'code' (the language, some NL) and the only possibility for accommodating the function of errors is to characterise them as "noise" to be eliminated. Successful communication is characterised as the hearer correctly discovering the message which the speaker intended to convey, and this is assumed to be the norm of what actually happens. This basic assumption underlies most psychological and pragmatic theories of interaction including the Interactive Alignment Model (Pickering and Garrod, 2004, see below), Gricean pragmatics (Grice, 1975) and Relevance Theory (Sperber and Wilson, 1995) which assume an underlying literal meaning enhanced by context-specific pragmatic inferences to uncover the *speaker*'s intention.

However, the actions of participants in dialogue form a system of coupled components (see, e.g., De Jaegher and Di Paolo, 2007) with the result that *feedback mechanisms*, like constant error indication and adjustment, are crucial for the stability, maintenance, and self-organisation of the system. Given the moment-by-moment need for action coordination, participants do not need explicit representations of others' or their own mental states, as correctly assumed in deep learning models, and neither do they need to converge on a shared 'code' or shared criteria of success. Instead, their conceptions and contributions need to be complementary to sustain a social practice whose normative character is defined externally to their own private or explicit rationalisations of their behaviour.

Rethinking our conception of successful communication away from shared codes or sufficiently similar mental representations puts the flexibility and dynamism of NL at the heart of communication. As Healey et al. (2018a) state "[i]nstead of thinking of effective communication as formulating a "perfect" message, it becomes about finding optimal ways to uncover and address misunderstandings". We go further and do not characterise these practices as uncovering 'misunderstanding' or 'miscommunication', terms suggesting that they are somehow in opposition to some common understanding and common ground. Instead, we characterise successful coordination (rather than "communication") as the local, incremental resolution of inevitable perturbations in the self-organisation of a complex dynamical system enabling people to contribute to larger social organisations that constitute their ecological niche ('form of life'). From a psychological perspective, the rapidity and highly incremental nature of turn-taking exchanges in dialogue (Levinson and Torreira, 2015; Sacks et al., 1974) shows that intractable exhaustive reasoning about some optimal local outcome is not what participants aim for (cf. Frank and Goodman, 2012). Instead, practices of navigating through, and local adjustment to, an incrementally evolving landscape of affordances provided by the ecological niche and participants' own actions enable the forms of distributed cognition observed in dialogue (e.g. Dingemanse, 2020). Transferring this insight to the domain of language technology, this assumption partially explains the limited success of language models in mimicking many aspects of human performance. We attribute the substantial current shortcomings of such models to the limited variety of data they are exposed to, i.e., lack of multimodal data (see e.g. Hanjie et al., 2021; Hill et al., 2020,?; Ruis et al., 2020; Röder et al., 2021; Lappin, 2021), lack of ability to actively interact with the data (cf. Li et al., 2017; Lewis et al., 2017) so lack of feedback, and their lack of physical embodiment (see e.g. Pustejovsky and Krishnaswamy, 2021). From this perspective, we suggest that progress in modelling human dialogue and conversational AI requires a radical reconception of NLs as mechanisms for (inter)action.

2.1 Human-human dialogue

The simplifications of characterising communication as attempts to adjust the replication of mental states are shown to be inadequate when we consider dialogue as shown in example (1), taken from the British National Corpus (BNC: Burnard, 2000). Units of meaning are co-created incrementally (Kempson et al., 2016; Hough et al., 2015) by multiple interlocutors using incomplete utterances (e.g. line 7 – Purver et al., 2011), with phenomena such as cross-person compound contributions (where one person continues another's utterance, as in lines 7 and 8 – Lerner, 1991; Howes, 2012), repairs (e.g. the clarification requests in lines 4 and 6 – Sacks et al., 1974; Purver, 2004), and disfluencies (e.g. the pause and restart in line 9 – Hough, 2015) – seen as 'performance errors' in traditional linguistics – becoming crucial in the sense-making activities of the participants.

(1)
a. **J:** Can you think of any catalysts?
b. **A:** Er is it potassium permanganate?
c. **J:** <unclear>
d. **A:** What
e. **J:** Pla <pause> a duck billed
f. **A:** Pardon?
g. **J:** A duck billed
h. **A:** Platypus.
i. **J:** And it's not platypus it's <pause> sounds like a type of pen.
j. **A:** Platinum.
k. **J:** Right, platinum. (BNC; file FMR 728-737)

This short extract in which a chemistry tutor (J) prompts a student (A) to answer the question posed in line 1, neatly illustrates the characteristic divergence and convergence that is key to driving dialogue forwards. From a standard individualistic perspective, one can characterise the exchange as indicating that, from J's perspective, A's response in line 2 is not the expected answer – it is divergent with it. A finally produces the expected answer (thus demonstrating convergence with J's expectations) in line 10. This is a valid way of describing the process and it might be the way that a single participant might rationalise or abstract the dialogue process into a narrative that they construct post hoc.

However, from a realistic modelling perspective, it neglects the fact that both participants operate in a context (a 'teaching context') that imposes normative constraints in what their actions should be aiming at as they perform the roles assigned to them by that sociocultural convention that constitutes the practice they are enacting. There are no 'teacher' or 'student' roles outside this socially-afforded context. The practice the participants engage in thus constitutes their (temporary) identities and action possibilities afforded to them. So, both participants' actions are now subsumed under the overall normative perspective that their actions should be relevant to the elicitation of some particular answer to a question posed by J, with both of them operating as a coherent system performing complementary actions towards that goal and compensating for each other's failings to contribute appropriately.

This normativity is imposed to the participants because there is a joint goal, not only between the participants but including a goal-driven process of the societal 'form of life' in general pervading the interaction. On the other hand, none of the two participants on their own has an overview of exactly what this overall goal consists in and how it can be achieved even though they become aware of their obligations and opportunities as they are enacting their roles. In effect, the cognition required for achieving this goal is distributed (Hutchins, 1995) not only across the participants' individual capacities but, crucially, the sociocultural environment that provides the state space and the normativity, correctness or incorrectness, of their joint action trajectory.

This distributed and systems perspective shows that the ecological sociocultural environment in which the interaction takes place directs the participants' unreflective but, nevertheless, fluent navigation towards the goal. The to and fro where A's indications of how far they can reach with respect to

contributing to the goal is explicitly conveyed not only by assertions but also by clarification requests and completable utterances (Gregoromichelaki et al., 2020b), elsewhere characterised as "fragments". As is shown here, this is not a case of communication 'breaking down' but of opportunities to engage in further enactment of exactly the 'teaching' practice they are engaged in. J compensates by revealing affordances that might be obscured from A's perception by making manifest more local, perhaps in another context seemingly irrelevant, "incorrect", affordances, which are functional in this particular environment to allow both of them to reach a point of sufficient satisfaction of their mission: after a cue in line 5 fails to elicit the required convergence, J exploits the predictability of the compound noun phrase 'duck-billed platypus' to get A to produce the first syllables of the answer to the original question. This is an illustration in miniature of the learning and developmental process that Gibson (1966) calls the 'education of attention'.

This management of the divergent and convergent contexts with respect to the normative imperatives of the sociocultural environment is incrementally and locally managed, with a hierarchy of joint goals and subgoals emerging in an unplanned and opportunistic fashion as each participant makes contributions that create opportunities for the other to make timely and appropriate responses and compensatory moves when such goals seem to be threatened (Howes and Eshghi, 2021). Such uncertainty with respect to what the activity exactly consists in, what concepts are relevant, how the practice is going to develop, in general, what the affordances are, is built-in in our awareness and visible in the talk's surface: teacher and student can only have probabilistic expectations as to what they are required to do moment-by-moment and this is the most effective strategy they should adopt given the usual uncertainty of the environment (otherwise, their behavioural adjustments will be threatened by "overfitting"). But they can trust the process unreflectively because they have available strategies, built into NL practices, that will correct and adjust their performance based on the feedback received.

In this dialogue, there is an asymmetry between the speakers, as J, the tutor, is both the expert, and more socially powerful than A, the student. But, in fact, this asymmetry is endemic, diagnostic of not just all child/adult (Duveen and Psaltis, 2013; Kunert et al., 2011) or expert/non-expert exchanges (such as tutoring dialogues or doctor/patient consultations, Lu et al., 2007; Pilnick and Dingwall, 2011), but all interactions. Differences in experiences, cultural background, individual physiology, and social communities all contribute to differences in our NL use, meaning that we never share the "same" language as anybody we nevertheless successfully interact with (Clark, 1998). This raises an important practical question: How can we communicate successfully when individual differences in language use are not the exception but the norm?

3 Language as action

We argue that the answer to this question relies on reconceptualising NL as a set of skills for interaction (Kempson et al., 2016; Gregoromichelaki et al., 2019, 2020b). This recasts language use in *actionist* terms, in parallel with recent actionist theories of perception (Nöe, 2004; Bickhard, 2009). Actionism holds that perception is not a series of snapshots of scenes in the world leading to their inferential manipulation as representations in the brain (Marr, 1982). Rather, perception is engagement with the world – an embodied agent activity and achievement.

The motivation for this perspective starts with the assumption that, in order to survive, organisms have to play an active part in controlling their environment and keeping it within desirable states (thus embodying systemic principles like self-maintenance, self-organisation, autopoiesis, see, e.g. Di Paolo, 2008; Di Paolo and De Jaegher, 2012). For an organism to exert such control, its adaptation to its environment equips it with abilities to perceive predictable relationships between its actions and ensuing perceptual stimulations (*sensorimotor contingencies*) since the purpose of perception/action is to ensure agent effective adaptability.

Under this view, adaptive exploration and exploitation of environmental resources makes use of the agent's practical and embodied know-how of such sensorimotor contingencies, i.e., direct

perception-action links (see, e.g, Buhrmann et al., 2013; Maye and Engel, 2011) rather than brain-internal cognitive inferential or representational resources. Sensorimotor contingencies are lawful regularities in the dynamic relation between the agent and the environment, the ecological niche, patterns of dependence of changes in the sensory input as a function of an agent's movements (Gibson, 2014). Consequently, the information agents perceive about entities and their potential for interaction outcomes is expressed in terms of *predictions* and it is perspectival in the sense that it is agent-relative. It is also mediated through the invocation of complex regular patterns, *constraints* (Barwise and Perry, 1983; Rączaszek-Leonardi and Scott Kelso, 2008), originating from social as well as natural learning experiences but not via internal skull-bound world models. Various such learned expectations (hierarchically-organised sequences of predictions) based on memorised holistic patterns of experience (*policies*) are built up through reiterated interactions with the environment and are then deployed in subsequent encounters (see also Bickhard, 2009). For human agents, in addition, learning to perceive refers to what is offered through their direct time-extended interactions with the sociocultural environment, which is a significant constitutive part of the human ecological niche. Therefore "perception" of an entity will then be primarily constituted by the set of expectations it invokes concerning the possible interactions enabled through it (its *affordances*), not some objective individuation or categorisation as a type of entity. Agents are not passive perceivers but act to realise the predictions (anticipations) they expect to receive as feedback from the environment, thus predictions can also serve as goals (see e.g. Gregoromichelaki et al., 2020c; Friston et al., 2012a). Such predictions can be confirmed or disconfirmed through the feedback so that they are the basic source of learning This view is intended to replace the static, internalist-inferential view of "perception" as the association of stimuli with mental symbols stored and recovered as propositional knowledge.

Analogously, competence with NLs does not require an abstract representational level or language of thought (Fodor, 1975), but can be viewed in terms of the linguistic and non-linguistic actions (utterances and, e.g., gestures) that can be performed in particular situations. In any type of engagement with others or the environment, an agent acts via NL means in order to perceive the predicted consequences of their interactions, instead of constructing and refining representations of these interactions to serve as guidance for its action. Such predictions are generated by means of the agent's embodied sensorimotor knowledge of the relevant sociocultural niche, i.e., by routinised anticipations (the 'grammar' in a Wittgensteinian sense, e.g., Forster, 2009), of how its various actions will change features of the sociomaterial world. For individual agents, such predictions are shaped and constrained by what is licensed within the current sociomaterial context, i.e., within the *normativity* of the socially-distributed nature of the grammar.

This means that no individual agent can be solipsistically aware of the significance of its own action: by observing its consequences (*feedback*), the very act of speaking (or writing) in a particular context reveals to participants (potentially abstractions over) the normatively constrained triggers of actions for the words used as well as generating structured anticipations of further possible developments, the latter thereby becoming further affordances within that conversational exchange. Thus a concrete action has both backward effects, in that it shapes the dialogue history under a particular conceptualisation, and forward effects, i.e., it opens up new trajectories in the current landscape of affordances. This is a more radical version of the empirically derived notion of the three-position "architecture" of conversation in Conversational Analysis (CA) or the retroactive and proactive effects of utterances (see e.g. Arundale, 2008, 2020). Since any action interpretation in dialogue has provisional status, only a probabilistic distribution over effects is ever possible. Hence, so-called 'repair' processes (i.e. feedback) are not confined to the highly noticeable explicit attempts, like asking for clarifications or correction, but it is a constant feature of interaction.

3.1 Exploiting probabilistic uncertainty in interactions

It is now becoming widely accepted that certainty either over interpretations or action outcomes is neither a feasible goal nor a criterion of success for human interaction. Both uncertainty and the

variety of multiple affordances in the human ecological niche introduce complexity due to the fact that agents do not perceive only one affordance at a time. Agents always perceive a continuously restructured dynamic landscape of affordances that consists of various possibilities for action soliciting their attention. Cisek & Kalaska (2010) propose that 'affordance competition' is resolved by humans and animals through active moment-to-moment exploration of the field of available affordances without realising an overall plan of action but by being drawn towards the most rewarding predicted outcomes.

Regarding the contribution of individual agents, the Skilled Intentionality Framework (Rietveld et al., 2018) reconstructs Friston's framework that is underpinned by the Free-energy Principle and active inference (Friston, 2010, 2011; Mathys et al., 2011) in non-representational, ecological, and action-oriented terms. Based on Bayesian statistics and machine learning approximate Bayesian inference, the free-energy principle is a proposal for modelling living self-organising systems like humans and other animals. The framework built around this principle assumes that living organisms are equipped with generative models generating top-down predictions about causes of received perceptual input from their environment. In order for such organisms to maintain themselves successfully in their environment, they constantly seek to reduce the prediction error that ensues due to discrepancies between their predicted sensory input and the actual ("bottom up") input they receive from the environment. Long-term reduction of prediction error ('minimisation of free energy') can be achieved by either changing the generative model (*perceptual inference*) or acting in the world to change the sensory input received (*active inference*). In the domain of human cognition, Friston's framework has received entirely solipsistic interpretations conceiving of the generative model as inducing brain-internal representations encoding information about an inaccessible external environment (Hohwy, 2013) in the same way that a neural network or distributional language model can be construed as a system performing learning and inference in isolation of its environment and hence facing the 'symbol grounding problem' (Harnad, 1990; Gao et al., 2018; Kottur et al., 2016).

However, a more plausible non-cognitivist interpretation is that an agent's generative model reflects the attunement of the agent's embodiment to its physical environment, for example, by establishing the regulation of its metabolic needs ('homeostasis'). For a more complex social agent, the generative model, in addition, incorporates embodied assumptions of normativity, i.e., the regular, expected ways of acting in the practices the agent participates in (e.g. Kirchhoff and Froese, 2017; Bruineberg et al., 2018a,b). From this perspective, rather than encoding information about an inaccessible environment, neural states contribute to the embodied capacities of changing the environment through action. The goal of active inference is to steer an agent's interactions with the ecological niche in such a way that the agent's actions harmonise with the affordances of the sociomaterial environment. Perceptual inference under this interpretation affects the agent's internal (endogenous) dynamics and can be conceptualised in terms of inducing patterns of action-readiness. Since, at any moment, a whole landscape of affordances confronts complex agents, there needs to be a way for the agent to select the relevant set of affordances that is predicted to yield the most rewarding outcome. In the Ecological Psychology literature, such relevant affordances are termed 'solicitations' to distinguish them and emphasise the agent's perspective and contribution to the determination of affordances, which are environment-agent relations. The Skilled Intentionality Framework proposes that the solicitation of multiple complex affordances towards humans can be modelled as triggering states of 'action readiness' (Frijda et al., 2014) within individual organisms. These are affective states, rather than the explicit formation of 'goals', 'intentions' and the like. Thus perceptual inference regulates action readiness as the agent is motivated to act based on its disattunement with the environment (its prediction error) which has an emotional effect on the agent's awareness.

Perceiving (i.e. predicting) complex nested structures of potential affordances constitutes, in our terms, *conceptualisations* of the situation as offered by the grammar and perceived by an agent at a particular time. Developing competence with the grammar, in the sense of an agent being solicited by appropriate action-inducing potential, requires training and developing skills. For human agents, this is accomplished through participation in 'practices', i.e., coordinated patterns of behaviour of

multiple individuals, within which NL interactivity is arguably the canonical case. Individuals or groups of individuals can then respond selectively to relevant (sets of) affordances since they have become attuned to the normativity of each particular situation. As a result, they act under the guidance of resolving 'affective tensions', i.e., emotional responses like feelings of discontent or dissatisfaction, rather than "rational" deliberations through propositional beliefs/intentions. Such feelings of tension are aroused by the discrepancies (overwhelming prediction failure, i.e., prediction error) between a concrete situation and the embodied skills of perceiving the norms of the situation type that the agent(s) have acquired by training. Agents resolve such tensions by resorting to their expertise and acting accordingly. Their familiarity with the interactive environment allows them to intervene and restore perception of the expected affordances of the situation type.

4 Action coordination in dialogue

On the view proposed here, NL behaviours are understood as practices, with their normativity underpinned by a set of conditional actions (the 'grammar') inducing ongoing emergent flows that can be approximated in more individualistic, abstract, and detached terms as the often-studied notions of context, content, intentions, speech acts and the like. On the present view, NLs, both in terms of syntactic structure and conceptualisation potential, are first and foremost coordinative action-control devices both with respect to the environment and other individuals; and a grammar formalism is duly determined directly in terms of defining the normative constraints (i.e. setting out and traversing the landscape of predicted affordances) that operate top-down to guide such action (see also Trafford, 2017; Zadrozny, 2020).

Affordances which, under our interpretation are publicly available resources, trigger motivations for action within agents (*solicitations*, e.g. Dreyfus, 2013). However, affordances are not, as standard, simply properties of the environment. Instead they are relations between agent abilities and what the current sociomaterial environment reliably makes available. This means that the shifting set of affordances in dialogue concerns the collective potential of the interactants, rather than individual perspectives whose meshing needs to be explicitly negotiated/represented. Interlocutors thus acquire a joint perspective as long as they operate as a system with its own self-organisation underpinned by prediction error minimisation (as modelled within the Free Energy Principle framework in its ecological/enactive interpretation, e.g., Bruineberg et al., 2018a). So the local and shifting landscape of affordances and the state and abilities of the agents involved determine at each moment a demarcated 'field of affordances', i.e. a subset of the landscape of affordances that are perceived as relevant by the agents. This provides for a joint conceptualisation of the current action potential with minute adjustments at each subsentential stage resulting in the appearance of planned rational action at the macro-level and as strategically introduced repair of intention recognition failures as in (2) and (3).

(2) (a) A: so …umm this afternoon …
(b) B: let's go watch a film
(c) A: yeah

(3) (a) A: I'm pretty sure that **the**
(b) B: **programmed visits**?
(c) A: programmed visits, yes, I think they'll have been debt inspections. [BNC KS1 789-791]

However, the function of what have been characterised as overt repairs is not some extraordinary feature of just some dialogue exchanges. The function and maintenance of a complex dynamic system requires constant interaction with the environment and adjustment of the participants' action/perception by reducing their independent potential while non-summatively maximising their joint capacities, otherwise it will just be the juxtaposition of two independent agents acting on their own.

4.1 Repair

The function of feedback in a coupled system is a primary regulatory factor in subsuming the individual components under a system architecture. Independently, the components will have available a multitude of degrees of freedom. In order to interact successfully and tractably, degrees of freedom need to be mutually constrained and this is achieved by the agents performing complementary and compensatory actions in the service of joint action (De Jaegher and Di Paolo, 2007; Paolo et al., 2018). So balancing and counterbalancing a complex but unified process can only be achieved by continuous work that ensures the self-organisation of the system. Most research on dialogue considers repair – specifically clarification requests such as the "what?" and "pardon?" in (1) – to index misunderstandings between individuals (mismatches between people's takes on the dialogue). In our view, however, repair as a separate category of constructions (Clark, 1996), turns out to be an artefact of assuming that interlocutors aim for the establishment of shared common world "representations", with speech acts contributing propositional contents (Poesio and Rieser, 2010; Ginzburg, 2012) in the service of reasoning and planning. These assumptions, which we argue are fallacious when interaction is properly characterised as skilled action use, also underpin the currently popular Rational Speech Acts model (RSA; Frank and Goodman, 2012; Goodman and Frank, 2016), which assumes that speakers reason over others' (presumed) intentions. A global RSA model does not seem to be computationally tractable (e.g. Cohn-Gordon et al., 2018), while the general availability of local repair mechanisms have been demonstrated to remove the need for such higher-order modelling in agent simulations (Van Arkel et al., 2020). This is not to deny that people can reason over others beliefs, desires and intentions (BDI). Rather, we claim that this is a higher-order skill not a necessary foundation for successful interaction (Gregoromichelaki et al., 2011).

This stance also inverts the usual assumptions about *backchannels*, which are considered to be "positive" feedback, signalling understanding (Fujimoto, 2007). On our view, a backchannel passes up an opportunity for so-called "repair" (Schegloff, 1993) or, in our terms, transforms the field of affordances in a monotonic manner. Such signals therefore acquire their myriad functions as a direct consequence, depending on the action in progress when the backchannel is produced. For example, if the speaker is telling a story, a backchannel may function as a *continuer*; if giving directions, it may *acknowledge* identification of a landmark; and if offering an opinion, it may indicate *agreement*. This position – supported by experimental evidence (Howes et al., 2012; Healey et al., 2018b; Mills, 2007; Mills and Healey, 2006) – means that rather than treating backchannels as multiply ambiguous, and completely opposite to clarification requests, we can unify them as *procedural* mechanisms for managing the types of transformations induced moment-by-moment in the field of affordances (Howes and Eshghi, 2021).

We turn now to a formalism that, we argue, captures such a notion of repair as a natural consequence of the incremental and domain-general architecture in terms of affordances that is assumed to underlie NL grammar.

5 Dynamic Syntax and Type Theory with Records (DS-TTR)

Dynamic Syntax (DS; Cann et al., 2005; Kempson et al., 2001, 2016) is a constraint-based (or model-theoretic, Pullum and Scholz, 2001) grammar architecture that models the dynamic, real-time, incremental interpretation of word-sequences (comprehension) or linearisation of contents (production) relative to a fine-grained concept of dialogue context (see Sec. 5.4 below). The DS syntactic engine, including the lexicon, is underpinned by a specialised version of Propositional Dynamic Logic (PDL), which is a multimodal logic able to express probabilistically licensed transition events (*actions*) among the states of a dynamic system (Sato, 2011; see Fig. 1 where outgoing edges/actions from each node form a learnable (Eshghi et al., 2013b) probability distribution conditioned on the current state or DS tree). As a result, DS is articulated in terms of conditional and goal-driven actions whose accomplishment either gives rise to expectations of further actions, tests the environment for further

contextual input, or leads to abandonment of the current strategy due to its being unviable in view of more competitive alternatives (see Fig. 1). In current versions of DS, words, morphology, and syntax are all modelled as *affordances*, i.e., indicators of opportunities for (inter-)action (Gregoromichelaki, 2018; Gregoromichelaki et al., 2019, 2020b,a). Both participants' opportunities for action, as well as their perspectives, are modelled in a unified model of the whole system, rather than assuming that the grammar is an individualistic mechanism inside one person's head. Participants' interactions are modelled as incrementally opening up a range of options so that selected alternatives can be pursued either successfully or unsuccessfully: even though a processing path might look highly favoured initially, due to the changing conditions downstream, it might lead to an impasse so that processing is aborted and backtracking to an earlier state is required (Sato, 2011).

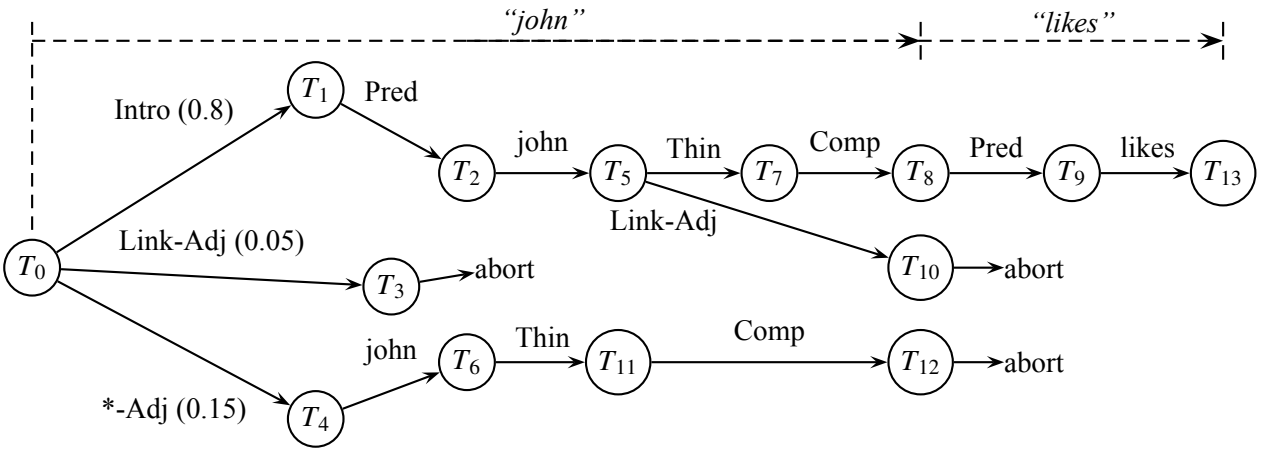

Figure 1: DS-TTR parsing as a Directed Acyclic Graph (DAG): actions (edges) are probabilistic transitions between partial trees (nodes).

Given these inherent properties, DS has lent itself particularly well to dialogue modelling and analysis in the past decade or so (see Purver et al., 2006; Gargett et al., 2009; Gregoromichelaki et al., 2011; Howes, 2012; Eshghi et al., 2015; Kempson et al., 2016; Howes and Eshghi, 2021, among others). Dialogue is modelled as the incremental and interactive composition of action sequences triggered by words either from oneself (in production) or an interloculor (in comprehension) in an incrementally evolving context, enabling unitary explanations of ellipsis (Kempson et al., 2015), self-repair (Hough and Purver, 2012), split utterances (Howes et al., 2011; Howes, 2012; Kempson et al., 2016), clarification requests (Gargett et al., 2009; Eshghi et al., 2015) and other feedback (Howes and Eshghi, 2021).

5.1 Type Theory with Records (TTR)

Recent efforts (e.g., Eshghi et al., 2012; Purver et al., 2011, 2010) have incorporated TTR (Cooper, 2012, 2005) and probabilistic versions of TTR (e.g., Hough, 2015; Hough and Purver, 2014b, 2017; Hough et al., 2018) as the conceptualisation formalism assumed by DS. Here we assume a version with types ('concepts') reinterpreted in DS dynamic terms as types of PDL actions (programs) (Gregoromichelaki et al., 2020b,a) – it is within this so-called DS-TTR fusion that we express our models below.

TTR is an extension of standard type theory, and has been shown to be useful in contextual and semantic modelling in dialogue (see e.g. Ginzburg, 2012; Fernández, 2006; Purver et al., 2010, among many others), as well as the integration of perceptual and linguistic semantics (Larsson, 2015; Dobnik et al., 2012; Yu et al., 2016). With its rich notions of underspecification and subtyping, TTR has proved crucial for DS research in strongly incremental semantic specification (Purver et al., 2011; Hough, 2015), as well as specification of richer concepts of dialogue context (Purver et al., 2010; Hough,

2015). Furthermore, Hough and Purver (2014a, 2017) use a probabilistic variant of TTR (Cooper et al., 2015) in combination with DS to flesh out a model of probabilistic inference for incremental reference processing and it has been shown, on this basis, how robotic design can benefit from modelling the perception of affordances of objects in the environment (Hough et al., 2018, 2020). But DS-TTR is in principle compatible with any version of probabilistic versions of TTR and other such frameworks (e.g. Cooper et al., 2014)

5.1.1 TTR: a quick formal introduction

In TTR, conceptual structures are specified as *record types*, which are sequences of *fields* of the form $[\ l : T\ ]$ containing a label l and a type T. Record types can be witnessed (i.e. having an instantiation, hence 'true') by *records* of that type, where a record is a sequence of label-value pairs $[\ l = v\ ]$. We say that $[\ l = v\ ]$ is of type $[\ l : T\ ]$ just in case v is of type T.

$$R_1 : \begin{bmatrix} l_1 & : & T_1 \\ l_{2=a} & : & T_2 \\ l_{3=p(l_2)} & : & T_3 \end{bmatrix} \quad R_2 : \begin{bmatrix} l_1 & : & T_1 \\ l_2 & : & T_{2'} \end{bmatrix} \quad R_3 : []$$

Figure 2: Example TTR record types

Fields can be *manifest*, i.e. defined in terms of a singleton type e.g. $[\ l : T_a\ ]$ where T_a is the type of which only a is a member; here, we write this as $[\ l_{=a} : T\ ]$. Fields can also be *dependent* on fields preceding them (i.e. higher) in the record type (see Fig. 2).

The standard subtype relation $\sqsubseteq$ can be defined for record types: $R_1 \sqsubseteq R_2$ if for all fields $[\ l : T_2\ ]$ in R_2, R_1 contains $[\ l : T_1\ ]$ where $T_1 \sqsubseteq T_2$. In Fig. 2, $R_1 \sqsubseteq R_2$ if $T_2 \sqsubseteq T_{2'}$, and both R_1 and R_2 are subtypes of R_3. This subtyping relation allows semantic information to be incrementally specified, i.e. record types can be indefinitely extended with more constraints: this inherent property has been the central reason for turning towards TTR for a formalism in which unfolding conceptual structures are represented in incremental parsing and generation.

Record Types as interaction potentials In this chapter, we follow Gregoromichelaki et al. (2020b), and argue that under the actionist perspective on successful communication as coordinative action – see earlier Sec. 3 – linguistically-relevant RTs should not be identified with Austinian propositions (i.e., a situation being of a particular type), as in Cooper (2005); Cooper and Ginzburg (2015) following tradition in NL Semantics. Instead, with the aid of Dynamic Syntax, we propose their reformulation as dynamically constructed ad hoc conceptualisations of situations inducing further actions (predictions) capturing, and enabling, the formation of fields of affordances (see Sec. 6 for an operationalisation of this idea). On this view, RTs are not taken to classify perceptual input or sensory information (cf. Larsson, 2011; Dobnik et al., 2012; Yu et al., 2016), but instead trigger action policies inducing predictions for further interaction. What they classify are therefore (inter)action potentials, thus allowing agents to predict and causally associate what they do or say with the outcomes that these actions are likely to have in their environment, with this crucially including how an interlocutor may or may not respond. As we will see below in Sec. 6, the DS-TTR hybrid also allows agents to use exploration through trial-and-error to *learn* the probabilistic associations between what they say and what is likely to happen afterwards (*reinforcement learning*), leading to particular action/word sequences becoming routinised as ways of bringing about particular perlocutionary effects.

5.2 Parsing and generation of linguistic actions

In DS-TTR, parsing or generating a string of words or non-verbal tokens, induces some organisation of a state space of activity possibilities (a 'field of affordances') in combination with top-down actions ensuing from preexisting skills and dispositions of the participants involved (the 'grammar') (cf. Zadrozny, 2020). This either transforms the existing state space, adds new structures to it, or removes existing paths through it. Locally, the immediate path trajectory moves through a tree-shaped

state space with nodes as states traversed by means of constraints expressed by the modal operators (e.g. $\langle\downarrow\rangle$, $\langle\uparrow\rangle$, $\langle\uparrow_*\rangle$) of a modal tree logic (the Logic of Finite Trees; LOFT: Blackburn and Meyer-Viol, 1994) expressing topological relations among current or future anticipated nodes. The tree-shaped organisation reflects the conceptualisation structure induced by the unfolding utterance in terms of function-argument articulations. More globally, the state space is presented as a directed acyclic graph (ICS-DAG, Interaction Control States DAG) that records possible paths of actions in a landscape defined by what the grammar, acting as a controller of the normativity pertaining to linguistic actions, allows as predictions of future interaction possibilities.

DS-TTR trees are always binary-branching when they have reached a stable organisation because they underpin the dynamic and incremental computation implemented as a combination of functors with their argument. However, as information becomes gradually available, intermediate stages will involve 'structural uncertainty', where so-called 'unfixed nodes' will be constructed triggering predictions and search for their appropriate accommodation in the binary tree organisation. Tree nodes thus correspond to terms in the lambda calculus and function application actions are conventionally indicated with argument nodes appearing on the right and functor nodes to the left. Node states can also host constraints on what input/output is predicted to occur indicated as labels. For example, an argument node might be one of the lowest types in the hierarchy of types ($Ty(e)$), standing for a generic type of entity) and then its sister might be of type ($Ty(e \rightarrow t)$) that receives that type as input and returns a type t ($Ty(t)$) as output (the label component t should not be taken as indicating a truth-evaluable proposition as in formal semantics frameworks since, for example, questions, imperatives, non-finites etc. can all be characterised as $Ty(t)$). Functor-argument structures can be built recursively on that basis (e.g. $Ty(e \rightarrow (e \rightarrow t))$). These type labels thus indicate what is possible for the node type to be expanded into or act as constraints as to what further actions can be taken when the pointer appears on such a node. For example, lexical entries are introduced with IF-THEN conditional actions that make reference to expectations regarding these type labels. So these types are indicators of the structural function-argument organisation of the complex conceptual structure that is being built. On the other hand, record types of Type Theory with Records (Fig. 4) expand such types to a more fine-grained articulation of content and appear also on tree nodes. But, since such types can also be defined as graph structures, we assume here that they can be introduced and built incrementally by means of the PDL apparatus of DS, so that what appears on the node is an abbreviation of an embedded subgraph of the higher level tree graph (see also Kempson et al., 2001, ch.9). Both trees and labels can be *partial* in every respect, introduced initially by prediction, in the form of unsatisfied (indicated with a ?) so-called *requirements* for any element defined within the formalism (e.g. $?Ty(e)$, is a requirement for future development to $Ty(e)$). The satisfaction of predictions can only be launched by proceeding from a specific point on the tree and this is indicated by a *pointer*, $\diamond$, labelling the node currently under development and relative to which any input/output can be defined. The potentially variable position of the pointer relative to a local tree accounts for variable licensed word-order possibilities in each language. The purpose of the grammar is both to induce and satisfy these predictions by licensing (consuming or producing) linguistic actions thus providing a normative perspective in the parsing/generation process. Thus knowledge of the grammar is literally knowledge of "how to go on" (Wittgenstein, 1953, para. 154) in an interaction. Since the possibilities that the grammar makes available are probabilistically ranked, the notion of grammaticality is very local (i.e., non-sentential, word-by-word transition probabilities) and gradient (see e.g. Lappin, 2021) and potentially defined for any combination of lexical actions, with higher 'surprisal' values assigned to unusual, i.e., not yet routinised, combinations (cf. Lau et al., 2017, 2020).

The dynamics of what constitutes 'syntax' in DS-TTR, is defined in terms of conditional packages of *actions*: procedural specifications for updates of the state space. Action sequences can be retrievable as uninterruptable chunks (*macros*). So-called *computational macros* are invoked without any linguistic input triggering their execution and only the pointer's presence at a node satisfying the conditional constraints included in the macro is necessary; and *lexical macros* are language-specific

action policies corresponding to and triggered by specific lexical tokens. All action macros are presented in an IF…THEN…ELSE format and correspond to transition edges (formalised as action PDL operators) along states of the tree or the ICS DAG (see e.g. Figs. 7 and 8). Formally, macros are composed sequences of PDL atomic actions (formalised as (multi)modal operators) such as `make`, `put` and `go`, which reflect state space updating operations. For example, `make` creates a new node, `go` moves the pointer there, and `put` decorates the pointed node with a prediction regarding some node label.

Computational macros form a small, fixed set. Some enforce the overarching constraints imposed by the lambda calculus and the modal logic tree formalism (LOFT: Blackburn and Meyer-Viol, 1994): for example, Elimination, performs beta-reduction of a node's daughters, and annotates the mother node with the result, while Thinning removes satisfied requirements. Other computational actions enable the fundamental predictivity and dynamics of DS-TTR, e.g. Completion, which moves the pointer up and out of a sub-tree once all requirements therein are satisfied; and Anticipation which moves the pointer from a mother node to a daughter node with any unfulfilled requirements thus expecting the resolution of a prediction in the immediate next step. While the former set of actions are *inferential*, thus not adding any new information to the trees, the latter set introduce alternative parse paths, thus capturing structural ambiguity: Completion for example, precludes any further development of the current sub-tree because it moves the pointer up and out of it. The successful parse or generation of a word w_1 thus amounts to finding a sequence of computational actions (possibly empty) leading to a tree that satisfies the preconditions of the lexical action for w_1. The search process through predicted future actions and the history of both taken and abandoned action possibilities is recorded in the ICS DAG, with (partial) trees as nodes, and actions as edges.

Lexical actions are associated with word forms in a DS-TTR lexicon. Like computational actions, these are state space update macros composed of sequences of atomic actions. Fig. 3 shows an example for a proper noun, *John*. The action checks whether the pointed node (marked as ◇) has a prediction (in DS-TTR terms, *requirement*) for the occurrence of type e; if so, it satisfies this prediction with providing type e, introduces the conceptual potential associated with *John* (see section 5.1 for details) and the bottom restriction $\langle\downarrow\rangle\bot$ (meaning that the node cannot have any daughters). Otherwise (if there is no prediction of $?Ty(e)$), the action aborts, meaning that the word *John* cannot be parsed in the context of the current tree.

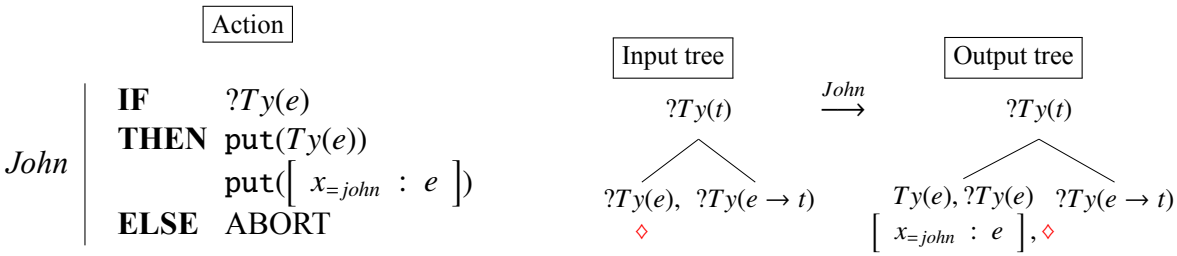

Figure 3: Lexical action for the word 'John'

Fig. 4 shows "John arrives", parsed incrementally, starting with an empty tree, with only the root node's daughters predictively introduced without any lexical grounding, and ending with a complete tree. The intermediate steps show the effects of Completion, which moves the pointer up and out of a complete node, and of Anticipation, which moves the pointer down from the root to its functor daughter.

The DS-TTR framework integrates various forms of uncertainty as an explanatory factor for syntactic/semantic phenomena. As an illustration of syntactic uncertainty, we display in Fig 5 the (condensed) steps involved in beginning the parsing of a standard long-distance dependency, *Who hugged Mary?*. The basic idea implemented in the DS-TTR modelling of such dependencies is that the sentence-initial phrase is introduced with recorded constrained uncertainty as to which role it will eventually play downstream in the tree-construction process. While the process continues, the so-called 'unfixed node' hosting the underderspecified *wh*-content (formalised as a metavariable in need

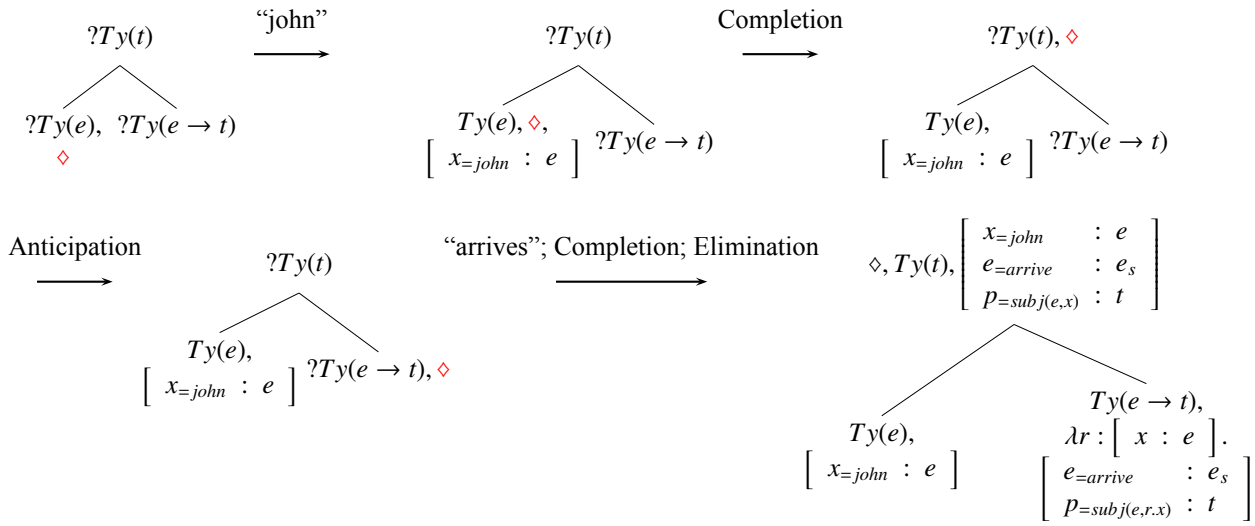

Figure 4: Incremental parsing in DS-TTR: *"John arrives"*

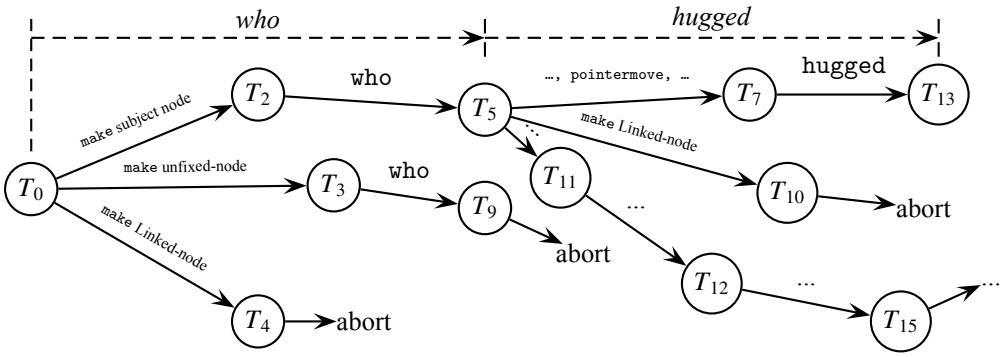

Figure 5: Processing *Who hugged*

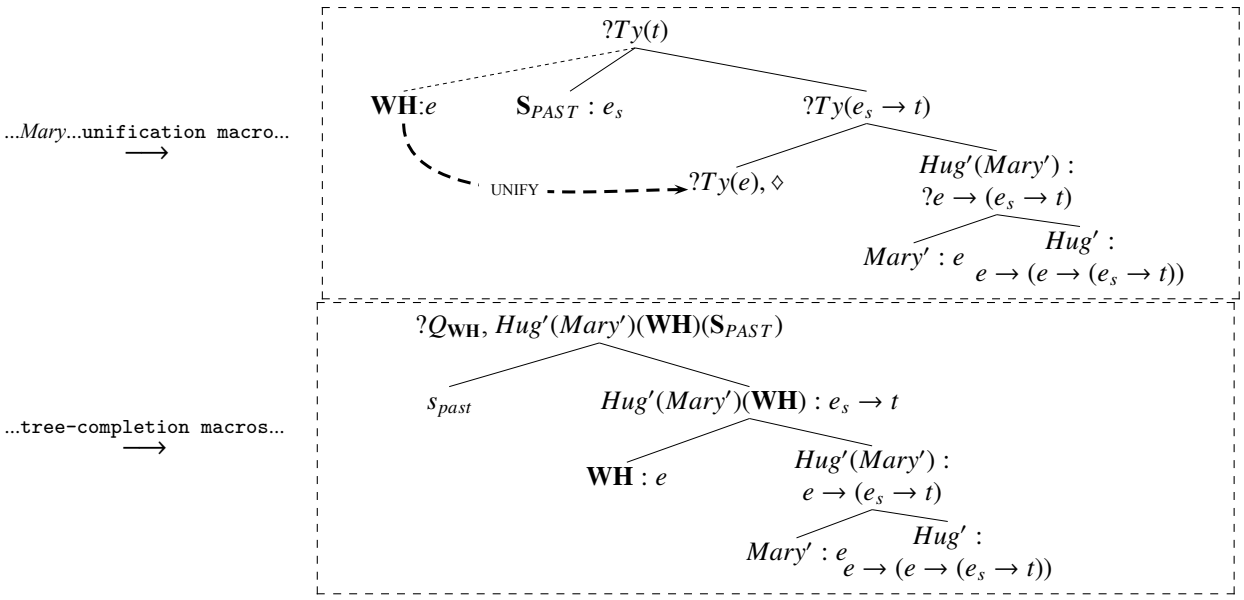

Figure 6: Structural uncertainty in DS-TTR: last steps of processing *Who hugged Mary?*

of substitution) needs to be kept in memory awaiting its resolution. The resolution of the structural uncertainty regarding the position of the unfixed node as subject will be provided in the next steps (see Fig. 6 for one path in the DAG traversal) but the content resolution of the metavariable will normally needs to be provided by the interlocutor. This structural and content uncertainty is an example of how to resolve a number of other syntactic/semantic puzzles regarding the processing of pronouns, anaphors, *wh*-elements, clitics, and various so-called "movement" operations[1]. Here the task starts with a set of probabilistically-weighted predicted *Interaction Control States* (ICSs) represented in the ICS DAG. At this stage, which might be, let's assume, the first utterance in a dialogue, the DAG landscape displays all the potential opportunities for parsing or producing verbal actions, prompting lexical actions as licensed by the grammar of English. These potential actions are assumed to be "virtually present" for the participants even though they are not all eventually actualised.[2] Either participant might then take the initiative to begin the articulation of an utterance while the other is in a state of preparedness checking whether the path pursued by the other interlocutor conforms to their expectations or whether they need to take over and compensate for their lack of coordination (Eshghi et al., 2015). Many alternative processing paths unfold at each step as affordances of the sociomaterial environment are taken up or are gradually abandoned (see also Sato, 2011; Eshghi et al., 2013b; Hough, 2015).[3]

5.3 Conceptualisation as state transitions

The conceptual structure being built here is indefinitely extendible (see Cooper, 2012) and "non-reconstructive" in the sense that it is not meant as a passive inner model of the world (see also Clark, 2017a,b) but as a means of interaction, that is contact, with the world via the predictions generated regarding subsequent processing. Accordingly, the affordances that constitute the conceptual structure are viewed as relational (see also Chemero, 2009; Bruineberg et al., 2018a): a pairing of (aspects of) the world with a (joint) perspective, namely, those affordances of the sociomaterial world that are accessible relative to the agent(s)' relevant sensorimotor skills shaped by prior experiences and the ecological niche.[4] Here we assume a perspectival construal of types as accessible affordances to an agent or group of agents. Following standard assumptions in ecological psychology and phenomenology, it is part of the force of an affordance that the perceiving/acting agent becomes aware that they are manipulating the world from a particular point of view. This awareness is enabled as part of the agent's sensorimotor knowledge of regularities and lawful variations regarding the changes in the environment that are caused by the agent's own actions as opposed to actions/events affecting the agent. When multiple agents are coupled as a temporarily assembled agentive system, but also in cases where experts use tools or patients use prostheses, the collective perception/action possibilities that emerge for the newly-formed systemic unit are not the result of simple summation of what is possible for the individual components but a new perspective for each individual which incorporates their function as a component of the overarching system (Di Paolo and De Jaegher (2012)). Thus the joint landscape

[1] The detailed justification of DS-TTR as a grammar formalism is given elsewhere (Kempson et al., 2001, 2011, 2016, 2017; Eshghi et al., 2011, a.o.).

[2] For relevant notions of "virtual presence", see Noë (2012); DeLanda (2013)

[3] A more realistic graph would also include the possibilities of non-verbal actions, not only gestures, but also physical voluntary actions like, for example, the physical response to a command or request. It is our claim that any "speech act" can be performed non-verbally (see, e.g., Clark, 2012; Gregoromichelaki and Kempson, 2015 and (4)-(5)). Accordingly, physical and grammatical NL actions readily compose with each other exactly because they perform meshing contributions in human interaction (Gregoromichelaki, 2018):

(4) She played [playing tune on the piano] not [playing another tune on the piano]

(5) OK, let's do it together. So we have [arm movement demonstration] and then we go [leg movement demonstration]

[4] In this actionist and externalist perspective, we diverge from standard construals of TTR as in Ginzburg (2012), Cooper, forthcoming.

of affordances can be much more or much less affordances depending on "enabling" or "disabling" couplings. In both cases, agents are able to perceive this new regime and generally capable to adjust their contributions in complementary ways (Mills and Gregoromichelaki, 2010; Mills, 2014).

The relativisation of the structure of human conceptual types against practice-based abilities has normative implications, in that the agent(s) might fail to achieve what is genuinely afforded to them by the sociomaterial environment, or the agent(s) might fail to take up the multitude of affordances that have been perceived as potential ("virtual") paths of action. Moreover, given that they engage with real properties of the sociomaterial context (see also Pickering and Garrod, 2021), the consequences of misapplying their abilities will be detectable by the agents themselves as 'error signals' when their predictions are falsified. Such failure is inevitable and constant and it is, in fact, the source that leads to further finer-grain differentiations in the agents' sociomaterial environment so that local adjustment and long-term learning and adaptation are the outcomes (Bickhard (2009); cf. Friston (2010, 2011)).

Given this requisite dynamicity and world grounding, type (concept) labels, like Hug' or $Arrive'$ here stand for abbreviations of triggers for complex sets of action potentials embedded under the DAG ICS nodes as nested affordances. Such labels then constitute additional ICS choice points in the generation of further potential paths within the DAG. Given this view of concepts, what individuates each such label is its distinguished provision of sets of available actions realisable in the next steps within the field of affordances (the DAG). To take a "syntactic" type as an example, type t is differentiated from type $(e_s \rightarrow t)$ in that the former (minimally) leads to the prediction of a left daughter of type e_s and a right daughter of type $(e_s \rightarrow t)$ whereas the latter leads to the prediction of e and $(e \rightarrow (e_s \rightarrow t))$. This is what differentiates these types, not their distinct labels. Within the grammar, such types either contribute tests in the conditional procedures that implement the operation of grammatical and extra-linguistic actions or trigger searches for appropriate words, or expand the current structure and annotations with the anticipation of further developments. But, even more pertinently here, such types do not have any model-theoretic content beyond the transitions they allow or curtail in the traversal of the states of the PDL model that underpins DS-TTR. Similarly, we take concept labels such as Hug' or $john'$[5] as triggering access to nested structures of potential actions regarding aspects of (mental or physical) interaction with an event of hugging or interacting with John, some of which will be taken up and others abandoned. As such, the types induced by the grammar (aka 'concepts') are mainly constituted by subpersonal mechanisms, however, the results of their operation can be brought to consciousness by processes of reification for purposes of, e.g., linguistic negotiation, explicit planning, theory construction, or teaching.

The actionism foundation of DS-TTR suggests that the sensorimotor knowledge-as-action underpinning to cognition implicates conceptual understanding from the earliest stages of perceptual access (unlike, e.g., existential phenomenology – Dreyfus, 2013 – and related views). However, conceptual abilities do not, as in standard models, proceed via an intermediate cognitive stage before initiating the control of action, for cognition is not seen as separate from the sensorimotor grounding of agent performance. Under this view, concepts are not the rich internal representational structures of standard views – they are skills. For our purposes, we argue that in perceiving some entity and identifying it as a dog, it is not a static retinal image that becomes associated with the application of the 'Dog' type. Instead, memorised patterns of current and past interactions are invoked to construct ad hoc a pattern of predicted interactions that differentiates the particular entity in the current context through its particular set of affordances as, e.g., a threat or a rewarding experience with incrementally adjustable behaviour of approach or avoidance (Gregoromichelaki et al., 2019; Bickhard and Richie, 1983). On this view, conceptual understanding cannot be taken as static pattern-matching but is, instead, an achievement: it is time-extended, incremental, and based on trial-and-error rather than an automatic mapping of experience to internal categories or propositional knowledge.

Moreover, due to their basis in action, concepts are necessarily always fragile and incomplete,

[5] For the view that such entity concepts are tracking abilities allowing the accumulation of knowledge about individuals, see Millikan (2000).

amenable to modification by prediction-error minimisation: in general, the specification of action guidance must allow flexibility to fit different situations and changing conditions and, therefore, successful situated action execution depends on leaving some degrees of freedom unbound (Suchman, 1987). This is notably echoed in NL phenomena like the so-called "polysemy" or "coercion" phenomena where word meanings are notoriously shiftable even within a single context. It is also indicated by the now established assumption of gradient grammaticality or well-formedness (e.g. Lappin, 2021). The latter, in our terms, is reduced to the effects of social normativity on complex, ad hoc conceptualisation sense-making activities, as DS-TTR rejects the notion of a distinct level of syntactic representations. Such phenomena clearly reflect the fact that NL use involves learning and adjustment as an interactional system self-organises around the needs and goals of the interlocutors. As Mills (2011, 2014) and Mills and Gregoromichelaki (2010) argue, interlocutors encountering a novel situation, interactively and incrementally organise their joint and complementary predictions to establish via trial-and-error ad hoc routines for coordinating with each other. This is shown in experiments (e.g. Healey and Mills, 2006) where dyads of participants playing "the maze game", gradually develop group-specific procedural interdependence employing NL structures with highly ad hoc sequential-position-dependent meanings.

Given affordance competition, agents select their next actions based on possibilities (probabilistically) grounded on these types which function as 'outcome indicators' (Bickhard and Richie, 1983) so that the predictions yielded by these types might be reinforced (verified) or abandoned (fail) in the next steps. As long as they remain as live possibilities, the operations induced by the types will keep triggering flows of predictions for further (mental or physical) action even if particular paths of sequences of nested predictions are not taken up. Maintaining even abandoned options is required for the explicit modelling of conversational phenomena like clarification, self/other-corrections (6), etc. but also, quotation, code-switching, humorous effects and puns (Hough, 2015; Gregoromichelaki, 2018) (see, e.g., (7)).

(6) John went swimming with Mary, um. . . , or rather, surfing, yesterday.
['John went surfing with Mary yesterday']

(7) The restaurant said it served meals any time so I ordered breakfast during the Renaissance.
[Stephen Wright joke]

5.4 Overt feedback as pruning of action sequences

However, these live but unactualised predictions, in the case of dialogue, reach the limits of their (virtual) existence when it is no longer possible for either participant to backtrack successfully in order to extend or "repair" ICS node elements due to the fact that the relevant paths have decayed in the DAG history. Memory mechanisms are implicated in how far the currently active DAG records go. This decay and elimination can also be facilitated and induced by the explicit verbal efforts (aka "overt feedback") on the part of the interlocutors, which can be seen as an efficiency strategy to intervene to reduce DAG complexity and lessen the burden on memory requirements.

In DS-TTR, *context*, required for processing various forms of context-dependency – including pronouns, VP-ellipsis, self-repair and short answers – is considered the global state space ICS DAG, which encompasses the virtual field of affordances for the conversational participants (or a dynamically developing 'situation convention', Bickhard (2009)). Edges here correspond to DS-TTR actions – both computational and lexical macros – and nodes correspond to tree subspaces updated after the application of each action (Sato, 2011; Eshghi et al., 2012; Kempson et al., 2015) – see Fig. 1. Here, we take a coarser-grained view of the DAG with edges corresponding to words (sequences of computational actions followed by a single lexical action) rather than single actions, and dropping abandoned parse paths (see Hough, 2015, for details) – Fig. 7 shows an example.

As Eshghi et al. (2015); Howes and Eshghi (2017, 2021) show, the processing and integration of utterances that have been characterised as explicit feedback in dialogue can be captured using the ICS

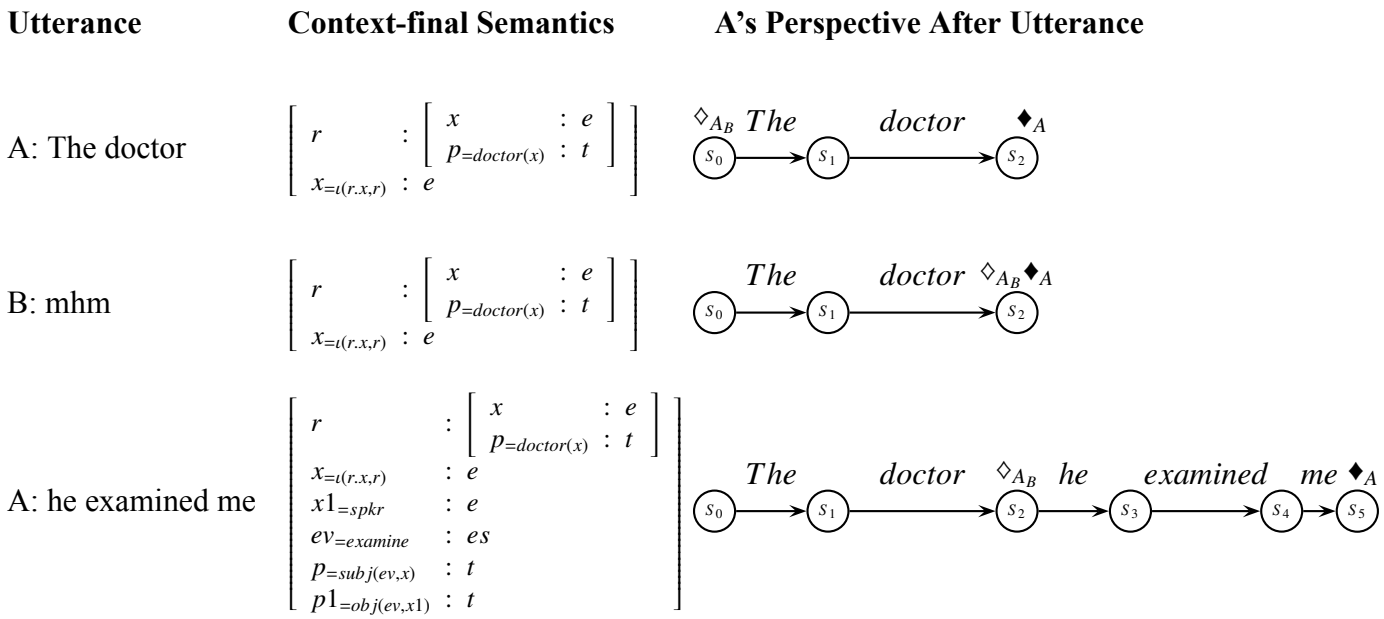

Figure 7: Backchannels as movement of coordination pointers on Interaction Control States (ICS); from A's perspective.

DAG, enhanced with the perspectival conception of affordances we discussed earlier (Sec. 5.3), in this case, implemented as two *coordination pointers*: the *self-pointer*, ♦; and the *other-pointer*, ◇. These pointers indicate the points up to which the dialogue participants have each marked the material as "grounded", i.e., liable to decay from memory storage.

Any action causes ICS pointer movement, and, as we said earlier, any action possibility includes the interlocutors' own perspective of the effect on the ICS. Such perspectives, which are crucial for demarcating self- and other-action may, as we will see below, provide divergent ICS trajectories for each participant with convergence as a result of clarification interaction and repair processes more generally. The self-pointer, $\blacklozenge_A$, on participant A's ICS view tracks the point to which A has given evidence for reaching. The other-pointer, $\lozenge_{A_B}$, tracks where the other participant, B, has given evidence for reaching. For example, an utterance produced by A will move A's self-pointer to the rightmost node of the ICS; on B's ICS perspective, it is the other-pointer that moves to the same location. On this model, the intersection of the path back to the ICS root from the self- and other-pointers is taken to be grounded, with the effect that parse or production search within this grounded pathway is precluded, thus removing the computational cost associated with finding alternative interpretation pathways, as well as formally explaining how conversations move forward.

This model has been shown to account for backchannels (Howes and Eshghi, 2017, 2021), clarificational exchanges and other corrections Eshghi et al. (2015). Clarification Requests (CR) cause branching on the ICS, where the current path is abandoned and another branch constructed – a subsequent response to the CR plus the acknowledgement of this response eventually realigns the two coordination pointers, and the interlocutors' individual ICS perspectives as a consequence.

Backchannels Fig. 7 is a step-by-step illustration of how the ICS with only A's perspective develops as the dialogue proceeds, and as B's backchannel, 'mhm' is processed. After producing the first utterance, A's self-pointer, $\blacklozenge_A$, is on s_2, the right-most node of her ICS so far. B's backchannel "passes the opportunity to repair" (Schegloff, 1982), thus moving A's other-pointer, $\lozenge_{A_B}$, to the same node and so grounding "the doctor". A's subsequent continuation creates new edges, and moves her self-pointer to the new right-most node. At this point, A's new utterance needs further feedback from B

to be grounded: divergence of pointer positions thus represents 'forward momentum' in conversation (elsewhere called *discursive potential*; Ginzburg, 2012).

Overt repair Fig. 8 shows an example of a clarificational exchange, as in 8:

(8) (a) A: The doctor examined me…
(b) B: Chorlton?
(c) A: no, Fitzgerald

It shows the incremental updates arising in the clarifier's perspective (B) in example (8), a case of a non-local CR which requires backtracking.[6] Initially, B successfully parses A's utterance, thus moving the other-pointer ($\Diamond_{B_A}$) to the right-most node of his DAG. Not having secured a referent for "the doctor" with enough certainty, he then aims to produce the CR, *Chorlton?*, which involves backtracking to "the doctor" node in order to produce it. At this juncture A's and B's perspectives have diverged: A's self-pointer ($\blacklozenge_A$) appears at the rightmost DAG edge, which B knows (hence $\Diamond_{B_A}$), while B has not grounded that edge. B's production of the CR causes A to have to parse it. This serves to re-align pointer positions for A and B, the result of which is both of them focussing on "the doctor"-subtree as the source of the misalignment.

A can now offer a confirmatory or a negative response to the CR. (c), in Figure 8 illustrates the latter case, with the utterance of *no* reflecting the abandonment of the "Chorlton?" branch, rather than the denial or rejection of a propositional content. This is followed by a correction of B's CR, thus forcing B's other-pointer ($\Diamond_{B_A}$) out of the "Chorlton?" branch, and inducing the construction of the new, "Fitzgerald" branch. At this juncture, B's self- ($\blacklozenge_B$) and other-pointers ($\Diamond_{B_A}$) are on different branches. This can be taken as representing the requirement for further action to be taken in order to realign pointer positions. Especially for B, whose pointer is now on an abandoned branch, this can constitute an obligation to ground the new information provided by A's repair *Fitzgerald*, thus accounting for the forward momentum created by the negative response. B's final backchannel, in 7(d), then serves to realign his two pointers, signalling acceptance to A, who, having processed the backchannel moves her other-pointer ($\Diamond_{A_B}$) to the same node, $s10$, thus ending the clarification sequence with the achievement of a realignment of A's and B's perspectives.

An alternative parsing path is illustrated in (e) of Figure 8. It represents the case where A, after the clarification in 7(b), confirms that the doctor is in fact Chorlton. This simply involves, for B, moving his other-pointer ($\Diamond_{B_A}$) to the end of the "Chorlton?" branch, thus confirming the referent of *the doctor* as Chorlton. This, unlike the negative response in 7(c) which necessitated rejecting already established branches and pointer divergence, ends the clarification sequence.

Both alternatives end up with A's and B's perspectives aligned as the result of repair and backchannelling and set for the continuation of the dialogue.

The account above puts structural, surface forms of context-dependency at the centre of the explanation of participant coordination and feedback in dialogue: various forms of context-dependent expression, from the weakest – backchannels, which have little or no semantic content, to the strongest – utterance continuations, all serve to narrow down the otherwise mushrooming space of interpretation pathways. Their pervasiveness is therefore not coincidental, but strategic, and serves to make dialogue computationally tractable.

This account gives formal rigour to the view expressed above under which language provides a set of interactional mechanisms – such as ellipsis, repair and backchannels – for dealing with the persistent potential for miscommunication (Healey et al., 2018b; Kempson et al., 2016).

[6] For space reasons we do not here include the clarification recipient's (A) point of view.

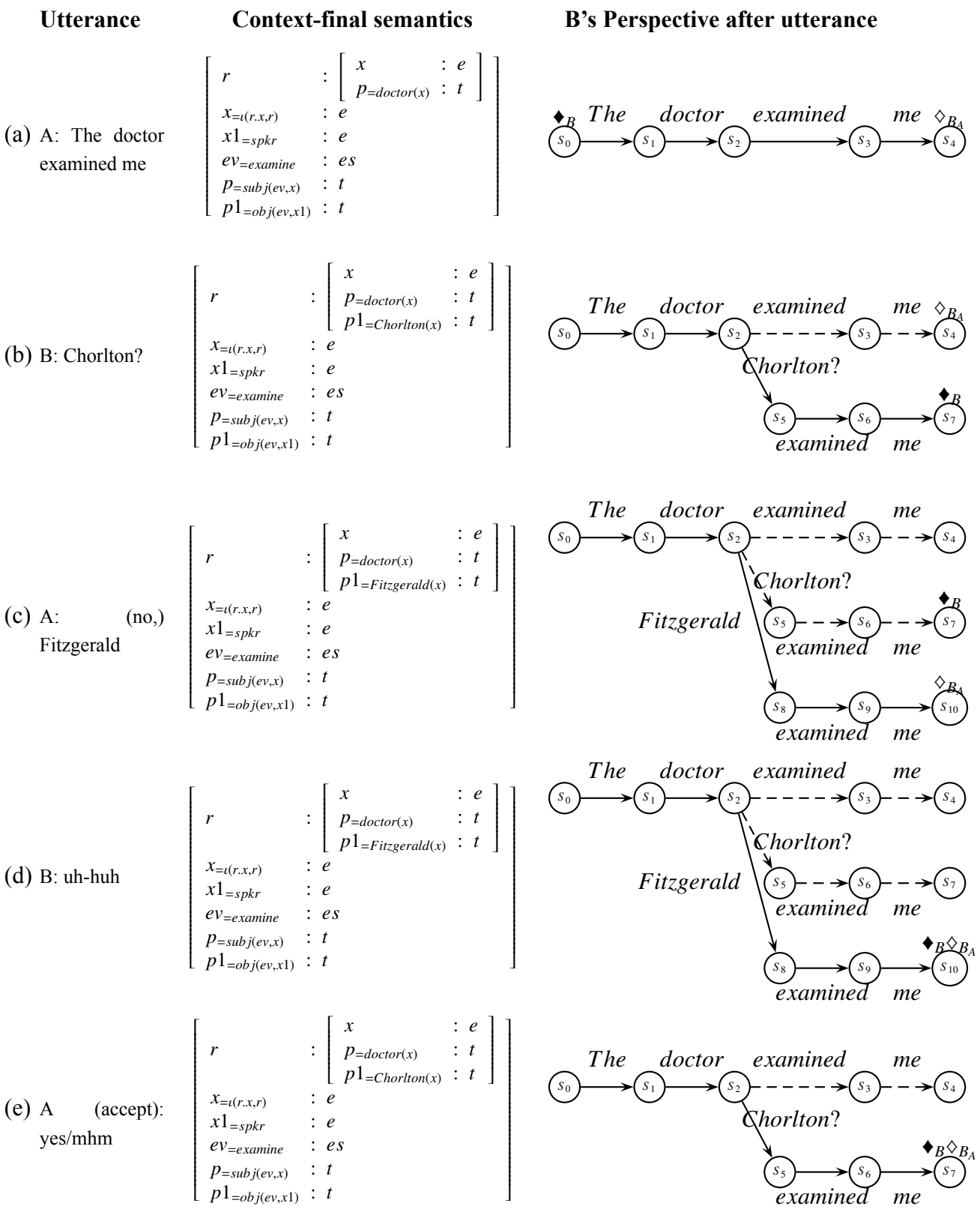

Figure 8: Overt Repair as editing action sequences - from B's perspective; turn (e) comes after (b)

6 Learning through affordance exploration

We said earlier (Sec. 1) that one of the main problems with rule-based models of dialogue, and NL in general, was that a rigid set of hand-crafted rules is applied to the processing and production of behaviour in interactions with the users. As a result, such systems lack flexibility to adjust their responses to various particular tasks by modifying their action policies under the receipt of feedback by the user or the environment. This brittleness is due to the fact that domain-specific and separate knowledge structures are assumed in their architectures lacking the ability to dynamically adapt to open-domain tasks. Low-level, e.g. syntactic and semantic NLU and NLG components, cannot interact and be influenced by the task at hand or the discourse context. Rigid categories of, e.g., intent detection or slot-filling, along with separated modules of dialogue state tracking and policy learning only allow for very task-specific behaviours, often erroneous when faced with ambiguity, noisy input, and less frequent user needs. DS-TTR characteristically does not distinguish between different modular capacities for embodied non-linguistic and linguistic action (Gregoromichelaki, 2013, 2018); neither does it postulate separate knowledge bases for theoretically demarcated linguistic areas like syntax, semantics, pragmatics (Gregoromichelaki et al., 2013). This is under the assumption that all the phenomena identified as separate and indicative of autonomy in these domains have been shown to be influenced by interactions across all presumed levels. It is shown that any update that is performed to adjust processing in the current context might need to take into account all aspects of that context, for example, even the notion of well-formedness has to be defined as context-sensitive and incremental (see e.g. (2)-(3) earlier). Therefore, given the uncertainty and rapid shifting of the sociomaterial environment, for agent behaviours to be adaptable so that they can intervene and avail themselves of opportunities, all such previously considered modularisations are cashed out uniformly in action terms as affordances underpinned by adjustable sensorimotor knowledge on the part of individual agents.

On the other hand, data-driven, end-to-end dialogue systems based on deep learning methods require large amounts of data and often fail to converge and generalise to best overall dialogue policies online, offering generic, uninteresting responses instead. This can be due to learning only simple local associations of input/output with lack of long-term goal-directed processing, ability to act jointly with other agents, and the high-frequency of uninformative responses in the training data. What is not usually implemented is the human tendency for forward-looking policies to exploit and explore the environment, including the conversational environment, for affordances, opportunities for action to receive rewards or avoid dangers. For this human characteristic to be implemented, agents have to be modelled as 'continual learners' (Roller et al., 2020). Exploring via trial-and-error the shifting landscape of affordances is crucial for the adaptability of agent systems (either sole embodied agents or groups of agents, Adolph, 2020; Veissière et al., 2020). Human conceptualisations of situations at hand, in the form of the solicitations perceived as states of action-readiness (see section 3.1 earlier), depends on building skills that arise from the accumulation of multimodal experiences and acquiring skill within practices ('language games') available in the particular 'form of life' inhabited by the agent. Language use according to these assumptions is no different since its function is to guide the perception and creation of such social conceptualisations via the establishment of grammar models, i.e., in our view, sets of actions (*macros*) that have been proved rewarding in previous interactions.

It is for these reasons that words, morphology, and syntax are all modelled as affordances in DS-TTR, i.e., indicators of opportunities for (inter)action and the source of normativity (notions of 'correctness/incorrectness' embodied in a grammar). As we saw earlier, such interactions incrementally open up a range of options for the interlocutors so that selected alternatives can be pursued either successfully or unsuccessfully: even though a processing path might look highly favoured initially, due to the changing conditions downstream, it might lead to failure so that processing is aborted and backtracking to an earlier state is required. The potential for failure or success relative to goals imbues the activities of the system, even though mainly subpersonal and perceived as affective states of 'action readiness', with a notion of normativity arising from the routinisation of action sequences retrievable

as chunks (macros). Such macros impose licensed expectations (predictions) that can in turn operate as triggers resulting in nested structures of affordances constraining potential interactions. This normative field of nested anticipations of further interactions built on the basis of prior trial-and-error efforts comes to constitute an instantiation of the *grammar* in particular concrete occasions. Such ad hoc grammars are what prompts or constrains the actions of the individuals participating in a dialogue. The grammar in this sense can be seen as an embodied generative model (e.g. Kirchhoff and Froese, 2017) allowing interlocutors to perform step-by-step a coordinated mapping from perceivable stimuli (phonological strings) to conceptual and physical actions or vice-versa.

In the following sections, we will consider two case studies of how the above ideas can work in practice. Firstly, we show how DS-TTR action policies can be learned through exploring environmental contingencies and acquiring skills in predicting suitable trajectories within the evolving landscape of affordances via Reinforcement Learning methods. Reinforcement Learning mechanisms can be subsumed under the more general framework of self-organisation via the Free Energy Principle and active inference (see e.g. Friston et al., 2012b; Tschantz et al., 2020) which we take here in its enactive interpretation as concerning states of action readiness. Secondly, we show how the 'education of attention' assumption about learning how to perceive affordances (Gibson, 1966) can be cashed out in DS-TTR terms. We focus on summarising the following areas of current research:

a. the so-called BABBLE method in Eshghi and Lemon (2014); Kalatzis et al. (2016); Eshghi et al. (2017) for bootstrapping interaction. This work combines DS-TTR with Reinforcement Learning, implementing and evaluating a method that allows fully incremental dialogue systems to be learned from small amounts of raw, unannotated dialogue data.

b. the work on grammar learning in Eshghi et al. (2013a,c) who present and evaluate a method for learning an incremental DS-TTR grammars from data in which utterances are paired with conceptualisation structures standing for the sense-making activities available to an agent.

6.1 Bootstrapping interaction: Learning how to do things with words

Eshghi and Lemon (2014), Kalatzis et al. (2016) and Eshghi et al. (2017) combine DS-TTR with Reinforcement Learning, implementing and evaluating a method that allows fully incremental dialogue systems to be learned from small amounts of raw, unannotated dialogue data. This work shows how a dialogue agent can learn to perform dialogue acts (or speech acts) together with their attendant, interactional structures within a particular domain of language use (i.e. a *language game*) without any of this being provided in advance in the form of supervision. The model relies on what the authors call 'babbling': the dynamic and local trial-and-error generation and composition of action sequences (macros) *in a particular context*, and in interaction with a simulated interlocutor, using a DS-TTR grammar. This babbling mechanism amounts to what Gregoromichelaki et al. (2020b) call the 'exploration of the field of affordances' in the agent's environment which includes the interlocutor. This leads to establishing conditional, probabilistic expectations about the outcomes of such low-level action sequences, for example, a question answered, a request fulfilled, some information given, a promise accepted etc. What is learned is thus probabilistic routines (macros) for producing desired *perlocutionary effects* in the agent's environment.

6.1.1 The BABBLE method

In this section we describe the BABBLE method for combining DS-TTR with Reinforcement Learning for learning Dialogue Management (DM) and Natural Language Generation (NLG) policies for a particular dialogue domain, and where these two problems are treated as a joint decision/optimisation problem.

The BABBLE method starts with two resources: a) a DS-TTR parser DS_{TTR} (either learned from data, as in Eshghi et al., 2013a, or constructed by hand), for incremental language processing, but also, more generally, for tracking the context of the dialogue using the DS_{TTR} model of feedback (Eshghi

et al., 2015; Howes and Eshghi, 2017, 2021); b) a set D of transcribed successful dialogues in the target domain.

We perform the following steps overall to induce a fully incremental dialogue system from D:

a. Automatically induce the Markov Decision Process (MDP) state space, S, and the dialogue goal, G_D, from D;
b. Automatically define the state encoding function $F : C \rightarrow S$; where $s \in S$ is a (binary) state vector, designed to extract from the current context of the dialogue, the semantic features observed in the example dialogues D; and $c \in C$ is a DS-TTR context, viz. a pair of TTR Record Types: $\langle c_p, c_g \rangle$, where c_p is the content of the current, *PENDING* clause as it is being constructed, but not necessarily fully grounded yet; and c_g is the content already jointly built and *GROUNDED* by the interlocutors (loosely following the DGB model of Ginzburg, 2012).
c. Define the MDP action set as the DS_{TTR} lexicon L (i.e. actions are words);
d. Define the reward function R as reaching G_D, while minimising dialogue length.

We then solve the generated MDP using Reinforcement Learning, with a standard Q-learning method: train a policy $\pi : S \rightarrow L$, where L is the DS_{TTR} Lexicon, and S the state space induced using F. The system is trained in interaction with a (semantic) simulated user, also automatically built from the dialogue data and described in the next section.

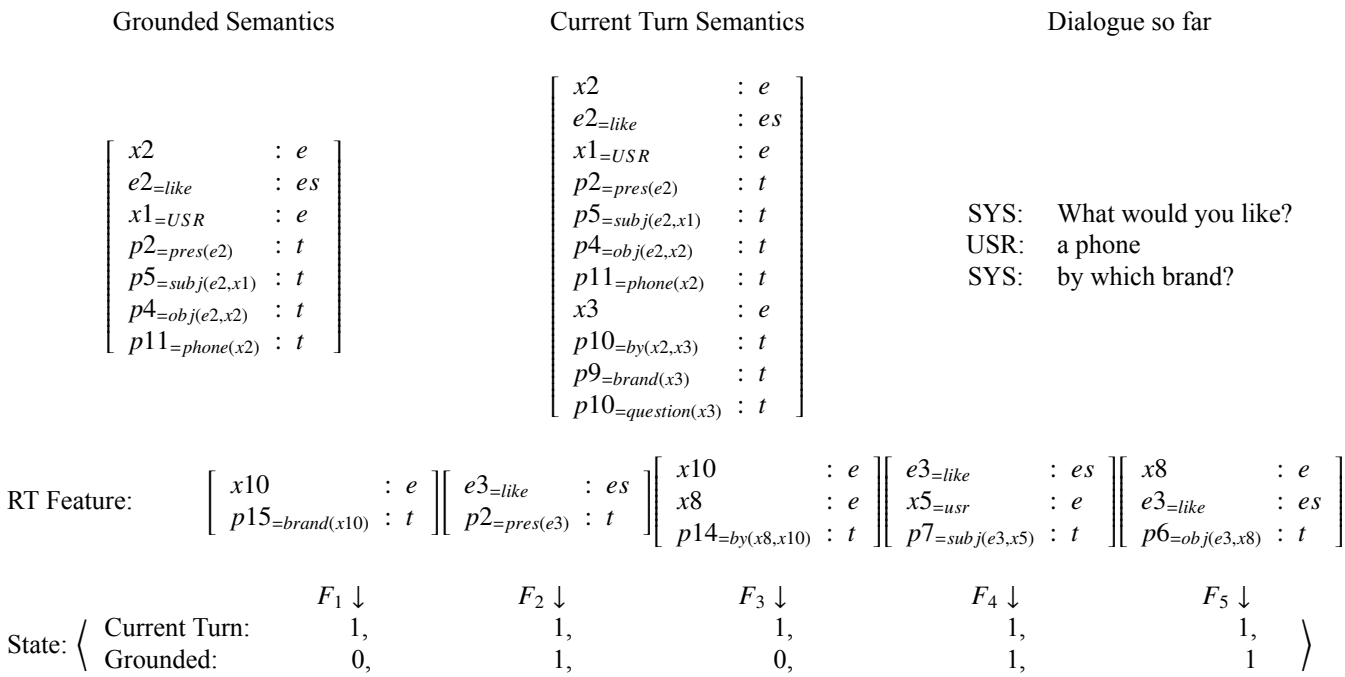

Figure 9: Semantics to MDP state encoding with RT features

The state encoding function, F As shown in Fig. 9 the MDP state is a binary vector of size $2 \times |\Phi|$, i.e. twice the number of the RT features. The first half of the state vector contains the grounded features (i.e. agreed by the participants) ϕ_i, while the second half contains the current semantics being incrementally built in the current dialogue utterance. Formally:
$s = \langle F_1(c_p), \ldots, F_m(c_p), F_1(c_g), \ldots, F_m(c_g) \rangle$;
where $F_i(c) = 1$ if $c \sqsubseteq \phi_i$, and 0 otherwise. (Recall that $\sqsubseteq$ is the RT subtype relation).

6.1.2 Simulating the interlocutor

The user simulation is in charge of two key tasks during training: (1) generating user turns given the domain-specific action triggers or contexts; and (2) word-by-word monitoring of the utterance

so far generated by the system during exploration (i.e. babbling grammatical word sequences) by the system. Both (1) and (2) use the full machinery of the DS_{TTR} parser, as well the state encoding function F, described above. They are thus performed based on the *context* of the dialogue so far, as generated by DS_{TTR}, the result of parsing or generation of word sequences (rather than, e.g. being based on string or template matching).

The rules required for (1) and (2) are extracted *automatically* from the raw dialogue data, D, using DS_{TTR} and F. The dialogues in D are parsed and encoded using F incrementally. For (1), all the user action triggers that trigger the user to generate a turn, $s_i = F(c)$ – where c is a DS-TTR context – immediately prior to any user turn is recorded, and mapped to what the user ends up saying in those contexts - for more than one training dialogue there may be more than one candidate (in the same context/state). The rules thus extracted will be of the form:
$s_{trig} \rightarrow \{u_1, \ldots, u_n\}$, where u_i are user turns.

Now note that the s_i's prior to the user turns also immediately follow system turns. And thus to perform (2), i.e. to monitor the system's behaviour during training, we only need to check further that the current state resulting from processing a word generated by the system, subsumes - is extendible to - one of the s_i. We perform this through a simple bitmask operation (recall that the states are binary). The simulation can thus semantically identify erroneous/out-of-domain actions (words) by the system. It would then terminate the learning episode and penalise the system immediately, aiding speed of training significantly.

6.1.3 Discussion

What is learned Using the method above, what is learned through RL exploration of lexical action pathways – trial and error generation or 'babbling' – is a policy mapping Record Types of TTR (dialogue contexts) to individual lexical actions or words, thus incrementally specifying what the agent should say/do in each of the contexts encountered enough times during training. Taken together with the ICS DAG (see above, Sec. 5.4), these contexts thus encode potentials for interaction to achieve some goal, that is, affordances in the agent's immediate environment.

Generalisation/bias The method described above has enabled prototype incremental dialogue systems to be bootstrapped automatically from small amounts of raw, unannotated dialogue data. For example, Eshghi et al. (2017) show that their resulting model can process 74% of the Facebook AI bAbI dataset even when trained on only 0.13% of the data (only 5 dialogues); and that it can in addition process 65% of bAbI+, a corpus they created by systematically adding self-corrections, restarts and hesitations to the bAbI dataset.

We argue that this generalisation capacity results from: (1) the predictive power of the underlying Dynamic Syntax incremental processing engine which provides constraints on how lexical actions can dynamically compose to form larger structures that simultaneously encode expectations about future possibilities; and (2) the inference power inherent within the Type Theory with Records (TTR) conceptualisation framework whose record types were used as action triggers. This allows equivalence classes to be formed during learning over different interaction potentials whereby alternative action pathways are captured as 'synonymous' viz. as having the same perlocutionary effect. This generalisation power thus results from the combined power the Dynamic Syntax syntactic engine on the one hand, and the TTR inference engine on the other.

Multi-modality As noted, the DS-TTR framework has already been used to integrate perceptual and linguistic semantics (see Yu et al. (2016); Hough et al. (2018, 2020)). Given the non-modularity assumption underpinning DS-TTR, and its attendant continuity between linguistic and non-linguistic actions (see above and Gregoromichelaki (2013, 2018)), it is only natural that learning through affordance exploration, or the BABBLE method outlined above, should also extend seamlessly to cross-modally grounded situations of utterance, or language games (e.g. manipulating objects together,

cooking together, and the like). In what we presented above, the interactive exploration of dialogue trajectories is used to learn the causal associations between what the agent might say, and how the dialogue would continue (e.g. the interlocutor providing a specific piece of information). But this discovered/learned perlocutionary effect need not be purely linguistic, but can equally be an effect observed in the non-linguistic (e.g. visual), physical or simulated environment of the dialogue (e.g. the interlocutor moving an object, going from one place to another or taking other actions in the world).

Future work will therefore explore integration of the BABBLE method with state of the art computer vision models (e.g. LXMERT (Tan and Bansal, 2019), Resnet-based models, or scene graph prediction models (Xu et al., 2017; Chen et al., 2019)), which can be used to infer high level representations of the visual scene. These observations will then constitute the non-linguistic context of the dialogue, and will form components of the reward function for RL. This in turn allows learning of dialogue trajectories that lead to particular task outcomes in the visual environment of the agent.

6.2 Grammar induction

As a second case study of how *affordance exploration* underpins learning, we turn to work on incremental grammar induction where essentially the same mechanism of trial-and-error composition of macros is used to learn DS-TTR grammars from data in a weakly supervised setting.

Eshghi et al. (2013a) describe a method for inducing probabilistic DS-TTR lexicons from sentences paired with DS-TTR trees (see below) representing the function-argument structure of conceptualisation potentials organising sense-making activities discriminated in a fine-grained manner with assignments of typing information. Here we follow the general logic of the enactive version of the Free Energy Principle and active inference in that we assume that perception (sense-making) consists in predicting the sensory outcomes of agents' own actions interacting with environmental affordances and using prediction errors to either amend the predictions accordingly or take further action to improve predictive accuracy. Eshghi et al. (2013b) then go on to extend this work to induce DS-TTR lexicons (a set of word forms associated with action macros) from the CHILDES corpus (MacWhinney, 2000). This work takes real child-directed utterances and pairs them with sense-making activity indicators (conceptualisations) in the form of TTR Record Types (see above, but also Cooper, 2005; Cooper and Ginzburg, 2015), thus providing weaker supervision. By assuming only the availability of a small set of general action sequence composition operations, reflecting the properties of the lambda calculus and the Logic of Finite Trees (LOFT, Blackburn and Meyer-Viol, 1994) that underpins DS-TTR, they ensure that the lexical actions learnt include the grammatical constraints and corresponding compositional structure of the language.

Their method exhibits incrementality in two senses: *incremental learning*, with the grammar being extended and refined as each new data point becomes available; an ensuing inherently *incremental, probabilistic grammar* for parsing and production, suitable for use in incremental dialogue systems (Purver et al., 2011) and for modelling human-human dialogue.

6.2.1 Problem statement

Our induction procedure now assumes as input:

- a known set of DS_{TTR} computational macros.
- a set of training examples of the form $\langle S_i, R_{T_i} \rangle$, where $S_i = \langle w_1 \dots w_n \rangle$ is an utterance in the language and R_{T_i} – henceforth referred to as the *target RT* – is the record type representing a situation conceptualisation, a set of available affordances, induced by S_i.

The output is a grammar specifying the possible lexical macros for each word form in the corpus. Given our data-driven approach, we take a probabilistic view: we take this grammar as associating each word form w with a probability distribution θ_w over lexical actions. For use in parsing, this distribution should specify the posterior probability $p(a|w, T)$ of using a particular action a to parse/generate a word form w in the context of a particular partial tree T.

6.2.2 Hypothesis construction by affordance exploration

The DS procedural framework is *monotonic*: actions can only *extend* the current (partial) tree T_{cur}, deleting nothing except satisfied requirements. Thus, lexical actions can be hypothesised by incrementally exploring the space of all monotonic, well-formed extensions T of T_{cur}, whose maximal conceptualisation affordances collected under R is a supertype of (extendible to / subsumes) the target R_T (i.e. $R_T \sqsubseteq R$). This gives a bounded space described by a DAG equivalent to that of section 5.4 - see Fig. 1: nodes are trees; edges are possible tree-building actions; pathways start from T_{cur} and end at any tree that can instantiate R_T. Edges may be either known computational actions or new *lexical hypotheses*. The space is further constrained by the properties of the lambda-calculus and the state transitions imposed by the constraints expressed in the modal tree logic LoFT (not all possible trees and extensions are well-formed).

General tree-building actions The lexical hypotheses comprising these DAG paths are divided into two general classes: (1) *tree-building* hypotheses, which hypothesise appropriately typed daughters to compose a given node; and (2) *content* hypotheses, which decorate leaf nodes with appropriate supertypes of R_T (non-leaf nodes then receive their content via beta-reduction/extension of daughters).

Tree-building actions can be divided into two general options: functional decomposition (corresponding to the addition of daughter nodes with appropriate types and formulae which will form a suitable mother node by beta-reduction); and type extension. We do not go into any details on the latter here, but note that possible type extensions constitute their own search space modelled using Record Type lattices (see Eshghi et al., 2013b; Hough and Purver, 2014a).

Figure 10 shows example *tree-building* action hypotheses which extend a mother node with a type requirement to have two daughter nodes which would (once themselves developed) combine to satisfy that requirement. On the left, a general rule in which a currently pointed node of some type X can be hypothesised to be formed of types e and $e \rightarrow X$ (e.g. if $X = e \rightarrow t$, the daughters will have types e and $e \rightarrow (e \rightarrow t)$). This reflects only the fact that DS-TTR trees correspond to lambda calculus terms, with e being a possible type. The other is more specific, suitable only for a type e node, allowing it to be composed of nodes of type cn and $cn \rightarrow e$ (where $cn \rightarrow e$ turns out to be the type of determiners), but again reflects only general semantic properties which would apply in any language.

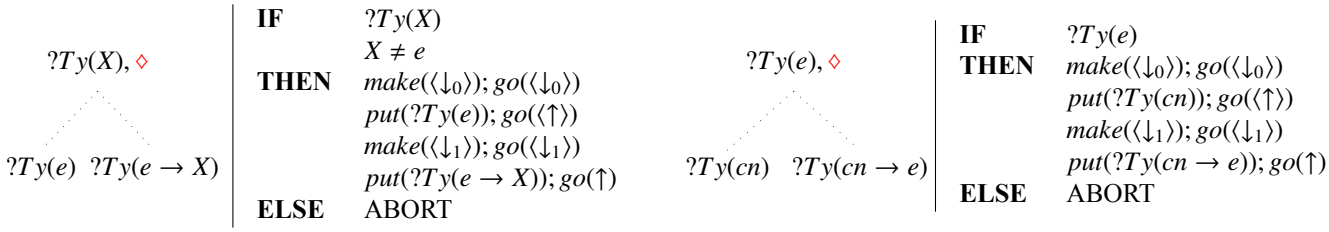

Figure 10: Target-independent tree-building hypotheses

6.2.3 Hypothesis splitting

Hypothesis construction therefore produces, for each training sentence $\langle w_1 \dots w_n \rangle$, all possible sequences of actions that lead from the axiom tree T_0 to the target tree T_t (henceforth, the *complete* sequences); where these sequences contain both lexical hypotheses and general computational macros. To form discrete lexical entries, we must split each such sequence into n sub-sequences, $\langle cs_1 \dots cs_n \rangle$, with each *candidate subsequence* cs_i, corresponding to a word w_i, by hypothesising a set of word boundaries.

Eshghi et al. (2013a,b) go on to describe how this splitting process can work and lead to distinct word hypotheses, w_i, and how the probability distribution $p(a|w, T)$ can be estimated using an incremental version of the Expectation Maximisation (or EM) algorithm.[7]

[7] We do not go into more detail here, but refer the interested reader to the original papers.

6.2.4 Discussion

Using the method outlined above, Eshghi et al. (2013b) show how grammars (normative action policies) can be learned from child-directed dialogue utterances annotated with conceptualisation potential here assumed to be provided by the multimodal environment surrounding the learner. The grammars learned are shown to have wide parsing coverage (92%), as well as good semantic accuracy (F-Score of 0.85).

The process described above of exploring the space of possible tree-building action sequences that extend some tree to another tree whose maximal conceptualisation potential subsumes the target set of affordances R_T is essentially the same as the 'babbling', or *affordance exploration* mechanism discussed above in Sec.6.1.1. While the mechanism is the same, there are at least two key differences here:

(i) In the problem above, of bootstrapping interaction, actions corresponded to words, and the underlying grammar was input to the learning. Here, action hypotheses are abstract: they merely specify very general procedures for extending the tree, albeit constrained by the properties of the lambda calculus and the modal tree logic, LoFT.

(ii) The trial-and-error generation or babbling mechanism of 6.1.1 was constrained by *interaction potentials*, i.e. possible responses in specific cases from the interlocutor. Here, the search space is constrained by R_T, under current assumptions, the conceptualisation trajectories opened up by reaching the goal: local action pathways that do not subsume further R_T trajectories (i.e. are not extendible to it: when $R_T \not\sqsubseteq R$) are abandoned.

As we said earlier (section 5.3), we don't take RTs as monolithic structures with symbols standing for entities in the world. Instead, following the enactive logic of the Free Energy Principle and active inference we construe such types in the same way as DS syntactic types, in the sense that they are individuated by means of the actions that they make available. So, for example, a type like Dog' expands into a set of sensorimotor contingencies (Nöe, 2004; Bickhard, 2009) arising from the agents' experience with dogs. These are expressed as nested structures introduced with associated requirements as standard in DS-TTR and constitute anticipations of potential interactions with the individual entity so characterised (Bickhard and Richie, 1983). In addition, as affordances, they also include the perspectival effects of the agent's action in proposing to conceptualise this entity by use of the particular word form associated with the Dog' type.

6.2.5 Limitations

Computational complexity As Eshghi et al. (2013b) themselves note, the overall time complexity of the algorithm outlined above of hypothesising all action sequences that subsume the goal R_T is exponential in the number of fields in the goal R_T. This is because the algorithm relies on the construction of incremental conceptualisation pathways that explore all trajectories leading from the empty type, R_ϵ, to the goal R_T; thus traversing all the super-types of R_T (all R_S such that $R_T \sqsubseteq R_S$) in all possible orders along the way. Since the number of super-types of R_T is exponential in the number of fields in R_T, this algorithm has an exponential time complexity.

To evaluate their algorithm, Eshghi et al. (2013b) thus had to limit their training data to shorter sentences which also had smaller conceptual structures, with fewer fields in the goal R_T.

Curriculum Learning Nevertheless, it is not reasonable to assume that acquisition of grammars in children starts from long, complex sentences or interactions with complex embedded structures. Instead, children start learning words as part of short and simple dialogue exchanges, within simple language games (e.g. getting something to eat, asking for an object, getting the other to attend to an object, giving and taking, etc.), with the process of language acquisition often viewed as correlated

with the length of child-produced utterances (Ginzburg and Kolliakou, 2009; Brown, 1973). Over time, they then use this knowledge in learning more complex syntactic and conceptual manipulations.

This *curriculum learning* strategy (Bengio et al., 2009) is compatible with the incremental hypothesis construction algorithm above: while Eshghi et al. (2013b) learn from each training example individually, starting from scratch each time, this does not have to be the case: action sequences learned from shorter utterances and simpler conceptual structures can be *reused* later in the context of more complex examples, thus exponentially reducing the search space and the complexity of learning from these more complex examples.

Future work will therefore explore a curriculum learning strategy to overcome the computational complexity of Eshghi et al.'s 2013b algorithm.

Learning from feedback In the formal and computational literature, the complexity of the grammar induction problem has currently been only assessed on the basis of the assumption of an independent syntactic and/or semantic structure that has to be learned on the basis of discovering rules and representations (although cf. Steedman, 2002, for a view of grammar as affordances in the BDI tradition). This formulation of the problem is ill-posed from the DS-TTR perspective, which seeks to formulate a more holistic model of grammar that includes both the generator and parser, their interactions, as well as the sociomaterial environment (see also Pickering and Garrod, 2021). This perspective allows the DS-TTR system to invoke learning regimes that embed learning of verbal actions within domain-general systems of action policy learning like RL. Here the assumption is that the added complexity of dependencies between synergistically organised components will prove to be not a liability but an advantage: this is the intuition pursued in the development of end-to-end architectures and their current successes as well as argued for in the psycholinguistic literature on language learning (e.g. Rączaszek-Leonardi et al., 2013, 2018). Additionally, as Lappin (2021) also notes, general RL methods are now being used by simulated agents to display generalising behaviours and one-shot learning of linguistic actions integrated into policies that respond to multimodal signals (see e.g. Hill et al., 2021)

These insights open new avenues for how to improve current computational architectures and NL theoretical frameworks by including features of human interactivity in the proposed models. It has been a traditional assumption in formal linguistics that human language acquisition cannot involve so-called "negative evidence", i.e., explicit corrections are neither provided to the child learning the language neither can they be processed as corrections given the child's capacities. However, these claims have been disputed and shown to be untenable. Saxton (1997); Saxton et al. (1998, 2005); Chouinard and Clark (2003); Clark and Lappin (2011) among others argue that negative feedback in the form of reformulations and corrections constitutes useful reliable input for inducing a grammar especially when the discourse context, i.e., the surrounding sociomaterial affordances, is simultaneously taken into account, which is what the uniform non-modular formulation of a DS-TTR grammar is trying to capture. Formal models of dialogue as well as grammar learning and induction methods now advocate taking into account feedback from an interlocutor or teaching partner (see, e.g., Ginzburg, 2012; Angluin and Becerra-Bonache, 2017; Liu et al., 2018). In future work with simulated agents, we plan to incorporate through further work with RL the impact of feedback, both corrective and confirming, in the induction of a full DS-TTR grammar from the types of sparse data and one-shot learning observable in human language acquisition.

7 Concluding discussion

In this chapter, we suggested that despite considerable recent advancements in deep learning methods for Natural Language Processing, progress in the area of dialogue modelling and Conversational AI has plateaued. We argued that this is because of the very wide predominance of the code model of communication under which agents are supposed to manipulate and transmit mental representations via NL to then be recovered and duplicated in the mind of the interlocutor. This construal of communication leads to passive agents and models of agents that do not learn interactional feedback mecha-

nisms, instead of fixed representations, with the result that they remain *static* during interaction or at prediction time. We further argued that this view should be replaced by a strongly enactive perspective on NLs, agent communication, and coordination in conversation. Research in robotics, psychology, and neuroscience, but also machine learning, currently converge on the perspective that prediction error minimisation is the mechanism under which agent performance is adjusted and developed dynamically to deal with the uncertainty and contingent nature of outcomes in everyday interactions with the environment and other agents. This perspective presupposes that agents do not carry around within their skulls explicit models of the world but, instead, they possess and gradually refine bodily skills (which includes mental skills) for dealing with the constantly changing circumstances of the surrounding sociomaterial environment. The affordances of the environment are revealed in real-time to interacting agents based on constant 'education of attention' processes that, crucially for humans, include NL-induced sociocultural conceptualisations of the material environment. This means that artificial agents should primarily be provided with embodiment (even if in simulation forms) and opportunities for interaction, along with learning and adjustment as is currently possible with deep learning methods.

We presented two case studies using Dynamic Syntax and Type Theory with Records (DS-TTR), an inherently action-based grammar formalism, showing how exploration of affordances and environmental communication contingencies using the grammar as a generative model enabling prediction induces learning. In one case, we showed how a dialogue agent can *learn to perform dialogue acts (or speech acts)* together with their attendant interactional structures without any of this being provided in advance in the form of supervision. What is learned in this case are conditional, probabilistic routines for producing desired *perlocutionary effects* in the agent's environment. In the second case study, we saw how the same trial-and-error generation mechanism for affordance exploration enables DS-TTR grammars to be learned from child-directed utterances.

Future DS-TTR work will explore lexical learning from live, multi-modal interaction whereby lexical entries for simple, proto-grammars can be induced from real-time feedback.It will also integrate the BABBLE framework with Deep Reinforcement Learning and state of the art computer vision techniques in a visually grounded setup whereby a conversational agent can learn to interact with and produce goal-directed effects in its physical (or simulated) environment.

References

Karen E. Adolph. An Ecological Approach To Learning In (Not And) Development. *Human development*, 63(Suppl 3-4):180–201, January 2020. ISSN 0018-716X. doi: 10.1159/000503823.

Dana Angluin and Leonor Becerra-Bonache. A model of language learning with semantics and meaning-preserving corrections. *Artificial Intelligence*, 242:23–51, January 2017. ISSN 0004-3702. doi: 10.1016/j.artint.2016.10.002.

R.B. Arundale. Against (gricean) intentions at the heart of human interaction. *Intercultural Pragmatics*, 5(2):229–258, 2008.

Robert B. Arundale. *Communicating & Relating: Constituting Face in Everyday Interacting*. Oxford University Press, January 2020. ISBN 978-0-19-093363-0.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-⊂main.9.

J. Barwise and J. Perry. *Situations and Attitudes*. MIT Press, Cambridge, MA, 1983.

Y. Bengio, R. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of ICML*, pages 41–48, 2009.

Mark H Bickhard. The interactivist model. *Synthese*, 166(3):547–591, 2009.

Mark H. Bickhard and D. Michael Richie. *On The Nature Of Representation: A Case Study Of James Gibson's Theory Of Perception*. Praeger, New York, 1983.

Patrick Blackburn and Wilfried Meyer-Viol. Linguistics, logic and finite trees. *Logic Journal of the Interest Group of Pure and Applied Logics*, 2(1):3–29, 1994.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258 [cs]*, August 2021.

Roger Brown. *A first language: The early stages*. Harvard University Press, 1973.

Jelle Bruineberg, Anthony Chemero, and Erik Rietveld. General ecological information supports engagement with affordances for 'higher' cognition. *Synthese*, 196(12):5231–5251, 2018a.

Jelle Bruineberg, Erik Rietveld, Thomas Parr, Leendert van Maanen, and Karl J Friston. Free-energy minimization in joint agent-environment systems: A niche construction perspective. *Journal of Theoretical Biology*, 455:161–178, October 2018b. ISSN 0022-5193. doi: 10.1016/j.jtbi.2018.07.002.

Thomas Buhrmann, Ezequiel Alejandro Di Paolo, and Xabier Barandiaran. A Dynamical Systems Account of Sensorimotor Contingencies. *Frontiers in Psychology*, 4, May 2013. ISSN 1664-1078.

Lou Burnard. *Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services, 2000.

Guendalina Caldarini, Sardar Jaf, and Kenneth McGarry. A Literature Survey of Recent Advances in Chatbots. *Information*, 13(1):41, 2022. doi: 10.3390/info13010041.

Ronnie Cann, Ruth Kempson, and Lutz Marten. *The Dynamics of Language*. Elsevier, Oxford, 2005.

Ana Paula Chaves and Marco Aurelio Gerosa. How Should My Chatbot Interact? a Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction*, 37(8):729–758, May 2021. ISSN 1044-7318. doi: 10.1080/10447318.2020.1841438.

Anthony Chemero. *Radical Embodied Cognitive Science*. MIT Press, Cambridge, MA, 2009.

Vincent Chen, Paroma Varma, Ranjay Krishna, Michael Bernstein, Christopher Re, and Li Fei-Fei. Scene graph prediction with limited labels. In *International Conference on Computer Vision*, 2019.

Michelle M Chouinard and Eve V Clark. Adult reformulations of child errors as negative evidence. *Journal of child language*, 30(3):637–669, 2003.

Paul Cisek and John F. Kalaska. Neural Mechanisms for Interacting with a World Full of Action Choices. *Annual Review of Neuroscience*, 33(1):269–298, 2010.

Alexander Clark and Shalom Lappin. *Linguistic Nativism and the Poverty of the Stimulus*. Wiley-Blackwell, 2011.

Andy Clark. Busting out: Predictive brains, embodied minds, and the puzzle of the evidentiary veil. *Noûs*, 51(4):727–753, 2017a.

Andy Clark. How to knit your own Markov blanket: Resisting the second law with metamorphic minds. In T. Metzinger & W. Wiese, editor, *Philosophy and Predictive Processing: 3. Frankfurt Am Main: MIND Group*. Johannes Gutenberg-Universität Mainz, 2017b. doi: 10.15502/9783958573031.

Herbert H. Clark. *Using Language*. Cambridge University Press, Cambridge, 1996.

Herbert H Clark. Communal lexicons. In Kirsten Malmkjær and John Williams, editors, *Context in Language Learning and Language Understanding*, chapter 4, pages 63–87. Cambridge University Press, Cambridge, 1998.

Herbert H Clark. Wordless questions, wordless answers. In De Ruiter, Jan P, editor, *Questions: Formal, Functional and Interactional Perspectives*, pages 81–100. Cambridge University Press, Cambridge, 2012.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. What Makes a Good Conversation? challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12. Association for Computing Machinery, New York, NY, USA, May 2019. ISBN 978-1-4503-5970-2.

Reuben Cohn-Gordon, Noah D. Goodman, and Christopher Potts. An Incremental Iterated Response Model of Pragmatics. *arXiv:1810.00367 [cs]*, October 2018.

Robin Cooper. Records and record types in semantic theory. *Journal of Logic and Computation*, 15 (2):99–112, 2005.

Robin Cooper. Type theory and semantics in flux. In Ruth Kempson, Nicholas Asher, and Tim Fernando, editors, *Philosophy of Linguistics*, volume 14 of *Handbook of the Philosophy of Science*, pages 271–323. North Holland (Elsevier), Amsterdam, 2012.

Robin Cooper and Jonathan Ginzburg. *Type Theory with Records for Natural Language Semantics**, chapter 12, pages 375–407. John Wiley & Sons, Ltd, 2015. ISBN 9781118882139. doi: 10.1002/9781118882139.ch12.

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. A probabilistic rich type theory for semantic interpretation. In *Proceedings of the EACL Workshop on Type Theory and Natural Language Semantics (TTNLS)*, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

Robin Cooper, Simon Dobnik, Shalom Lappin, and Staffan Larsson. Probabilistic type theory and natural language semantics. In *Linguistic Issues in Language Technology, Volume 10, 2015*, 2015.

Hanne De Jaegher and Ezequiel Di Paolo. Participatory sense-making. *Phenomenology and the Cognitive Sciences*, 6(4):485–507, December 2007. ISSN 1572-8676. doi: 10.1007/s11097-⊂007-⊂9076-⊂9.

Manuel DeLanda. *Intensive Science and Virtual Philosophy*. Bloomsbury, London, 2013.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-⊂1423.

Ezequiel Di Paolo. Extended Life. *Topoi*, 28(1):9, December 2008. ISSN 1572-8749. doi: 10.1007/s11245-⊂008-⊂9042-⊂3.

Ezequiel Di Paolo and Hanne De Jaegher. The interactive brain hypothesis. *Frontiers in Human Neuroscience*, 6:163, 2012. ISSN 1662-5161. doi: 10.3389/fnhum.2012.00163.

Mark Dingemanse. Resource-rationality beyond individual minds: the case of interactive language use. *Behavioral and Brain Sciences*, 43:e9, 2020. doi: 10.1017/S0140525X19001638.

Simon Dobnik, Robin Cooper, and Staffan Larsson. Modelling language, action, and perception in Type Theory with Records. In *Proceedings of the 7th International Workshop on Constraint Solving and Language Processing*, pages 51–63, 2012.

Hubert L. Dreyfus. The myth of the pervasiveness of the mental. In J. K. Schear, editor, *Mind, Reason, and Being-in-the-World*. Routledge, London, 2013.

Gerard Duveen and Charis Psaltis. The constructive role of asymmetry in social interaction. In Serge Moscovici, Sandra Jovchelovitch, and Brady Wagoner, editors, *Development as a Social Process: Contributions of Gerard Duveen*, pages 133–154. Routledge, Abingdon, UK, 2013.

Arash Eshghi and Oliver Lemon. How domain-general can we be? Learning incremental dialogue systems without dialogue acts. In *Proceedings of the 18th SemDial Workshop on the Semantics and Pragmatics of Dialogue (DialWatt)*, pages 53–61, 2014.

Arash Eshghi, Matthew Purver, and Julian Hough. DyLan: Parser for Dynamic Syntax. Technical report, Queen Mary University of London, 2011. EECSRR-11-05.

Arash Eshghi, Julian Hough, Matthew Purver, Ruth Kempson, and Eleni Gregoromichelaki. Conversational interactions: Capturing dialogue dynamics. In S. Larsson and L. Borin, editors, *From Quantification to Conversation: Festschrift for Robin Cooper on the occasion of his 65th birthday*, volume 19 of *Tributes*, pages 325–349. College Publications, London, 2012. ISBN 978-1-904987-91-2.

Arash Eshghi, Julian Hough, and Matthew Purver. Incremental grammar induction from child-directed dialogue utterances. In *Proceedings of the 4th Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 94–103, Sofia, Bulgaria, August 2013a. Association for Computational Linguistics.

Arash Eshghi, Matthew Purver, and Julian Hough. Probabilistic induction for an incremental semantic grammar. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 107–118, Potsdam, Germany, March 2013b. Association for Computational Linguistics.

Arash Eshghi, Matthew Purver, Julian Hough, and Yo Sato. Probabilistic grammar induction in an incremental semantic framework. In Duchier D. and Parmentier Y., editors, *CSLP, Lecture Notes in Computer Science*, volume 8114 of *Lecture Notes in Computer Science*, pages 92–107. Springer, Berlin, Heidelberg, 2013c.

Arash Eshghi, Christine Howes, Eleni Gregoromichelaki, Julian Hough, and Matt Purver. Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS)*, pages 261–271, London, UK, 2015. ACL.

Arash Eshghi, Igor Shalyminov, and Oliver Lemon. Bootstrapping incremental dialogue systems from minimal data: Linguistic knowledge or machine learning? In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2220–2230, 2017.

Raquel Fernández. *Non-Sentential Utterances in Dialogue: Classification, Resolution and Use*. PhD thesis, King's College London, University of London, 2006.

Joel E. Fischer, Stuart Reeves, Martin Porcheron, and Rein Ove Sikveland. Progressivity for voice interface design. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, CUI '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450371872. doi: 10.1145/3342775.3342788.

Jerry A Fodor. *The Language of Thought*, volume 5. Harvard University Press, Cambridge, MA, 1975.

Michael N. Forster. *Wittgenstein on the Arbitrariness of Grammar*. Princeton University Press, January 2009. ISBN 978-1-4008-2604-9.

Carol A. Fowler and Bert Hodges. Finding common ground: Alternatives to code models for language use. *New Ideas in Psychology*, 42:1–6, August 2016. ISSN 0732-118X. doi: 10.1016/j.newideapsych.2016.03.001.

Michael C. Frank and Noah D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012. doi: 10.1126/science.1218633.

Nico H Frijda, K Richard Ridderinkhof, and Erik Rietveld. Impulsive action: Emotional impulses and their control. *Frontiers in Psychology*, 5:518, 2014.

Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11 (2):127–138, February 2010. ISSN 1471-0048. doi: 10.1038/nrn2787.

Karl Friston. Embodied inference: Or "I think therefore I am, if I am what I think". In *The Implications of Embodiment: Cognition and Communication*, pages 89–125. Imprint Academic, Charlottesville, VA, 2011. ISBN 978-1-84540-240-2.

Karl Friston, Spyridon Samothrakis, and Read Montague. Active inference and agency: Optimal control without cost functions. *Biological Cybernetics*, 106(8):523–541, October 2012a. ISSN 1432-0770. doi: 10.1007/s00422-⊂012-⊂0512-⊂8.

Karl Friston, Spyridon Samothrakis, and Read Montague. Active inference and agency: Optimal control without cost functions. *Biological Cybernetics*, 106(8):523–541, October 2012b. ISSN 1432-0770. doi: 10.1007/s00422-⊂012-⊂0512-⊂8.

Donna T Fujimoto. Listener responses in interaction: A case for abandoning the term, backchannel. *Journal of Osaka Jogakuin College*, 37:35–54, 2007.

Jianfeng Gao, Michel Galley, and Lihong Li. Neural Approaches to Conversational AI. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1371–1374, New York, NY, USA, June 2018. Association for Computing Machinery. ISBN 978-1-4503-5657-2. doi: 10.1145/3209978.3210183.

Andrew Gargett, Eleni Gregoromichelaki, Ruth Kempson, Matthew Purver, and Yo Sato. Grammar resources for modelling dialogue dynamically. *Cognitive Neurodynamics*, 3(4):347–363, 2009.

James J. Gibson. *The Senses Considered as Perceptual Systems*. The Senses Considered as Perceptual Systems. Houghton Mifflin, Oxford, England, 1966.

James J Gibson. *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press, New York, 2014.

Jonathan Ginzburg. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford, 2012.

Jonathan Ginzburg and Dimitra Kolliakou. Answers without questions: The emergence of fragments in child language. *Journal of Linguistics*, 45(3):641–673, 2009. doi: 10.1017/S0022226709990053.

Noah D. Goodman and Michael C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829, 2016.

Eleni Gregoromichelaki. Grammar as action in language and music. In M. Orwin, R. Kempson, and C. Howes, editors, *Language, music and interaction*, pages 93–134. College Publications, 2013.

Eleni Gregoromichelaki. Quotation in Dialogue. In Paul Saka and Michael Johnson, editors, *The Semantics and Pragmatics of Quotation*, Perspectives in Pragmatics, Philosophy & Psychology, pages 195–255. Springer International Publishing, Cham, 2018. ISBN 978-3-319-68747-6.

Eleni Gregoromichelaki and Ruth Kempson. Joint Utterances and the (Split-) Turn-Taking Puzzle. In *Interdisciplinary Studies in Pragmatics, Culture and Society*, pages 703–743. Springer, 2015.

Eleni Gregoromichelaki, Ruth Kempson, Matthew Purver, Gregory J. Mills, Ronnie Cann, Wilfried Meyer-Viol, and Patrick G. T. Healey. Incrementality and intention-recognition in utterance processing. *Dialogue and Discourse*, 2(1):199–233, 2011.

Eleni Gregoromichelaki, Ruth Kempson, Christine Howes, and Arash Eshghi. On making syntax dynamic:. In Ipke Wachsmuth, Jan de Ruiter, Petra Jaecks, and Stefan Kopp, editors, *Alignment in Communication: Towards a New Theory of Communication*, pages 57–85. John Benjamin, 2013.

Eleni Gregoromichelaki, Christine Howes, Arash Eshghi, Ruth Kempson, Julian Hough, Mehrnoosh Sadrzadeh, Matthew Purver, and Gijs Wijnholds. Normativity, meaning plasticity, and the significance of Vector Space Semantics. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue*, London, United Kingdom, sep 2019. SEMDIAL.

Eleni Gregoromichelaki, Stergios Chatzikyriakidis, Arash Eshghi, Julian Hough, Christine Howes, Ruth Kempson, Jieun Kiaer, Matthew Purver, Mehrnoosh Sadrzadeh, and Graham White. Affordance competition in dialogue: the case of syntactic universals. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue*, 2020a.

Eleni Gregoromichelaki, Christine Howes, and Ruth Kempson. Actionism in syntax and semantics. In *Dialogue and Perception - Extended Papers from DaP2018*, volume 2 of *CLASP Papers in Computational Linguistics*, pages 12–27. GUPEA, Gothenburg, 2020b.

Eleni Gregoromichelaki, Gregory James Mills, Christine Howes, Arash Eshghi, Stergios Chatzikyriakidis, Matthew Purver, Ruth Kempson, Ronnie Cann, and Patrick G. T. Healey. Completability vs (In)completeness. *Acta Linguistica Hafniensia*, 52(2):260–284, July 2020c. ISSN 0374-0463. doi: 10.1080/03740463.2020.1795549.

H. Paul Grice. Logic and conversation. *Syntax and Semantics*, 3(S 41):58, 1975.

Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

Austin W. Hanjie, Victor Zhong, and Karthik Narasimhan. Grounding Language to Entities and Dynamics for Generalization in Reinforcement Learning. *arXiv:2101.07393 [cs]*, June 2021.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990. ISSN 0167-2789. doi: 10.1016/0167-⊂2789(90)90087-⊂6.

Patrick G. T. Healey and Gregory J. Mills. Participation, precedence and co-ordination in dialogue. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 1470–1475, 2006.

Patrick G. T. Healey, Jan Peter de Ruiter, and Gregory J. Mills. Editors' introduction: Miscommunication. *Topics in Cognitive Science*, 10(2):264–278, 2018a. doi: https://doi.org/10.1111/tops.12340.

Patrick G. T. Healey, Gregory J. Mills, Arash Eshghi, and Christine Howes. Running repairs: Coordinating meaning in dialogue. *Topics in Cognitive Science*, 10(2):367–388, 2018b. ISSN 1756-8765.

Felix Hill, Olivier Tieleman, Tamara von Glehn, Nathaniel Wong, Hamza Merzic, and Stephen Clark. Grounded Language Learning Fast and Slow. *arXiv:2009.01719 [cs]*, October 2020.

Felix Hill, Olivier Tieleman, Tamara von Glehn, Nathaniel Wong, Hamza Merzic, and Stephen Clark. Grounded language learning fast and slow. In *International Conference on Learning Representations (ICLR)*, 2021.

Jakob Hohwy. *The Predictive Mind*. Oxford University Press, November 2013. ISBN 978-0-19-968273-7.

Julian Hough. *Modelling Incremental Self-Repair Processing in Dialogue*. PhD thesis, Queen Mary University of London, 2015.

Julian Hough and Matthew Purver. Processing self-repairs in an incremental type-theoretic dialogue system. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (SeineDial)*, pages 136–144, Paris, France, September 2012.

Julian Hough and Matthew Purver. Probabilistic type theory for incremental dialogue processing. In *Proceedings of the EACL 2014 Workshop on Type Theory and Natural Language Semantics (TTNLS)*, pages 80–88, Gothenburg, Sweden, April 2014a. Association for Computational Linguistics. ISBN 978-1-937284-74-9.

Julian Hough and Matthew Purver. Lattice theoretic relevance in incremental reference processing. In *Proceedings of the RefNet Workshop on Psychological and Computational Models of Reference Comprehension and Production*, Edinburgh, August 2014b.

Julian Hough and Matthew Purver. Probabilistic record type lattices for incremental reference processing. In Stergios Chatzikyriakidis and Zhaohui Luo, editors, *Modern Perspectives in Type-theoretical Semantics*, pages 189–222. Springer, 2017.

Julian Hough, Casey Kennington, David Schlangen, and Jonathan Ginzburg. Incremental semantics for dialogue processing: Requirements, and a comparison of two approaches. In *Proceedings of the 11th International Conference on Computartional Semantics (IWCS)*, London, UK, 2015.

Julian Hough, Lorenzo Jamone, David Schlangen, Guillaume Walck, Robert Haschke, et al. Towards a Types-as-Classifiers Approach to Dialogue Processing in Human-Robot Interaction. In *Proceedings of the Workshop on Dialogue and Perception (DaP 2018), Gothenburg*, 2018.

Julian Hough, Lorenzo Jamone, David Schlangen, Guillaume Walck, and Robert Haschke. *Dialogue and Perception - CLASP Papers in Computational Linguistics*, volume 2, chapter A Types-As-Classifiers Approach to Human-Robot Interaction for Continuous Structured State Classification, pages 28–40. Centre for Linguistic Theory and Studies in Probability (CLASP), Department of Philosophy, Linguistics and Theory of Science (FLOV), University of Gothenburg, February 2020.

Christine Howes. *Coordination in Dialogue: Using Compound Contributions to Join a Party*. PhD thesis, Queen Mary University of London, 2012.

Christine Howes and Arash Eshghi. Feedback relevance spaces: The organisation of increments in conversation. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*. ACL, 2017.

Christine Howes and Arash Eshghi. Feedback relevance spaces: Interactional constraints on processing contexts in Dynamic Syntax. *Journal of Logic, Language and Information*, 2021. Accepted for publication.

Christine Howes, Matthew Purver, Patrick G. T. Healey, Gregory J. Mills, and Eleni Gregoromichelaki. On incrementality in dialogue: Evidence from compound contributions. *Dialogue and Discourse*, 2(1):279–311, 2011.

Christine Howes, Patrick G. T. Healey, Matthew Purver, and Arash Eshghi. Finishing each other's ... responding to incomplete contributions in dialogue. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, pages 479–484, Sapporo, Japan, 2012. ISBN 978-0-9768318-8-4.

Edwin Hutchins. *Cognition in the Wild*. MIT Press, 1995. ISBN 978-0-262-58146-2.

Dimitrios Kalatzis, Arash Eshghi, and Oliver Lemon. Bootstrapping incremental dialogue systems: Using linguistic knowledge to learn from minimal data. In *Proceedings of the NIPS 2016 Workshop on Learning Methods for Dialogue*, Barcelona, 2016.

Ruth Kempson, Wilfried Meyer-Viol, and Dov Gabbay. *Dynamic Syntax: The Flow of Language Understanding*. Wiley-Blackwell, Oxford, 2001.

Ruth Kempson, Eleni Gregoromichelaki, and Christine Howes(eds.). *The Dynamics of Lexical Interfaces*. Studies in Constraint Based Lexicalism. CSLI, Stanford, CA, 2011.

Ruth Kempson, Ronnie Cann, Arash Eshghi, Eleni Gregoromichelaki, and Matthew Purver. Ellipsis. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*. Wiley-Blackwell, Oxford, 2015.

Ruth Kempson, Ronnie Cann, Eleni Gregoromichelaki, and Stergios Chatzikiriakidis. Language as mechanisms for interaction. *Theoretical Linguistics*, 42(3-4):203–275, 2016.

Ruth Kempson, Ronnie Cann, Eleni Gregoromichelaki, and Stergios Chatzikyriakidis. Action-Based Grammar. *Theoretical Linguistics*, 43(1-2):141–167, 2017.

Michael D. Kirchhoff and Tom Froese. Where There is Life There is Mind: In Support of a Strong Life-Mind Continuity Thesis. *Entropy*, 19(4):169, April 2017. doi: 10.3390/e19040169.

Stefan Kopp and Nicole Krämer. Revisiting human-agent communication: The importance of joint co-construction and understanding mental states. *Frontiers in Psychology*, 12:597, 2021. ISSN 1664-1078. doi: 10.3389/fpsyg.2021.580955.

Satwik Kottur, Ramakrishna Vedantam, Jose M. F. Moura, and Devi Parikh. Visual Word2Vec (vis-w2v): Learning Visually Grounded Word Embeddings Using Abstract Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4985–4994, 2016.

Richard Kunert, Raquel Fernández, and Willem Zuidema. Adaptation in child directed speech: Evidence from corpora. *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue (LosAngelogue)*, 2011.

Shalom Lappin. *Deep Learning and Linguistic Representation*. CRC Press, April 2021. ISBN 978-1-00-038032-3.

Staffan Larsson. The TTR perceptron: Dynamic perceptual meanings and semantic coordination. In *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2011 - Los Angelogue)*, pages 140–148, September 2011.

Staffan Larsson. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 25(2):335–369, 2015.

Jey Han Lau, Alexander Clark, and Shalom Lappin. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41(5):1202–1241, 2017. ISSN 1551-6709. doi: 10.1111/cogs.12414.

Jey Han Lau, Carlos Armendariz, Shalom Lappin, Matthew Purver, and Chang Shu. How Furiously Can Colorless Green Ideas Sleep? sentence Acceptability in Context. *Transactions of the Association for Computational Linguistics*, 8:296–310, June 2020. ISSN 2307-387X. doi: 10.1162/tacl_a_00315.

Gene H. Lerner. On the syntax of sentences-in-progress. *Language in Society*, pages 441–458, 1991.

Stephen C. Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6:731, 2015. ISSN 1664-1078. doi: 10.3389/fpsyg.2015.00731.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-⊂1259.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-⊂main.428.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. End-to-end task-completion neural dialogue systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 733–743, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. Dialogue Learning with Human Teaching and Feedback in End-to-End Trainable Task-Oriented Dialogue Systems. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2060–2069, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-⊂1187.

Ryan Thomas Lowe, Nissan Pow, Iulian Serban, Laurent Charlin, Chiao-Wei Liu, and Joelle Pineau. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue Discourse*, 8: 31–65, 2017.

Xin Lu, Barbara Di Eugenio, Trina C Kershaw, Stellan Ohlsson, and Andrew Corrigan-Halpern. Expert vs. non-expert tutoring: Dialogue moves, interaction patterns and multi-utterance turns. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 456–467, 2007.

Ewa Luger and Abigail Sellen. "like having a really bad pa": The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 5286–5297, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450333627. doi: 10.1145/2858036.2858288.

Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, New Jersey, third edition, 2000.

David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. WH Freeman and Company, San Francisco, 1982.

Colin Matheson, Massimo Poesio, and David Traum. Modeling grounding and discourse obligations using update rules. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Seattle, Washington, April 2000.

Christoph Mathys, Jean Daunizeau, Karl Friston, and Klaas Stephan. A Bayesian Foundation for Individual Learning Under Uncertainty. *Frontiers in Human Neuroscience*, 5:39, 2011. ISSN 1662-5161. doi: 10.3389/fnhum.2011.00039.

A. Maye and A. K. Engel. A discrete computational model of sensorimotor contingencies for object perception and control of behavior. In *2011 IEEE International Conference on Robotics and Automation*, pages 3810–3815, 2011.

Ruth Garrett Millikan. *On clear and confused ideas: An essay about substance concepts*. Cambridge University Press, Cambridge, 2000.

Gregory J. Mills. *Semantic co-ordination in dialogue: The role of direct interaction.* PhD thesis, Queen Mary University of London, 2007.

Gregory J. Mills. The emergence of procedural conventions in dialogue. In *IProceedings of the 33rd Annual Conference of the Cognitive Science Society*, 2011.

Gregory J. Mills. Dialogue in joint activity: Complementarity, convergence and conventionalization. *New Ideas in Psychology*, 32:158–173, 2014.

Gregory J. Mills and Eleni Gregoromichelaki. Establishing coherence in dialogue: Sequentiality, intentions and negotiation. In *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue*, 2010.

Gregory J. Mills and Patrick G. T. Healey. Clarifying spatial descriptions: Local and global effects on semantic co-ordination. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL)*, Potsdam, Germany, September 2006.

Robert Mirski and Mark H. Bickhard. Conventional minds: An interactivist perspective on social cognition and its enculturation. *New Ideas in Psychology*, 62:100856, August 2021. ISSN 0732-118X. doi: 10.1016/j.newideapsych.2021.100856.

Roger K. Moore. Is Spoken Language All-or-Nothing? implications for Future Speech-Based Human-Machine Interaction. In Kristiina Jokinen and Graham Wilcock, editors, *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, Lecture Notes in Electrical Engineering, pages 281–291. Springer, Singapore, 2017. ISBN 978-981-10-2585-3. doi: 10.1007/978-⊂981-⊂10-⊂2585-⊂3_22.

Bill Noble and Vladislav Maraev. Large-scale text pre-training helps with dialogue act recognition, but not without fine-tuning. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*. Association for Computational Linguistics, 2021.

Alva Nöe. *Action in Perception*. MIT Press, first edition, 2004.

Alva Noë. *Varieties of presence*. Harvard University Press Cambridge, MA, 2012.

Ezequiel A. Di Paolo, Elena Clare Cuffari, and Hanne De Jaegher. *Linguistic Bodies: The Continuity between Life and Language*. MIT Press, November 2018. ISBN 978-0-262-03816-4.

Hae Won Park, Mirko Gelsomini, Jin Joo Lee, and Cynthia Breazeal. Telling stories to robots: The effect of backchanneling on a child's storytelling. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 100–108. ACM, 2017.

Martin Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226, 2004.

Martin J Pickering and Simon Garrod. *Understanding Dialogue: Language Use and Social Interaction*. Cambridge University Press, Cambridge, 2021.

Alison Pilnick and Robert Dingwall. On the remarkable persistence of asymmetry in doctor/patient interaction: A critical review. *Social science & medicine*, 72(8):1374–1382, 2011.

Massimo Poesio and Hannes Rieser. Completions, coordination, and alignment in dialogue. *Dialogue and Discourse*, 1:1–89, 2010.

Geoffrey Pullum and Barbara Scholz. On the distinction between model-theoretic and generative-enumerative syntactic frameworks. In G. Morrill, P. Le Groote, and C. Retore, editors, *Proceedings of the 4th International Conference on Logical Aspects of Computational Linguistics (LACL)*, pages 17–43. Springer, 2001.

Matthew Purver. *The Theory and Use of Clarification Requests in Dialogue*. PhD thesis, University of London, 2004.

Matthew Purver, Ronnie Cann, and Ruth Kempson. Grammars as parsers: Meeting the dialogue challenge. *Research on Language and Computation*, 4(2-3):289–326, 2006.

Matthew Purver, Eleni Gregoromichelaki, Wilfried Meyer-Viol, and Ronnie Cann. Splitting the 'I's and crossing the 'you's: Context, speech acts and grammar. In P. Łupkowski and M. Purver, editors, *Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue*, pages 43–50, Poznań, June 2010. Polish Society for Cognitive Science.

Matthew Purver, Arash Eshghi, and Julian Hough. Incremental semantic construction in a dialogue system. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, pages 365–369, Oxford, UK, January 2011.

James Pustejovsky and Nikhil Krishnaswamy. Embodied Human Computer Interaction. *KI - Künstliche Intelligenz*, 35(3):307–327, November 2021. ISSN 1610-1987. doi: 10.1007/s13218-⊂021-⊂00727-⊂5.

Erik Rietveld, Damiaan Denys, and Maarten Van Westen. Ecological-enactive cognition as engaging with a field of relevant affordances. *The Oxford handbook of 4E cognition*, page 41, 2018.

Joanna Rączaszek-Leonardi and J. A. Scott Kelso. Reconciling symbolic and dynamic aspects of language: Toward a dynamic psycholinguistics. *New Ideas in Psychology*, 26(2):193–207, August 2008. ISSN 0732-118X. doi: 10.1016/j.newideapsych.2007.07.003.

Joanna Rączaszek-Leonardi, Iris Nomikou, and Katharina J. Rohlfing. Young Children's Dialogical Actions: The Beginnings of Purposeful Intersubjectivity. *IEEE Transactions on Autonomous Mental Development*, 5(3):210–221, September 2013. ISSN 1943-0612. doi: 10.1109/TAMD.2013.2273258.

Joanna Rączaszek-Leonardi, Agnieszka Dębska, and Adam Sochanowicz. Pooling the ground: Understanding and coordination in collective sense making. *Frontiers in Psychology*, 5:1233, 2014. ISSN 1664-1078. doi: 10.3389/fpsyg.2014.01233.

Joanna Rączaszek-Leonardi, Iris Nomikou, Katharina J. Rohlfing, and Terrence W. Deacon. Language Development From an Ecological Perspective: Ecologically Valid Ways to Abstract Symbols. *Ecological Psychology*, 30(1):39–73, January 2018. ISSN 1040-7413. doi: 10.1080/10407413.2017.1410387.

Frank Röder, Ozan Özdemir, Phuong D. H. Nguyen, Stefan Wermter, and Manfred Eppe. The Embodied Crossmodal Self Forms Language and Interaction: A Computational Cognitive Review. *Frontiers in Psychology*, 12:3374, 2021. ISSN 1664-1078. doi: 10.3389/fpsyg.2021.716671.

Stephen Roller, Y.-Lan Boureau, Jason Weston, Antoine Bordes, Emily Dinan, Angela Fan, David Gunning, Da Ju, Margaret Li, Spencer Poff, Pratik Ringshia, Kurt Shuster, Eric Michael Smith, Arthur Szlam, Jack Urbanek, and Mary Williamson. Open-Domain Conversational Agents: Current Progress, Open Problems, and Future Directions. *arXiv:2006.12442 [cs]*, July 2020.

Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M. Lake. A Benchmark for Systematic Generalization in Grounded Language Understanding. *arXiv:2003.05161 [cs]*, October 2020.

Harvey Sacks, Emmanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974.

Yo Sato. Local ambiguity, search strategies and parsing in Dynamic Syntax. In E. Gregoromichelaki, R. Kempson, and C. Howes, editors, *The Dynamics of Lexical Interfaces*. CSLI Publications, Stanford, CA, 2011.

Matthew Saxton. The contrast theory of negative input. *Journal of Child Language*, 24(1):139–161, 1997.

Matthew Saxton, Bela Kulcsar, Greer Marshall, and Mandeep Rupra. Longer-term effects of corrective input: An experimental approach. *Journal of Child Language*, 25(3):701–721, October 1998. ISSN 1469-7602, 0305-0009. doi: 10.1017/S0305000998003559.

Matthew Saxton, Carmel Houston–Price, and Natasha Dawson. The prompt hypothesis: Clarification requests as corrective input for grammatical errors. *Applied Psycholinguistics*, 26(3):393–414, July 2005. ISSN 1469-1817, 0142-7164. doi: 10.1017/S0142716405050228.

Emanuel A Schegloff. Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. *Analyzing discourse: Text and talk*, 71:93, 1982.

Emmanuel A. Schegloff. Reflections on quantification in the study of conversation. *Research on Language and Social Interaction*, 26:99–128, 1993.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016.

Lifeng Shang, Zhengdong Lu, and Hang Li. Neural Responding Machine for Short-Text Conversation. *arXiv:1503.02364 [cs]*, April 2015.

Claude E. Shannon and Warren Weaver. *A Mathematical Model of Communication*. University of Illinois Press, Champaign, IL, 1949.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado, May 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-⊂1020.

Dan Sperber and Deirdre Wilson. *Relevance: Communication and Cognition*. Blackwell, Oxford, second edition, 1995.

Mark Steedman. Plans, affordances, and combinatory grammar. *Linguistics and Philosophy*, 25(5): 723–753, 2002.

Lucy A Suchman. *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press, Cambridge, UK, 1987.

Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-⊂1514.

James Trafford. Reconstructing intersubjective norms. *Phenomenology and Mind*, 13:176–182, 2017. ISSN 2239-4028 (Online); 2280-7853 (Print). doi: 10.13128/Phe_Mi-⊂22440.

Alexander Tschantz, Beren Millidge, Anil K. Seth, and Christopher L. Buckley. Reinforcement Learning through Active Inference. *arXiv:2002.12636 [cs, eess, math, stat]*, February 2020.

Jacqueline Van Arkel, Marieke Woensdregt, Mark Dingemanse, and Mark Blokpoel. A simple repair mechanism can alleviate computational demands of pragmatic reasoning: Simulations and complexity analysis. In *the 24th (Virtual) Conference on Computational Natural Language Learning (CoNLL 2020)*, pages 177–194. ACL, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Samuel P. L. Veissière, Axel Constant, Maxwell J. D. Ramstead, Karl J. Friston, and Laurence J. Kirmayer. Thinking through other minds: A variational approach to cognition and culture. *Behavioral and Brain Sciences*, 43, 2020. ISSN 0140-525X, 1469-1825. doi: 10.1017/S0140525X19001213.

Oriol Vinyals and Quoc Le. A neural conversational model. *arXiv*, (1506.05869v3), 2015.

L. Wittgenstein. *Philosophical Investigations, Trans. G.E.M. Anscombe*. Oxford: Blackwell, 1953.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. Transfertransfo: A transfer learning approach for neural network based conversational agents. *ArXiv*, abs/1901.08149, 2019.

Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Yanchao Yu, Arash Eshghi, and Oliver Lemon. Training an adaptive dialogue policy for interactive learning of visually grounded word meanings. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 339–349, Los Angeles, 2016.

Wlodek W. Zadrozny. Towards Coinductive Models for Natural Language Understanding. Bringing together Deep Learning and Deep Semantics. *arXiv:2012.05715 [cs]*, December 2020.

Wlodek W. Zadrozny. Abstraction, Reasoning and Deep Learning: A Study of the "Look and Say" Sequence. *arXiv:2109.12755 [cs]*, September 2021.