How infants learn word meanings and propositional attitudes: a neural network model

Alistair Knott

1 Tomasello's account of word learning and pragmatic development

Michael Tomasello's influential model of language development (Tomasello, 2000; Tomasello, 2003) emphasises the role of infants' pragmatic understanding of the world in supporting their learning of language. For all humans, whether adult or child, language has a pragmatic function: we communicate linguistically in order to further social goals (for instance to share our beliefs and desires with others, or to ascertain the beliefs and desires of others) or to pursue joint undertakings (for instance to collaborate in a shared task). Tomasello imagines an infant observing a speaker producing an utterance directed at a hearer (possibly the infant herself). He proposes that the infant's understanding of the intentions of the speaker and hearer in this context plays a crucial role in her ability to learn an association between the linguistic form of the utterance and its meaning. The infant interprets the meaning of the utterance in the light of her understanding of the speaker's current goals, and of how the hearer features in these goals. At some point during development, the infant becomes aware of the general fact that human agents perform actions in service of goals. In Tomasello's account, learning this fact is a precondition for learning language. When the pragmatically-aware infant observes a speaker producing an utterance, she attempts to infer the speaker's goals, and then forms hypotheses about how the words in the utterance further these goals.

In Tomasello's model, an early instance of pragmatic learning in language development concerns the learning of the meanings of individual words. Words are symbols that denote concepts. While some theories of language development assume the relationship between words and concepts is founded simply in the existence of regular associations or co-occurrences between words and concepts in the infant's mind, for Tomasello it is fundamentally pragmatic in origin, and has to be learned by infants in pragmatic analyses of speech events. What infants must learn is that that *uttering words can serve to evoke representations in the mind of the hearer*. It is only after learning this general fact that infants can properly begin to learn the meanings of particular words.

Tomasello sees two pragmatic ablities as prerequisites for learning word meanings. One is the ability to establish **joint attention** with an observed agent, i.e. to attend to the same object the observed agent is attending to. When an infant has acquired this ability, a speaker's gaze will direct her attention to particular objects. Since speakers often visually attend to the situations they describe linguistically, the infant can learn that words can likewise serve to direct attention to arbitrary concepts. The other important pragmatic ability is the ability to infer the **communicative intentions** of an observed agent, i.e. to infer the goal underlying the agent's communicative actions. The infant must learn that there is a special class of actions (communicative actions) which have communicative effects rather than physical effects. Communicative actions are physical actions, which are directed at another agent, who is physically present in the communicative situation. But unlike regular physical actions, their effect is on the agent's mental state rather than on his physical state: specifically, they evoke representations in the agent's mind. In spoken language, these actions are articulatory gestures that realise phonological word forms. Tomasello argues that the infant

Dept of Computer Science, University of Otago. e-mail: alik@cs.otago.ac.nz

must be able to identify the special communicative effects of such actions before she can learn to associate specific actions with specific effects.

Tomasello advances both conceptual and empirical arguments for his proposal about word learning. The conceptual arguments turn on the question of what it means for words to be symbols: as just summarised, Tomasello argues that the meanings of words are more than just concepts which are regularly associated with them. The empirical arguments are of three types. Firstly he argues that infants acquire the social-pragmatic skills needed to learn words (joint attention and the ability to recognise the intentions underlying communicative actions) around the age of 9–12 months—and that this is also the age at which infants start to learn word meanings (see e.g. Tomasello, 1995). Secondly he argues that these social-pragmatic skills mark a key difference between humans and their closest evolutionary cousins, great apes (see e.g. Tomasello and Herrmann, 2010). Finally, he cites evidence that human infants systematically interpret speakers' utterances with reference to their inferred intentions (see e.g. Diesendruck *et al.*, 2004).

While there are good grounds for Tomasello's proposal about the role of social-pragmatic abilities in learning word meanings, the exact nature of these abilities is less clear. For one thing, it is unclear how an infant *represents* the communicative intentions of a speaker, or the special properties of the actions which achieve communicative intentions. We do not have good models of how intentions of any kind are represented in the brain. There are promising models of the cognitive states which store prepared physical actions (e.g. Miller and Cohen, 2001) or prepared sequences of physical actions (e.g. Averbeck *et al.*, 2002), but these are some way from providing a full model of the outcomes which agents intend when they perform actions. And the representation of *communicative* intentions presents particular problems. A communicative intention is an intention to bring about a certain mental state in the mind of a hearer—for instance, to make the hearer entertain a certain belief about the world. We do not have good models of how agents represent the current beliefs of other agents, let alone their intended beliefs. In Tomasello's model, infants draw inferences about the communicative intentions of speakers, and use these (along with the ability to establish joint attention) to help learn word meanings—but Tomasello does not make any suggestions about how these inferred intentions are represented in the infant brain. Moreover, he does not make any suggestions about how the development of social-pragmatic abilities in infants leads to their ability to learn word meanings. Tomasello's account proposes that infants' social-pragmatic understanding influences their ability to learn word meanings—but he does not give an account of the neural mechanisms through which this influence is exerted.

In one sense it is prudent to express the social-pragmatic theory of language learning at a high level, given that we know so little about the neural representations of intentions and mental states. On the other hand, if our interest is in understanding how the brain represents intentions and mental states, Tomasello's developmental account provides potentially useful information. Presumably infants' earliest representations of the mental states and communicative intentions of other agents are fairly simple, and become more complex as development progresses. It may be easier to model early proto-mental states and proto-communicative-intentions than to model the mature representations that eventually develop—and a model of the representations which emerge early in development may contain clues about the mature representations that emerge later.

In this paper I will introduce a neural network model that addresses both how infants represent the communicative actions of observed agents, and how this representation supports the learning of individual word meanings. I will begin in Section 2 by reviewing some of the difficulties to be tackled in formulating a model of communicative action representations. In Section 3 I will present a neural network model of vocabulary learning and its interaction with the development of a simple concept of communicative actions. In Section 4 I will discuss how this model suggests some solutions to the problems inherent in representing communicative actions, and I will conclude with a discussion of the model in Section 5.

2 Neural representations of communicative actions

Consider a situation in which a mother tells her infant that a dog is chasing a cat. In Tomasello's model, the infant represents this action of her mother in a way that highlights its special communicative status. What

Fig. 1: (a) Representation of a single proposition by association of concepts with roles. (b) Problematic representation of nested propositions.

might this representation look like? I will discuss two issues that need to be addressed, that both relate to the fact that communicative actions *express propositions*: the mother tells her infant [that P is the case].

The first issue is specific to models of neural representation. Communicative action representations feature *nested* propositions. The outer proposition is about the act of communication (the mother tells the infant something); but the material that is communicated is also a proposition in its own right (a dog is chasing a cat). Nested propositions are hard to represent in a neural network. Representing a simple proposition involves activating a collection of concepts (for instance, MOTHER, INFANT, TELL), but it is also important to indicate the roles these concepts play in the depicted action: for instance that TELL indicates the ACTION-TYPE, MOTHER is the AGENT of the action, and INFANT is its TARGET or PATIENT. There are several ways of binding concepts to roles. Most straightforwardly, we could create direct associations between concepts and role labels (Chang, 2002), as shown in Figure 1a. This works well for simple propositions, but it is problematic if there is a nested proposition. The roles played by concepts in the nested proposition must also be represented; in our example, where the communicated proposition is that a dog is chasing a cat, we must associate DOG with AGENT, CAT with PATIENT and CHASE with the ACTION-TYPE. Simply overlaying these associations on top of those defining the outer proposition (as shown in Figure 1b) creates problems: there is nothing in this representation to indicate *which* agent and patient participate in *which* action. Many solutions to this problem have been devised (see e.g. Plate, 2003; van der Velde and de Kamps, 2006). For the moment I just want to note that nested propositions pose special problems for neural networks, which require special solutions.

The second issue to be addressed concerns how the relation between the outer and inner propositions should be encoded. The outer proposition describes the act of *expressing*, or *representing*, the inner proposition. The verb *tell* links a person (the speaker) to a represented state of affairs in the world: in philosophical parlance, it expresses a **propositional attitude** of the speaker towards a certain state of affairs. An agent can adopt different attitudes towards a represented proposition: he can believe it, desire it, doubt it, fear it—he can also *tell* it to someone else. In each case, the agent's attitude is 'about' a proposition: to use philosophical terminology once again, it has **intentionality** (Brentano, 1874; for discussion see Dennett, 1989, Jacquette, 2004). This relation of 'aboutness' between an agent's attitude and its propositional content is notoriously hard to define. But there are some well-known properties of statements about propositional attitudes that any account of this relation must capture. Firstly, when we assert that agent *A* adopts a given attitude towards proposition *P*, we commit ourselves to the existence of agent *A* and her adoption of this attitude, but we are not committed to the truth of *P*. For example, when I say that *Jane believes a dog is chasing a cat*, I am asserting that Jane exists, and has a certain belief—but I am *not* asserting the content of this belief, i.e. that a dog is chasing a cat. (I am not even committed to the *existence* of the dog and the cat. I can assert that Jane believes a unicorn is chasing a dragon without believing in unicorns and dragons myself.) Secondly, assertions about propositional attitudes are **intensional**: that is to say, their truth depends on the way the content of these attitudes is reported linguistically. For instance, assume the cat featuring in Jane's propositional attitude happens to belong to the Prime Minister, but that Jane does not know this. *Jane believes a dog is chasing a cat* is a true statement, but *Jane believes a dog is chasing the Prime Minister's cat* is not. In regular assertions about the physical world, the truth of a statement is not dependent on language in this way. For instance, if Jane is the Prime Minister's daughter, and Jane sneezed, then we can truly assert *The Prime Minister's daughter sneezed*: whether Jane knows she is the Prime Minister's daughter or not is irrelevant. To account for these properties of propositional attitudes, logicians traditionally adopted modal logic as a knowledge representation formalism, allowing reference to possible worlds other than the actual world in representations of propositional content, and in representations of the mean-

ing of words (see e.g. Montague, 1974). A more recent strategy is to model propositions and words as the cognitive states of agents, which may or may not reflect the current state of the world (see e.g. Gärdenfors, 2004): this is the approach I will take.

Statements about an agent's communicative actions express propositional attitudes in both the respects just described. The statement '*X* tells *Y* that *P*' asserts the existence of the speaker *X* and the hearer *Y*, and the the fact that a telling event occurred, but it does not assert that *P*. And clearly, the words that report the telling action have a bearing on the action which is reported: asserting that *Jane told her daughter the dog was chasing the cat* is different from asserting that *Jane told her daughter the dog was chasing the Prime Minister's cat*. At the same time, telling (and other communicative actions) are unusual as propositional attitudes. Whereas the prototypical attitudes (believing, desiring etc) are pure mental states, communicative actions are substantive actions: they have motor components as well as mental components. Identifying a communicative action involves processing a speaker's physical actions. For a linguistic action, these are typically articulatory gestures expressing a sequence of word forms. Word forms are associated with concepts, and the manner in which word forms are assembled conveys information about how these concepts are connected to form propositions, so the gestures which form the physical component of a communicative action collectively convey the propositional content of the action. A hearer who knows the language being used can recover the propositional content of the action from the gestures. Identifying the 'purely mental' propositional attitudes of an agent (e.g. beliefs and desires) is not so closely tied to the processing of a particular type of action. Mental states like desires and beliefs can be inferred from a variety of sources: for instance facial expressions or overt behaviour. Of course they can also be inferred from linguistic utterances: if someone says *P*, a strong default is to infer they believe that *P*. But communicative actions are unique in being *tied* to particular movements: they convey propositional attitudes conventionally through physical movements.

This close connection with physical movements makes the development of communicative action representations a natural first step in the development of propositional attitude representations. Tomasello's proposal that infants must learn to identify communicative actions at an early point during language learning thus fits well within an account of the development of mature propositional attitude representations. But as discussed in Section 1, Tomasello does not say anything about how infants represent communicative actions. In this section I have outlined two requirements for any model of the communicative action representations developed by infants. Firstly, it must provide a means for representing nested propositions. Secondly it must capture the intentionality of communicative actions: the elusive relation between a physical speaking action and the propositional content it expresses. I turn now to a computational implementation of Tomasello's developmental theory that aims to address the open questions about representations that it raises.

3 A neural network model of the role of communicative action concepts in word learning

In this section I will describe a neural network model of infant word learning which also gives an account of how infants develop simple representations of communicative actions. The model is intended to describe developmental processes occurring between the ages of around 10–18 months. In the model, infants' learning of word meanings and their development of communicative action representations bootstrap one another: learning word meanings facilitates the development of communicative action representations, which in turn facilitate the learning of word meanings. The model is called **pragmatic bootstrapping**; the network which implements the model is described in detail in Caza and Knott (2012). In the current paper my focus is on the representations of communicative actions which the network learns; technical details of the model can be found in Caza and Knott (2012).

Time		t_1	t_2	t_3	t_4	t_5
Concepts	AGENT	DADDY	COOKIE	DOG	MUMMY	CAT
	ACTION-TYPE	LAUGH		JUMP	TALK	RUN
Word forms		*puppy*	*tv*	*toy*	*cat*	
		ball	*tv*	*break*	*run*	

Table 1: Input data to the network: parallel streams of concepts and word forms

3.1 Input data

The input data for the network is a stream of word form representations, and a parallel stream of conceptual representations. These simulate the inputs being received by an infant. The conceptual representations are assumed to be delivered by the sensorimotor system; these reflect the infant's real-time visual, motor and auditory experience of the world. The word form representations are phonological encodings of words produced by mature speakers in the infant's current environment, again in real time. We assume that the infant is already able to identify individual words as phonological units, using the statistical learning abilities documented by Saffran *et al.* (1996). We also assume the infant is able to follow the gaze of an observed human agent—an ability which is also attested at 10 months, and which develops rapidly from 10-18 months (Butterworth and Jarrett, 1991). And we assume that infants regularly follow the gaze of observed speakers—an ability which also develops rapidly during this period (Baldwin *et al.*, 1996).

An example of the network's input data is shown in Table 1. We assume that the infant is able to identify simple episodes taking place in her environment, and to represent these as structures of concepts, perhaps using the kind of scheme illustrated in Figure 1a, each involving an agent performing an action. As the infant observes the world, she evokes a sequence of conceptual representations: at t_1 she observes daddy laughing, at t_2 she observes a cookie, at t_3 she observes a dog jumping and so on. At the same time, the infant is hearing words produced by mature speakers in the environment.

In this input data we introduce a weak correlation between words and conceptual representations. We assume the data represents a succession of situations, each of which contains a subset of the set of agents, performing a subset of the set of possible actions. The words heard by the infant in a given situation are more likely than chance to refer to the agents and actions in this situation, because speakers regularly talk about objects and actions in the current situation—so through a process of cross-situational learning (Siskind, 1996), the infant can slowly learn correct associations between concepts and words. But there is a large amount of noise in the mapping between concepts and words in the input data, so this cross-situational learning is very inefficient.

We also introduce another kind of regularity in the input data which provides a better opportunity to learn concept-word mappings. We assume that infants routinely follow the gaze of speakers, and that speakers often look at the objects and events they are talking about if they are physically present (see e.g. Yu and Ballard, 2007). If the infant happens to apprehend an episode in which a mature speaker is talking, then there is a brief moment afterwards when the correlation between concepts and words in the infant's mind is much stronger than usual: the infant is more likely than usual to be perceiving an episode that the speaker is talking about. This is illustrated at t_4 and t_5. At t_4, the infant apprehends her mother talking (MUMMY TALK). The infant then follows the speaker's gaze, and at t_5, perceives an episode in which a cat runs (CAT RUN). At the same time, in the input medium representing incoming word forms, the infant is representing the words produced by the speaker, in particular the words *cat* and *run*.[1] The moment just after the infant perceives a speaker talking therefore constitutes a particularly good opportunity to learn a mapping between words and concepts. For the infant, recognition of a talk action can be thought of *as a cue to engage in some word-meaning learning*.

It is important to say a little more about what is going on at times t_4 and t_5. At t_4, the infant observes a physical action, which in some ways is very similar to other motor actions such as jumping or laughing: an

[1] Our simulation considers only content words: we do not consider the issue of how the meanings of function words are learned, or how the infant learns the syntactic principles that map surface sequences of words with to episode representations. But these issues are the focus of a separate neural network model (see Takac *et al.*, 2012).

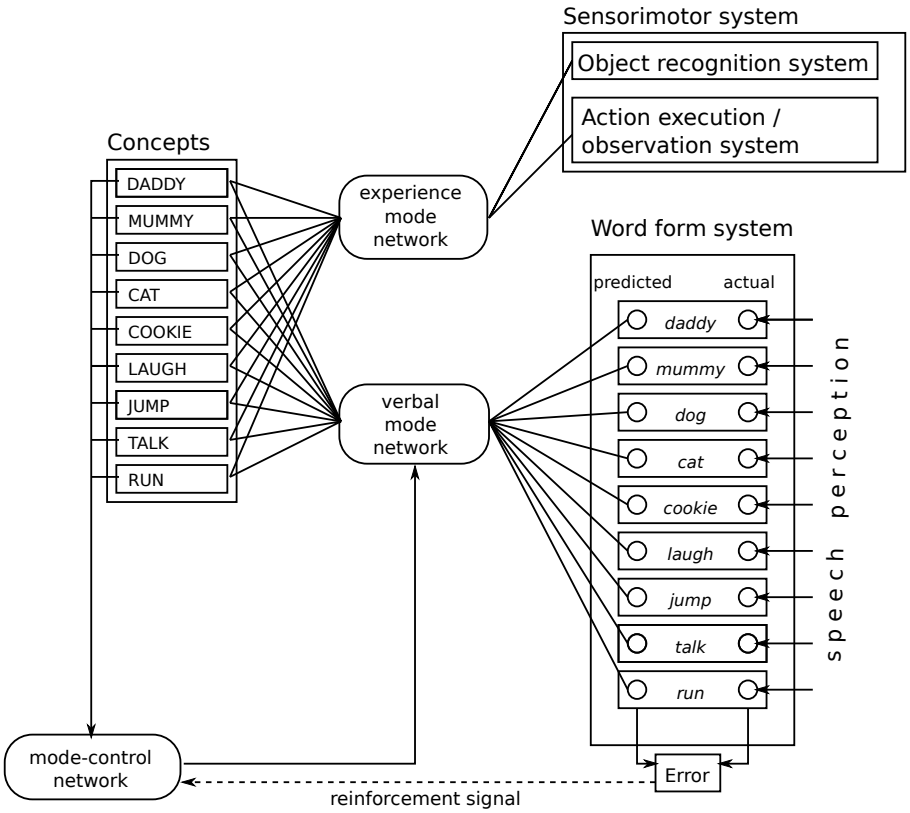

Fig. 2: Architecture of the word-learning network

agent is producing certain gestures which are of a recognisable type. In our model, the action type TALK simply represents a certain type of motor action: in itself it does not encode any of the special properties of communicative actions that were discussed in Section 2. However, the talking action is special in that it is represented twice by the infant: once as a semantic concept denoting a motor action taking place in the world, and once in a special medium holding motor actions that potentially convey meanings—namely the medium which holds word forms. We assume that this medium automatically processes phonological signals picked up by the infant. At the moment when the infant attends to a talking event (t_4), the signals encoded in this medium are constrained to be those produced by the speaker, because the semantic and phonological representations are derived from the same perceptual input. At the next moment (t_5), when the infant establishes joint attention with the speaker, we assume that the representations of the word forms produced by the speaker remain active, in some form of phonological working memory (see e.g. Baddeley *et al.*, 1998): thus in Table 1 the speaker's words *cat* and *run* are active at both t_4 and t_5.

3.2 Network architecture

The network's architecture is shown in Figure 2. In this section I will describe the two key features of the architecture.

3.2.1 Experience mode and verbal mode networks

One key feature of the network architecture is that conceptual representations (DADDY, MUMMY, RUN, JUMP etc) are linked separately to the sensorimotor system and to word forms. The **experience mode network** links concepts to the agent's perceptual and motor interfaces with the world. During normal experience, this network is engaged: through this network, when the agent perceives objects in the world, or activates motor programmes, this activates conceptual representations. However, we also envisage a separate network, the **verbal mode network**, which links the agent's conceptual representations to a specialised neural medium encoding word forms, or some other repertoire of atomic communicative gestures (for instance hand gestures in sign language). Our proposal is that the infant can *selectively* engage 'verbal mode', by turning on the connections in this network. When verbal mode is is disengaged, the stream of word forms arriving in the phonological input buffer is effectively ignored by the infant. When verbal mode is engaged, the infant can learn associations between currently active concepts and word forms—and later, when associations have been learned, the infant can use word forms to activate concepts by themselves.

3.2.2 The mode-control network

A second key feature of the architecture is a network which learns *when* to engage verbal mode, as a function of current experiences. We propose that there are certain moments when it makes particular sense for the infant to engage verbal mode. In particular, it makes sense to engage verbal mode *immediately after perceiving a talking action*. As discussed in Section 3.1, the perception of a talking action signals an imminent word-learning opportunity. After the infant observes a talking action, there is a brief period of time during which there is a particularly reliable mapping between active concepts and active word form representations: for the infant, this is an ideal moment to do some word learning, and therefore an ideal moment to engage verbal mode.

In our model, the infant learns to engage verbal mode through *reinforcement*, in roughly the same way she learns when to execute ordinary motor actions (Sutton and Barto, 1998). In regular operant learning, the agent experiences a sequence of perceptual stimuli, and is taught a specified mapping from these stimuli to motor responses by a *reward schedule* that rewards the agent whenever a particular action follows a particular perceptual stimulus. Initially, the agent executes actions from her motor repertoire at random. From time to time she executes an action which results in a reward: when this happens, she learns an association between this action and the perceptual stimulus that preceded it, so that the next time this stimulus appears, she is more likely to execute the associated action.

The task of learning when to engage verbal mode is performed by the **mode-control network**. When verbal mode is engaged, the concepts that are currently active are mapped to *predicted word forms*. These predictions are compared to the *actual word forms* active in the phonological input buffer, and an error term is generated reflecting the accuracy of the prediction. From this error term a reward signal is generated, which is used to train the mode-control network. If the error is low, the signal is a reward; if it is high, it is a punishment. (The magnitude of the punishment is relatively small compared to that of the reward.) The structure of the input data, together with this reward schedule, cause the mode-control network to learn to engage verbal mode immediately after perceiving a talking action, and in no other circumstance.

3.3 Learning in the verbal-mode and mode-control networks

At the start of training, weights in the mode-control and verbal-mode networks are initialised to random values. During training, the output of the mode-control network is annealed with noise, so that it initially engages verbal mode at random, but over time comes to deliver an output based on the activity of the concept units which provide its input.

Learning proceeds as follows. To begin with, the mode-control network engages verbal mode at random, after perceiving arbitrary episodes (e.g. CAT JUMP). In almost all cases, this leads to a small punishment, because the verbal-mode network has not yet learned any correct associations between concepts and words.

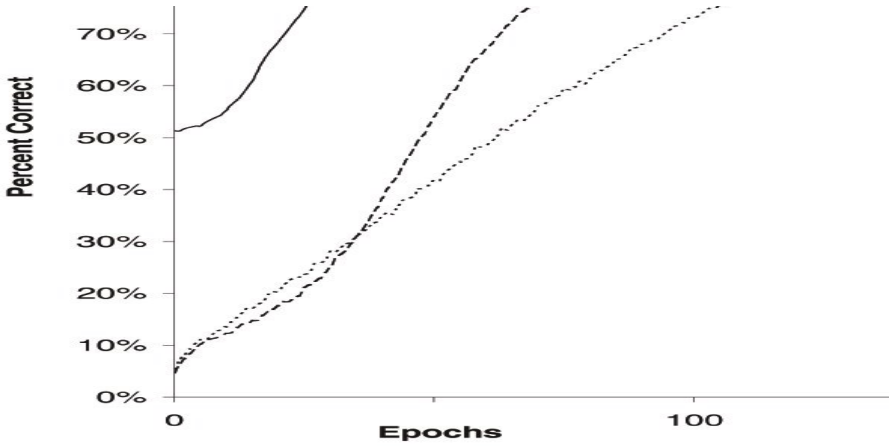

Fig. 3: Learning in the mode-control and verbal-mode networks in the model of Caza and Knott (2012)

But there is enough randomness early in learning that the network is turned on quite frequently nonetheless. When it is turned on, some cross-situational word learning occurs. This learning is very slow: the baseline mapping from concepts to words in the input data is extremely noisy, as discussed in Section 3.1; in addition, learning only happens at the moments when verbal mode is engaged, so the network is only exposed to a subset of the noisy input data. After a few correct concept-word mappings have been learned, however, there is a subtle change in the reward schedule. If verbal mode is engaged *after perception of a talk episode* (e.g. MUMMY TALK), there is an increased chance of a correct mapping between concepts and word forms at the next time point, and thus an increased chance of a large positive reward. The mode-control network thus starts to learn to engage verbal mode after perceiving talk episodes. Engaging verbal mode after perceiving other episode types continues to result in a small punishment, so the network begins to learn to engage verbal mode *only* after perceiving talk episodes. Once this happens, the verbal-mode network starts to receive more reliable training data, and it begins to learn words more efficiently. As it becomes better at predicting word meanings, talk episodes in turn become better predictors of reward, and the rate of learning increases in the mode-control network. In short, the verbal-mode and mode-control networks *bootstrap* one another.

Figure 3 charts learning in the network in two experiments. In the first experiment, learning happens in parallel in the mode-control and verbal-mode networks, as outlined above. In the second experiment, verbal mode is engaged at every time point, and the mode-control network is not used. The dotted line shows the percentage of words correctly learned at each epoch of training when the mode-control network is not used. Learning proceeds at a roughly constant rate in this condition. The dashed line shows the percentage of words correctly learned at each epoch when the verbal-mode and mode-control networks are bootstrapping one another. (The solid line charts learning in the mode-control network: the percentage of times this network engages verbal mode in response to a talk action.) In this condition learning of words is initially slower, because verbal mode is only engaged around half the time. But as the mode-control network learns to selectively enable verbal mode in response to perceived talk actions, the speed of word learning increases significantly. While word learning now only happens at a small proportion of time points, learning is still more efficient, because it is focussed on the time points which carry the best information about the mapping between concepts and words, namely communicative actions. As I will discuss in Section 4, this transition serves as an interesting model of how the development of communicative action representations enables infants to begin learning words efficiently at around the age of 12 months.

3.4 Behaviour of the network after learning

When training is complete, the mode-control network routinely engages verbal mode whenever a talk action is perceived, and the verbal-mode network reliably maps concepts onto the correct word forms. At this point we envisage some changes in the way verbal mode works. During training, verbal mode is engaged simultaneously with experience mode, so that the concepts evoked by experience are also those associated with word forms. Once training is complete, we suggest that verbal mode and experience mode become *alternatives* to one another, so that that if verbal mode is engaged, experience mode is disengaged. We also assume that the distinction between predicted and actual word forms disappears, and that the learned associations between concepts and word forms run in both directions, so that when verbal mode is engaged, word forms arriving in the phonological input buffer activate their associated concepts.

With these changes in place, there are two completely different ways of activating conceptual representations. In experience mode concepts are activated by sensorimotor experience of the world, and incoming words are ignored. In verbal mode, concepts are activated by incoming words, through associations learned by the verbal-mode network during training, and sensorimotor inputs are ignored.

4 Representations of communicative actions in the model

The neural network model outlined above offers an interesting account of the developmental processes taking place in infants at around the age of 12 months, which play a core role in Tomasello's theory of word learning. Tomasello's theory posits that infants only begin learning words in earnest when they acquire two social-pragmatic skills: an ability to establish joint attention, and an ability to identify and represent communicative actions and their underlying intentions. As discussed in Section 1, Tomasello does not consider in any detail what infants' representations of communicative actions look like, or the mechanism through which these representations help infants to learn word meanings. Caza and Knott's model makes some concrete proposals on both counts. In this section I will discuss these.

Firstly, what can we say about how communicative actions are represented in the neural network model? In one sense, clearly, communicative actions are simply represented as ordinary physical action episodes: MUMMY TALK is an episode, recognised through the same abilities to perceive objects and motor actions as are used to perceive ordinary episodes like DOG JUMP. But once the infant learns to routinely enter verbal mode after identifying a talk action, there is an interesting new element of structure to the pattern of concepts activated by a talk action. Each talk action activates a sequence of *two episode representations* in the conceptual system. The first representation is of the episode as a physical action (e.g. MUMMY TALK). This representation causes the mode-control network to engage verbal mode, so the next concepts to be activated will be those associated with the currently active word form units. These units are of course encodings of the word forms making up the utterance that has just been represented—so the concept units that become active next will reflect the *propositional content* of this utterance. In brief: when learning is complete in the network, each utterance which is perceived will be represented in a sequence of two patterns of activation in the concept units: the first representing the utterance itself, the second representing its propositional content.

Recall from Section 2 that any neural network model of communicative action representations must solve two difficult problems. First, it must allow the representation of nested propositions (e.g. 'Mummy said [the cat ran]'). Second, it must capture the elusive relation of intentionality between the outer proposition and the inner one (the fact that Mummy's saying action is 'about' the cat running). The network of Caza and Knott, when trained, offers an interesting solution to both these problems. It represents a proposition nested within another proposition very simply, as a *sequence of two propositions*. In our model the conceptual system can only represent one proposition (i.e. one episode) at a time: the propositional content of an utterance is represented entirely separately from the fact of the utterance itself, at the time point immediately after the utterance itself is represented. While the propositional content of the utterance is represented separately, the distinctive relation of 'aboutness' that links the utterance and its propositional content is also captured by the network. This relation is not represented declaratively—rather, it is cap-

tured by constraints on how episode representations can succeed one another in the trained network. These constraints are partly due to the network's own internal mechanisms: when a talk action is perceived, the network is constrained to engage verbal mode. But they are also partly due to the way the network interfaces with the external world. The same perceptual stimuli which are perceived by the action recognition system as a talk action are encoded by the speech perception system as word forms—so when verbal mode is established after a talk action is perceived, the word forms that activate concepts are constrained to be those produced by the perceived speaker.

As discussed in Section 2, representations of propositional attitudes have two distinctive characteristics: I will now consider whether the representations of communicative actions in Caza and Knott's trained network have these characteristics. Firstly, the representer of a propositional attitude is committed to the *fact* of the attitude (e.g. the fact that a given agent has a given belief), but not to its propositional content. Do the network's representations of communicative actions have this property? Consider an example communicative action 'Mummy says [the cat jumps]'. Note that the trained network evokes the episode representation MUMMY TALK directly from sensory experience, but this is not the case for the representation CAT JUMP: this representation has no relation to the network's sensory experience at all. The network does not implement any formal treatment of commitment, but if we assume a simple model, in which the network is only committed to the truth of the episodes it establishes in experience mode, then its representations of communicative actions correctly avoid commitment to the propositional content of communicative actions. Secondly, statements about propositional attitudes are intensional: their truth depends on the words that convey their propositional content. Communicative actions represented by the network certainly have this property. Their propositional content is activated in verbal mode, through associations between words and concepts. Even if a reliable observer knows that two words happen to designate the same individual in the world, there is no necessity that these words map onto the same concept in some arbitrary network: whether or not this is the case depends on the precise training that this network has received. Thus the network's representations of communicative actions seem to have many of the right properties to qualify them as representations of propositional attitudes.

To summarise: Caza and Knott's network extends Tomasello's social-pragmatic theory of infant word learning in two ways. Firstly it provides an account of the mechanism via which infants' understanding of communicative actions supports their learning of word meanings. Secondly, it provides the basis for a novel model of how communicative actions are represented, that goes some way towards capturing their distinctive properties as conveyors of propositional content.

5 Discussion

5.1 Communicative action representations as instances of semantic representations

It is interesting to compare the model of communicative action just proposed to other accounts of cognitive representation. The model shares several features with other more general proposals about the form of cognitive representations. A particularly interesting point of contact is with Ballard *et al.*'s (1997) model of **deictic routines**. Ballard *et al.* argue that the cognitive representations active in an agent at any given moment in time cannot be interpreted in isolation, as they often 'implicitly refer' to the agent's momentary deployment of perceptual and motor resources to his immediate environment. For instance, neural assemblies in the visual object categorisation system in inferior temporal cortex predominantly represent the stimulus at the current fixation point, or at the current locus of covert attention (Zhang *et al.*, 2011): in order to interpret these representations, we need to know what attentional action resulted in their activation. For Ballard *et al.*, cognitive representations are often given meaning by their position in a sequentially structured routine of cognitive operations—for instance a routine in which an agent attends to an object, then computes its grasp affordances, and then reaches for it. Each cognitive operation in the routine generates transitory cognitive representations—and often transitory motor states—which provide the conditions under which the next cognitive operation can be executed. The sequence of episode representations which collectively encode a communicative action in the current proposal can usefully be thought of as a deictic

routine of this kind. The episode representation MUMMY TALK, when active, enables execution of a cognitive operation which changes the way the cognitive system is deployed to the world; it is impossible to interpret the conceptual representations which are activated next in the network without making reference to this cognitive operation, and to the episode representation that triggered it, even though this is no longer active.

Ballard *et al.*'s model of deictic routines is extended by Knott (2012), who proposes that an agent perceives *all* concrete episodes through sensorimotor routines with canonical sequential structure, and represents all such episodes as prepared sensorimotor routines. (A computational model is provided by Takac and Knott, 2013.) In this account, semantic representations of concrete episodes are uniformly structured as sequences, and many principles of syntax are seen as deriving from constraints in the way sensorimotor operations can succeed one another sequentially. The idea that an utterance and its propositional content are represented at two distinct moments in time might seem unusual for theorists accustomed to thinking of semantic representations as static patterns of activity. But in the light of accounts like those of Ballard *et al.* (1997) and Knott (2012), communicative action representations are not exceptional in having sequential structure, but actually conform to the general pattern for semantic representations of episodes.

5.2 Towards a general model of propositional attitude representations

As noted in Section 1, infants do not develop sophisticated representations of propositional attitudes immediately: the early representations of communicative actions posited by Tomasello as having a role in word learning are presumably simple precursors to the representations that develop later in life. Indeed the model of propositional content representations put forward in the current paper is just a model of communicative actions: it is tailored to these actions, and does not attempt to provide an account of other types of mental state, such as belief or desire. However, since it provides an interesting account of the propositional content of utterances, it is interesting to consider whether it can provide the basis for a more general model of mental states. I will conclude with a few suggestions about this prospect.

Firstly, the idea that a propositional atttitude representation in general takes the form of a sequence of two simple proposition representations, separated by some mode-changing operation, is an interesting one. In a more general model of propositional attitudes, the suggestion would be that the first proposition in the sequence would encode the agent adopting the attitude, together with a special action or operation denoting the attitude in question: BELIEVE, WANT, REMEMBER and so on. In each case, activation of this special representation would trigger a change in cognitive mode, resulting in the activation of a new proposition in the conceptual system. Different attitudes would presumably trigger different cognitive modes: perhaps WANT would configure the conceptual system to represent an intention of the agent rather than the results of sensorimotor experience, for instance, while REMEMBER would configure the conceptual system to receive episode representations from long-term memory. I assume there would have to be hard-wired circuitry in the network to support each of the distinct propositional attitudes. In the case of REMEMBER, there is actually good evidence that memory retrieval involves the establishment of a special cognitive mode, implemented by specialised neural circuits (see e.g. Buckner and Wheeler, 2001; Buckner *et al.*, 2008).

Secondly, a more complete model of attitudes would need to allow the agent to represent his own attitudes as well as those of other agents. In the account of communicative action representations given in this paper, the communicative actions are always those of an observed agent. But the infant herself can execute communicative actions: how would the current model need to be extended to accommodate an account of producing utterances as well as of producing them? In a generation scenario, we have to imagine that the infant uses representations in the conceptual system to plan her own actions—and that these actions can include the action of entering verbal mode for the purposes of speaking. Here, presumably, we must envisage a process which actively removes the infant's planned action of entering verbal mode ('ME TALK') as soon as it is achieved, and replaces it with a representation of the content to be produced. Where this content comes from is an interesting question. We must also envisage that conceptual representations activate word forms in this mode, and that word forms result in overt speech sounds. In the case of a more abstract propositional attitude like wanting or remembering, it is likely that the operations evoking one's own attitudes are somewhat simpler than those evoking those of another agent: for instance, while evoking one's

own desire in the concept units might just involve activating an interface to these units from one's own planning system, evoking the desire of another agent is likely to require perceptual inference mechanisms, and perhaps also specialised mechanisms for storing mental states of other agents.

A final interesting question concerns whether a mode-changing model of mental states supports arbitrarily deep nesting of mental states. For instance, imagine an agent who enters verbal mode (ME TALK) and then evokes the representation of another mode-changing operation—for instance ME WANT. Given the agent is in verbal mode, the operation WANT is presumably mapped to a word (*want*)—but is the WANT operation also executed in this mode?

All in all, the mode-changing model of communicative action representations seems to provide quite an interesting platform for the development of a more elaborate account of the representation of propositional mental attitudes. But it certainly raise more questions than it answers.

References

Averbeck, B., Chafee, M., Crowe, D., and Georgopoulos, A. (2002). Parallel processing of serial movements in prefrontal cortex. *PNAS*, **99**(20), 13172–13177.

Baddeley, A., Gathercole, S., and Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, **105**(1), 158–173.

Baldwin, D., Markman, E., Bill, B., Desjardins, R., Irwin, J., and Tidball, G. (1996). Infants' reliance on a social criterion for establishing word-object relations. *Child Development*, **67**, 3135–3153.

Ballard, D., Hayhoe, M., Pook, P., and Rao, R. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, **20**(4), 723–767.

Brentano, F. (1874). *Psychologie vom empirischen Standpunkte*. Duncker & Humblot, Leipzig.

Buckner, R. and Wheeler, M. (2001). The cognitive neuroscience of remembering. *Nature Reviews Neuroscience*, **2**, 624–634.

Buckner, R., Andrews-Hanna, J., and Schacter, D. (2008). The brain's default network: Anatomy, function and relevance to disease. *Annals of the New York Academy of Sciences*, **1124**, 1–38.

Butterworth, G. and Jarrett, N. (1991). What minds have in common is space: Spatial mechanisms for perspective taking in infancy. *British Journal of Developmental Psychology*, **9**, 55–72.

Caza, G. and Knott, A. (2012). Pragmatic bootstrapping: A neural network model of vocabulary acquisition. *Language Learning and Development*, **8**, 1–23.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, **26**, 609–651.

Dennett, D., editor (1989). *The Intentional Stance*. MIT Press/Bradford Books, Cambridge, MA.

Diesendruck, G., Markson, L., Akhtar, N., and Reudor, A. (2004). Two-year-olds sensitivity to speakers intent: an alternative account of Samuelson and Smith. *Developmental Science*, **7**(1), 33–41.

Gärdenfors, P. (2004). *Conceptual Spaces*. MIT Press, Cambridge, MA.

Jacquette, D., editor (2004). *Brentanos Concept of Intentionality*. Cambridge University Press.

Knott, A. (2012). *Sensorimotor Cognition and Natural Language Syntax*. MIT Press, Cambridge, MA.

Miller, E. and Cohen, J. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, **24**, 167–202.

Montague, R. (1974). The proper treatment of quantification in ordinary English. In R. Thomason, editor, *Formal philosophy: Selected Papers of Richard Montague*, pages 247–270.

Plate, T. (2003). *Holographic Reduced Representations*. CSLI Lecture Notes Number 150. CSLI Publications, Stanford, CA.

Saffran, J., Aslin, R., and Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, **274**, 1926–1928.

Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, **61**(1–2), 39–91.

Sutton, R. and Barto, A. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Takac, M. and Knott (2013). A neural network model of working memory for episodes. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, Berlin.

Takac, M., Benuskova, L., and Knott, A. (2012). Mapping sensorimotor sequences to word sequences: A connectionist model of language acquisition and sentence generation. *Cognition*, **125**, 288–308.

Tomasello, M. (1995). Joint attention as social cognition. In C. Moore and P. Dunham, editors, *Joint attention: Its origins and role in development*, pages 103–130. Erlbaum, Hillsdale, NJ.

Tomasello, M. (2000). The social-pragmatic theory of word learning. *Pragmatics*, **10**(4), 401–413.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Tomasello, M. and Herrmann, E. (2010). Ape and human cognition: What's the difference? *Current Directions in Psychological Science*, **19**(1), 3–8.

van der Velde, F. and de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, **29**, 37–108.

Yu, C. and Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, **70**(13-15), 2149–2165.

Zhang, Y., Meyers, E., Bichot, N., Serre, T., Poggio, T., and Desimone, R. (2011). Object decoding with attention in inferior temporal cortex. *Proceedings of the National Academy of Sciences of the USA*, **108**(21), 8850–8855.