Title

Modelling the co-development of word learning and perspective-taking

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 38(0)

Authors

Woensdregt, Marieke
Kirby, Simon
Cummins, Chris
et al.

Publication Date

2016

Peer reviewed

Modelling the co-development of word learning and perspective-taking

Marieke Woensdregt (m.s.woensdregt@sms.ed.ac.uk)[a]

Simon Kirby [a], Chris Cummins [b], Kenny Smith [a]

[a] Centre for Language Evolution, School of Philosophy, Psychology & Language Sciences, University of Edinburgh, 3 Charles Street, Edinburgh, EH8 9AD, United Kingdom

[b] Department of Linguistics and English Language, School of Philosophy, Psychology & Language Sciences, University of Edinburgh, 3 Charles Street, Edinburgh, EH8 9AD, United Kingdom

Abstract

Word learning involves mapping observable words to unobservable speaker intentions. The ability to infer referential intentions in turn has been shown to depend in part on access to language. Thus, word learning and intention-reading co-develop. To explore this interaction, we present an agent-based model in which an individual simultaneously learns a lexicon and learns about the speaker's perspective, given a shared context and the speaker's utterances, by performing Bayesian inference. Simulations with this model show that (i) lexicon-learning and perspective-learning are strongly interdependent: learning one is impossible without some knowledge of the other, (ii) lexicon- and perspective-learning can bootstrap each other, resulting in successful inference of both even when the learner starts with no knowledge of the lexicon and unhelpful assumptions about the minds of others, and (iii) receiving initial input from a 'helpful' speaker (who adopts the learner's perspective on the world) paves the way for later learning from speakers with perspectives which diverge from the learner's. This approach represents a first attempt to model the hypothesis that language and mindreading co-develop, and a first exploration of the implications for theories of word learning and mindreading development.

Keywords: word learning; perspective-taking; computational model; Bayesian inference;

Introduction

Word learning is a special case of associative learning, as one has to learn a mapping between something observable — a speaker's utterance — and something unobservable — the speaker's meaning. Word learning therefore requires inferring the speaker's referential intention (Waxman & Gelman, 2009), which in turn requires theory of mind (ToM). Learning about words and learning about minds are thus necessarily connected: language learners need to figure out not just the stable mappings between words and concepts (the lexicon) but also a way of inferring speaker intention, which is variable over time and depends on context and speaker-specific features.

In this paper we present evidence that language and ToM development go hand in hand, and we explore the implications of such a co-development by means of an agent-based model. As a test case we look specifically at the interaction between word learning and perspective-taking. Although perspective-taking cannot be equated with ToM, it is an instantiation of the latter and forms a good starting point for formalising the relation between language learning and ToM development.

Learning about words and minds

There is persuasive evidence consistent with the idea that learning about words and learning about minds are interrelated. In a study comparing children with autism (AD) to typically-developing (TD) children, Parish-Morris et al. (2007) showed that although 5-year-old AD children have some ability to use social cues (pointing and eye gaze) to direct their attention in word learning, they perform at chance when learning new words required inferring the speaker's intention, unlike language- and mental-age-matched TD children.

The reverse phenomenon has also been observed, namely that the development of ToM depends in part on having access to language. Deaf children of hearing parents, who lack consistent linguistic input, were shown to have delayed ToM development relative to deaf children of deaf parents, who receive sign language input from birth (Schick et al., 2007). Similarly, a study with TD children showed that simply training children on the use of mental state verbs with sentential complements accelerated their false belief understanding (Lohmann & Tomasello, 2003).

Thirdly, in a study comparing different age-groups of signers of the recently emerged Nicaraguan Sign Language, Pyers and Senghas (2009) showed that the bootstrap effect of language on ToM development continues on into adulthood. Pyers & Senghas found that the first cohort of signers (mean age 27), whose language had very limited mental state vocabulary, were worse at understanding false belief than the second cohort (mean age 17) who had more words for mental states. Moreover, a follow-up study two years later revealed that the first-cohort signers had improved in their false belief understanding and that this either followed or co-occurred with, but never preceded, an expansion of mental state vocabulary.

Finally, recent evidence suggests that mindreading and language skills co-develop. Brooks and Meltzoff (2015) showed that gaze-following in 10.5-month-old infants predicted their production of mental state terms at 2.5-years-old, and that these mental state terms in turn predicted the extent of their false belief understanding at 4.5-years-old, even though gaze-following did not directly predict false belief understanding. Thus, this shows evidence of an indirect relation between early sensitivity to social cues and later mindreading ability, mediated by language.

Models of word learning and perspective-taking

Words are used in complex environments, and each word could label any part of that complex environment. Worse, words can label objects and events which are not currently perceivable to the hearer and/or the speaker (e.g. events which are spatially or temporally distant from the time of speaking). Learners therefore face *referential uncertainty*: every time a word is used, there may be many meanings which a learner could infer as the word's intended meaning.

Computational models of word learning have explored several potential solutions to the problem of referential uncertainty, which could be roughly divided up into three kinds: (i) solutions using learning biases, (ii) social cues solutions, and (iii) intention-reading solutions.

Brute force statistical learning of word-referent associations is impossible if referential uncertainty is unbounded: if all logically possible meanings are equally-plausible candidates for the meaning of any word on any use, then no learner can learn the meaning of any word (an observation commonly attributed to Quine, 1960, in his work on radical translation). Experimental and observational studies have demonstrated that word learners use a number of heuristics to reduce referential uncertainty: learners assume that words refer to whole objects (Macnamara, 1972); they use argument structure and syntactic context to constrain the meaning of new words (Gillette et al., 1999); and they use knowledge of the meaning of other words to constrain hypotheses about the meaning of a new word, for example by assuming that words have mutually exclusive meanings (Markman & Wachtel, 1988). Models of cross-situational statistical learning suggest that brute-force cross-situational learning of large lexicons is possible under surprisingly high levels of referential uncertainty (Blythe, Smith, & Smith, 2010) or even under infinite referential uncertainty if word learners can use their heuristics to rank candidate meanings in terms of their plausibility (Blythe, Smith, & Smith, submitted).

In addition to exploiting linguistic context or their knowledge of likely word meanings, learners can use social cues, which are potentially highly informative in guiding word learning (see Paulus and Fikkert (2014) for eye-gaze and pointing and Yu and Smith (2012) for joint attention). Yu and Ballard (2007) formalised these mechanisms in a model of word learning that integrates the use of statistical regularities and social cues. They provided an associative model with information about which words and objects in a discourse stream were highlighted by social cues (prosody and joint attention), and simply increased the association weight of those items. They then tested the model on how well it could learn a lexicon from transcriptions of two videos of mother-child interactions from the CHILDES corpus. This 'hybrid' model was compared to a 'bare' statistical learning model, and statistical learners who exploited prosody or joint attention, but not both. Best performance was obtained with the model that integrated both types of social cue.

However, there is more to social interaction than just cues that direct attention. The ability to recognise that speech can convey unobservable communicative intentions comes online before children start talking (Vouloumanos, Onishi, & Pogue, 2012) and is used to guide their word learning (Parish-Morris et al., 2007). To formalise the role that inferring speaker intentions plays in word learning, Frank, Goodman, and Tenenbaum (2009) designed a Bayesian model that simultaneously infers word-object mappings and speaker intentions, and tested this model on the same CHILDES videos used by Yu and Ballard (2007). Rather than re-weighting items based on social cues, Frank et al. assume that learners posit an extra unobserved variable mediating between the objects in the physical context and the words that the speaker produces: the speaker's referential intention. The learner then evaluates all possible lexicon hypotheses based on the prior probability of that lexicon and the likelihood of a word given that lexicon and the speaker's referential intention, where the intention hypotheses that are considered by the learner are simply all possible subsets of the objects present in the context, including an 'empty' intention.

This model has two advantages over other associative learning models. Firstly, it can represent the possibility of 'empty intentions', where the word does not refer to any physically present object. Secondly, it can distinguish between words that can be used referentially and words that are used exclusively 'non-referentially', where non-referential (e.g. function) words are simply left out of the lexicon. Frank et al. (2009) show that this model outperforms several alternative statistical learning models (including Yu and Ballard's), both when tested on the lexicon they learned and on the referential intentions they inferred (given their lexicon).

Although these various models constitute important first steps towards modelling the role of intention-reading in word learning, they treat the ability to utilise social cues or infer intentions as a given and fixed capacity, present from the start of word learning. In real-world learning, the ability to learn words and the ability to infer mental states (including referential intentions) improve as a child grows older. As described above, this improvement is partly accounted for by a co-development of language and intention-reading. Below, we will describe a model that takes these considerations into account: rather than modelling word learning as a combination of associative learning with social cues or uninformed intention representations, we provide a model which allows for the co-development of word learning and perspective-taking.

The current model: Integrating development of word learning and perspective-taking

Model description

We model referential intentions as a result of the interaction between a set of attributes of the world — the context — and an attribute of the speaker — the perspective. This perspective can be interpreted in a literal sense, where objects that are spatially or temporally closer to the agent are more salient (see figure 1). Importantly however, it can equally serve as a

model for the sum of an agent's knowledge and beliefs about the world that determine what topics of conversation will be most salient to them in a given situation. The latter is the sort of perspective that requires full-blown ToM to be inferred. All that matters here is that there is a function that maps from the attributes of the world to an agent's saliency distribution over potential topics, and that the agent has a hidden variable (their perspective) that is a parameter in this function.

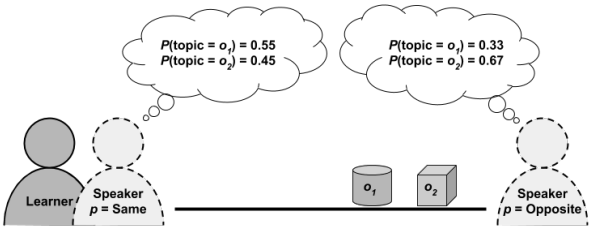

Figure 1: Diagram of how speaker perspective gives rise to referential intention. The speaker on the left is only slightly more likely to choose object 1 (o_1) over object 2 (o_2) as a referent, since both are approximately equidistant. The speaker on the right however is twice as likely to choose object 2 than object 1 since o_2 is twice as close as o_1. The learner has their own perspective on the world and learns with an egocentric bias; assuming that the speaker shares their perspective.

The variables that the learner can observe are the context and the speaker's utterance (see figure 2). The variables that are unobservable are the speaker's perspective, the speaker's referential intention, and the speaker's lexicon. The learner's task is to infer the speaker's perspective and the lexicon based on the same data: the speaker's word use in different contexts.

This model differs from that outlined in Frank et al. (2009) in that it posits an extra unobservable variable: the speaker's perspective, which together with the context determines the speaker's referential intention. Given a specific hypothesis about the speaker's perspective, the learner can compute a prediction of how likely it is that the speaker will refer to a given object in a given context (i.e. how salient the object is for the speaker). Subsequently, given a specific hypothesis about what the lexicon is, the learner can turn this prediction about likely referents into a prediction of likely utterances.

We assume, unlike the models of word learning described above, that all objects that are part of the world are possible referents in every learning context: thus, simple associative cross-situational learning alone will not be able to solve the problem of referential ambiguity. The learner can get around this problem by inferring the speaker's perspective: a hypothesis about this perspective is the only information available that can render the probability distribution over possible referents non-uniform, which in turn allows the learner to infer the most likely word-object mappings. Specifically, this is achieved by incrementing the posterior belief in lexicon hypotheses in proportion to how salient the object that is associated to the utterance in that lexicon is for the speaker, given the perspective hypothesis under consideration.

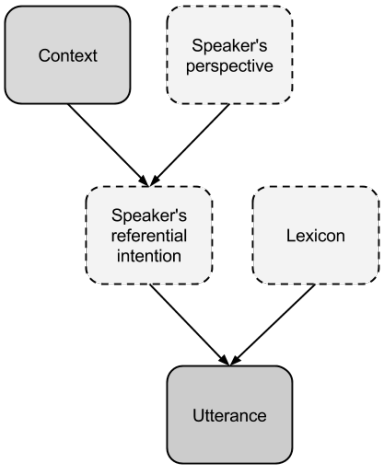

Figure 2: Diagram of the current model. Variables in dark grey and solid lines are observable to the learner, variables in light grey and dashed lines are unobservable. The learner's task is to infer the speaker's perspective and the lexicon based on observations of the speaker's word use in context.

Note that in this model no lexicon hypothesis can be evaluated without simultaneously positing a perspective hypothesis, and vice versa. Thus the complete hypothesis space for the learner consists of all possible combinations of lexicon hypothesis and perspective hypothesis (with the potential of representing different perspectives, and indeed different lexicons, for different speakers). Learning in this model is implemented as Bayesian inference according to the definitions described below.

Posterior The task of the learner in this model[1] is to find the lexicon hypothesis l and perspective hypothesis p that have the highest posterior probability given data D, as shown in equation 1.

$$P(l, p \mid D) \propto P(D \mid l, p)P(l, p) \quad (1)$$

The perspective hypothesis p represents a single parameter in an intention function that maps from the context to the speaker's referential intention. This referential intention is based on the saliency of the objects in the context, which is defined as the inverse of the distance between the speaker's perspective and the object's location (see figure 1). These saliency values are then normalized over all objects in the context, rendering a probability distribution over all objects

[1]We describe the model in terms of a learner who assumes that a single lexicon and a single speaker perspective will account for all of their data: the same model can straightforwardly be extended to model a learner who allows that different speakers might have different lexicons and different perspectives; later we present results for a learner who entertains multi-perspective hypotheses.

that determines how likely the speaker is to choose them as intended referent. This distribution is then used to generate the speakers referential intention.

The learner does not need to infer the intention function itself, only the perspective parameter. This model thus simulates the situation where the learner is 'born' with the ability to represent mental states, but has to learn how to make predictions about the *content* of another agent's mind on the basis of the context. More specifically, the learner is born with a model of how a context will give rise to a speaker's referential intention, given the speaker's perspective, but has to infer from data exactly what the speaker's perspective is.

Likelihood The likelihood of a set of data D is:

$$P(D \mid l, p) = \prod_{d \in D} P(w_d \mid l, p, c_d) \tag{2}$$

where each data point d consists of a context c and a word w that was uttered by the speaker in that particular context. The likelihood of a single word w_d is defined in equation 3.

$$P(w_d \mid l, p, c_d) = \sum_{o \in c_d} P(i_o \mid p, c_d) P(w_d \mid i_o, l) \tag{3}$$

where o stands for object and i_o for the probability that object o will be the intended referent given the perspective hypothesis p.

Thus, the probability of a particular word being uttered in a particular context is equal to the product of the probability of that word being uttered for a given object (according to lexicon hypothesis l) and the probability of that object being the intended referent (according to perspective hypothesis p), summed over all objects.

In the simulations described below all lexicon hypotheses that are considered consist simply of discrete binary mappings between words and objects — in other words, if there are two objects and two possible words, there are nine possible lexicons (object 1 maps to word a or word b or either, and object 2 independently maps to word a or word b or either). Thus, the probability of a given word being uttered for a given intended referent is given by equation 4

$$P(w_d \mid i_o, l) = \begin{cases} \dfrac{1}{|w_o|} & \text{if } w_d \text{ maps to } o \text{ in } l \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $|w_o|$ is the number of words that map to object o in lexicon l.

Prior For all simulations described below, we assume that learners have a neutral prior over lexicons and an egocentric prior over perspectives. That is, the learner starts out assuming that all lexicons are equally probable, and that other agents share their own perspective. Over all combinations of lexicon and perspective prior, the 'composite prior' is simply the product of the two, as shown in equation 5.

$$P(l, p) = P(l)P(p) \tag{5}$$

Simulation results

All simulation results described in this section show what happens in the very simple case where the learner gets input from one or two speakers in a world where there exist only two possible referents (objects) and two words. The set of lexicon hypotheses consists of all functionally distinct ways of mapping two words onto two objects (nine lexicons in total, as described above). The set of perspective hypotheses consists of the two most extreme possibilities: either the speaker's perspective is the same as the learner's own perspective, or it is exactly the opposite. The learner's hypothesis space consists of all possible combinations of lexicon hypothesis and perspective hypothesis.

In a first set of four simulations we explore the influence that perspective-learning and lexicon-learning have on each other. We compare three different cases: (i) the target lexicon is unambiguous (i.e. each object is associated with a distinct word) but the learner is unable to learn that speakers might have a perspective that is different from their own (which we achieve by setting the prior probability of the 'other' perspective to 0); (ii) the learner is initially egocentric yet can learn that speakers can have a perspective that differs from their own (which we achieve by setting the prior probability of the 'other' perspective to 0.1, and the 'own' perspective to 0.9), but the target lexicon is partly ambiguous (e.g. object 1 maps to both word a and word b, while object 2 maps only to word b); (iii) same as in (ii) but with a fully ambiguous lexicon (both objects map to both words); and (iv) the learner can learn that the speakers can have a different perspective from the learner, as in (ii) and (iii), and the target lexicon is unambiguous, as in (i).

Situation (iv) thus simulates a typically-developing child in a normal language environment (under the assumption that words are effectively unambiguous in their linguistic context: Piantadosi, Tily, & Gibson, 2012) — we refer to this as the Typical condition. Situation (i) simulates a word learner with a strongly impaired (or absent) ToM — we refer to this as the No ToM condition. Situation (ii), which we refer to as the Partly Ambiguous Lexicon condition, simulates a typically-developing word learner in an environment where the target lexicon is such that a speaker's utterances are rather uninformative about their referential intentions. This scenario could be compared to the case of deaf children who grow up with hearing parents (i.e. without sign language), since although such parents do exhibit communicative behaviour that could reveal something about their communicative intentions, this is less explicit and more ambiguous than linguistic data (Schick et al., 2007). Finally, situation (iii), which we refer to as the Uninformative Lexicon condition, is an extreme form of this case, where there is a complete absence of behaviour that is informative about the speaker's intentions. This is a case analogous to one in which a reliable language has yet to emerge in a population.

Figure 3 shows the learning results for the four different situations described above. Several interesting learning dynam-

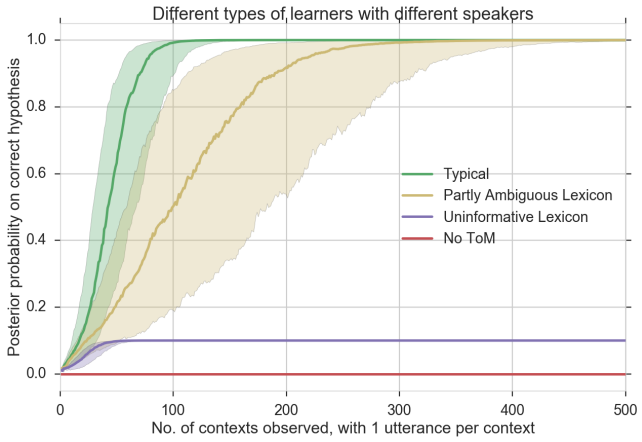

Figure 3: Learning curves for different learners in different learning situations. Learning is measured as the amount of posterior probability assigned to the correct hypothesis, where 1.0 is ceiling. Lines show median over 1000 runs, shaded area shows first and third quartile.

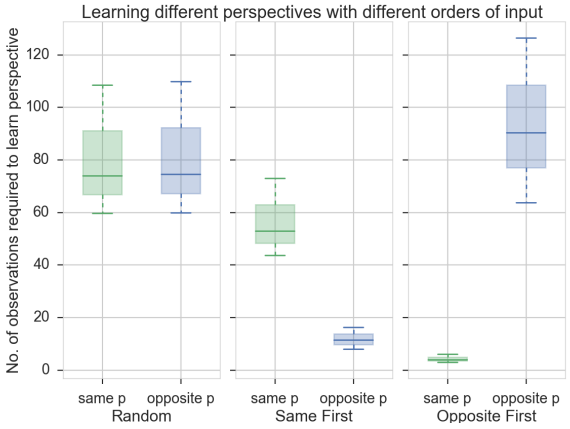

Figure 4: Amount of observations required for learning different speaker perspectives under different input conditions: Random, Same First and Opposite First. Successful learning is defined as > 0.99 posterior probability on correct hypothesis, and the lexicon is learned fully in all conditions before the learner enters the second input phase. Boxes show median, first and third quartile over 100 runs.

ics are apparent. Firstly, inferring the correct lexicon is impossible when the learner cannot infer the correct perspective of the speaker that they get input from (No ToM condition). Secondly, inferring the speaker's perspective becomes more difficult when there is a less direct mapping between their referential intention and their behaviour (Partly Ambiguous Lexicon condition). However, learning in this case is still eventually successful: the ability to infer perspective gives a way into learning the lexicon, thus making it easier to deal with lexical ambiguity. Thirdly, inferring the speaker's perspective becomes impossible when the speaker's behaviour gives no information at all about their intention (Uninformative Lexicon condition). Finally, learning happens most quickly and successfully when the learner is both able to represent different perspectives and the speakers' lexicon is unambiguous (Typical condition).[2]

In a second set of three simulations we present the effect of order of input on lexicon and perspective learning. These simulations are similar to the ones described above, except that the learner receives input from two different speakers who have two different perspectives: one speaker shares the learner's perspective, the other has the opposite perspective. We present the learning results in three different situations: (i) the speaker is randomly picked on each trial, but both speakers get to speak for an equal number of contexts (Random condition); (ii) the learner receives the first half of their input from the speaker that shares their perspective, and the second half from the 'opposite perspective' speaker (Same First condition); and (iii) the learner receives the first half of input from the opposite perspective speaker and the second half from the same perspective speaker (Opposite First condition).

[2]These results are qualitatively similar for learning about larger lexicons of e.g. 3x3 and 4x4 objects and words.

As figure 4 shows, the difference in the amount of observations that is required to learn the opposite perspective is bigger between the two conditions (Same First vs. Opposite First) than the difference in the amount of observations required to learn the same perspective in the two conditions. This means that receiving input from a 'helpful' speaker (a speaker who shares the learner's perspective) first paves the way for later learning about perspectives that are different from the learner's own.[2]

The mediating factor that gives rise to this effect is the lexicon, since the only thing that is different about the learner after having learned the same perspective first is their knowledge of the lexicon. (Which, in all simulations shown in figure 4, is fully learned before the learner enters the second input phase.) This effect relies on the lexicon being shared among members of the population. Language as a convention is what allows the learner to bootstrap knowledge of other's perspectives based on starting with a familiar speaker first.

Discussion

We presented an agent-based model that simulates the co-development of word-learning and perspective-taking through Bayesian inference. This model is different from previous models of word learning in that all objects that are part of the world are considered as potential referents at each learning episode, rendering brute-force cross-situational learning impossible. However, the learner can overcome this referential uncertainty by learning about the speaker's perspective. Both the lexicon and the perspective are learned using the same data (the speaker's word use in context).

This model gives rise to several potentially interesting co-development dynamics. Firstly, lexicon-learning and

perspective-learning are strongly interdependent: learning the one cannot happen without some knowledge of the other. Secondly, lexicon- and perspective-learning can bootstrap each other, resulting in successful inference of both variables even when the learner starts out with an inappropriate egocentric bias and no knowledge of the lexicon whatsoever. Finally, the results show that receiving input from a helpful speaker first paves the way for later learning from speakers whose perspective differs from the learner's — the helpful speaker provides data which facilitates learning of the lexicon, which then facilitates learning of the perspective of other less well-aligned speakers (on the assumption that the lexicon is shared among speakers).

To our knowledge, this is the first computational model that does not simply incorporate pragmatic inference as a tool to infer word meaning (Frank et al., 2009), but rather incorporates pragmatic inference as a developing skill that interacts bi-directionally with word learning. Thus, this model is a first step towards formalising the hypothesis that language and mindreading co-develop.

The simulation results of this model described here replicate several empirical findings. Firstly, it mirrors the finding that word-learning depends partly on the inference of mental states (Parish-Morris et al., 2007). Secondly, it mirrors the finding that the development of mindreading depends partly on vocabulary development (Lohmann & Tomasello, 2003; Pyers & Senghas, 2009; Schick et al., 2007). Finally, it generates the developmental prediction that learning from a helpful speaker who shares the child's perspective early on in life will aid vocabulary development, and that this in turn will help the child to learn about alternative perspectives later on.

Several aspects of this model are however very simplistic. Firstly, the learner in this model is 'born' with a ToM. Rather than having to infer the full function that maps from a context to a speaker's referential intention, the learner only has to infer the speaker's perspective. In real life children have to develop not only the ability to infer the content of mental states, but also the underlying ability to represent *that* the content of others' minds is different from that of their own. Future work with this model could incorporate a more realistic model of ToM development that could mimic more closely the stages of ToM development we see in real children.

Secondly, the relation between observations of words and learning about perspectives is very direct. Each word-object mapping that is learned helps with inferring perspective because it allows the learner to evaluate their prediction of referential intent based on their perspective hypothesis. It is not yet clear what the role of language learning is in driving the development of ToM in the real world — this might have to do with access to discourse, explanations or representations of mental states (see e.g. Lohmann & Tomasello, 2003; Pyers & Senghas, 2009; Schick et al., 2007).

Despite these simplifications, this model forms a first exploration into the co-development dynamics of language and ToM.

References

Blythe, R. A., Smith, A. D. M., & Smith, K. (submitted). Word learning under infinite uncertainty.

Blythe, R. A., Smith, K., & Smith, A. D. M. (2010). Learning times for large lexicons through cross-situational learning. *Cognitive Science*, *34*, 620–42.

Brooks, R., & Meltzoff, A. N. (2015). Connecting the dots from infancy to childhood: A longitudinal study connecting gaze following, language, and explicit theory of mind. *Journal of Experimental Child Psychology*, *130*, 67–78.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–85.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*, 135–76.

Lohmann, H., & Tomasello, M. (2003). The Role of Language in the Development of False Belief Understanding: A Training Study. *Child Development*, *74*(4), 1130–44.

Macnamara, J. (1972). The cognitive basis of language learning in infants. *Psychological Review*, *79*, 1–13.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meaning of words. *Cognitive Psychology*, *20*, 121–57.

Parish-Morris, J., Hennon, E. a., Hirsh-Pasek, K., Golinkoff, R. M., & Tager-Flusberg, H. (2007). Children with autism illuminate the role of social intention in word learning. *Child Development*, *78*(4), 1265–87.

Paulus, M., & Fikkert, P. (2014). Conflicting Social Cues: Fourteen- and 24-Month-Old Infants' Reliance on Gaze and Pointing Cues in Word Learning. *Journal of Cognition and Development*, *15*(1), 43–59.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*, 280-91.

Pyers, J. E., & Senghas, A. (2009). Language Promotes False-Belief Understanding: Evidence From Learners of a New Sign Language. *Psychological Science*, *20*, 805–12.

Quine, W. V. O. (1960). *Word and Object*.

Schick, B., de Villiers, P., de Villiers, J., & Hoffmeister, R. (2007). Language and theory of mind: a study of deaf children. *Child Development*, *78*(2), 376–96.

Vouloumanos, A., Onishi, K. H., & Pogue, A. (2012). Twelve-month-old infants recognize that speech can communicate unobservable intentions. *PNAS*, *109*, 12933–7.

Waxman, S. R., & Gelman, S. a. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, *13*(6), 258–63.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*, 2149–65.

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–62.