Multiple variable cues in the environment promote accurate and robust word learning

Padraic Monaghan (p.monaghan@lancaster.ac.uk)
James Brand (j.brand@lancaster.ac.uk)
Rebecca L.A. Frost (r.frost1@lancaster.ac.uk)
Department of Psychology, Lancaster University, Lancaster, LA1 4YF, UK

Gemma Taylor (g.taylor4@salford.ac.uk)
Department of Psychology, Salford University, Salford M5 4WT, UK

Abstract

Learning how words refer to aspects of the environment is a complex task, but one that is supported by numerous cues within the environment which constrain the possibilities for matching words to their intended referents. In this paper we tested the predictions of a computational model of multiple cue integration for word learning, that predicted variation in the presence of cues provides an optimal learning situation. In a cross-situational learning task with adult participants, we varied the reliability of presence of distributional, prosodic, and gestural cues. We found that the best learning occurred when cues were often present, but not always. The effect of variability increased the salience of individual cues for the learner, but resulted in robust learning that was not vulnerable to individual cues' presence or absence. Thus, variability of multiple cues in the language-learning environment provided the optimal circumstances for word learning.

Keywords: word learning; multiple cues; strategies; gesture; prosody; cross-situational learning.

Cues for word learning

Learning how words relate to objects, actions, properties, or relations in the world is a complex task. One of the key difficulties is that word learning provides few explicit constraints on which words can relate to particular aspects of the environment (Quine, 1960). Thus, in acquiring vocabulary, children must resolve a many-to-many (possibly even an infinite-to-infinite) mapping between words in utterances and elements of the environment around them. So how do children solve this task?

There are two proposals for how learning word-referent mappings can be constrained. The first is that children have internal biases that apply to language learning situations that limit possible referents to words (Markman, 1994). For instance, mutual exclusivity refers to the assumption in word learning situations that each referent has only one name, leading children to pair an unnamed object with a novel word (Markman & Wachtel, 1998). In terms of limiting referents, children seem to be biased to linking a word with a whole object rather than a part of an object (Macnamara, 1982), and may more readily form categories of objects with similar shape which are referred to by the same word (Baldwin, 1992).

The alternative proposal for resolving the many-to-many mapping problem in word learning is that the *environment*, rather than the learner, contains many properties that assist in constraining possible mappings (MacWhinney, 1991). Though a single learning situation contains many possible words and many possible referents for those words, over multiple situations, children may observe that there are co-occurrences between particular words and particular elements of the environment. Yu and Smith (2007) showed that learners are able to exploit such cross-situational statistical relations between words and referents. However, the statistical associations are noisy in real-world child-directed speech settings (Yu & Ballard, 2007), and so additional cues in the environment are likely to assist further in constraining learning.

One possibility is distributional information in terms of co-occurrences between words. In English child-directed speech, determiners reliably precede nouns in complex utterances (Monaghan & Mattock, 2012), and these distributional cues can assist the child in knowing which potential words in an utterance are likely to refer to objects in their environment (Fitneva, Christiansen, & Monaghan, 2009). Other distributional cues that are readily available to children can also provide information about verb categories, and function versus content word distinctions (Childers, 2011; Christiansen & Monaghan, 2016).

Prosodic information is another cue to assist in reducing the many-to-many mapping problem, not only providing information about different grammatical categories (Christiansen & Monaghan, 2016) but also indicating speaker focus in a learning situation: Messer (1981) found that approximately 50% of child-directed utterances with a learning goal had the referring word reaching the highest amplitude.

For further reducing the possibilities for the intended referent, gestural cues provide additional cues to constrain word learning, with 15% of child-directed speech utterances accompanied by gestures that guided the child to the object being referred to (Iverson, Capirci, Longobardi, & Caselli, 1999).

Combining cues for word learning

Individually, then, cues appear to be noisy but informative sources of information about intended referents. Thus, combining cues is likely to result in yet more robust and faster learning. There are several models for how multiple cues may interact for word learning.

First, cues may be additive, such that more information provides cumulative evidence about word-referent mappings. For instance, in a computational model, Yu and Ballard (2007) demonstrated that mapping accuracy improved with the addition of distributional cues.

However, an alternative model for how multiple cues may support learning is provided by Bahrick, Lickliter, and Flom's (2004) intersensory redundancy hypothesis. In this theoretical model, multiple cues that indicate the same structure in language (such as multiple cues indicating the word-referent mapping, for instance) enable the learner to realise that this relation is not random, but carries information about the stimuli. Consequently, cues that are correlated increase in saliency and are attended to more as learning proceeds.

However, this view of increased saliency from redundant cues only applies when there are overlapping cues to structure, and the distribution of cues in the learning environment may be very different. Monaghan et al. (2007) examined cues to grammatical categories of words across a range of languages. They found that distributional information provided, unsurprisingly, valuable information about the role of words in each language – for instance, in English words that belonged to the verb category tended to succeed "you" and precede "the", whereas words that belonged to the category of nouns tended to succeed "the", and precede "to". But, in addition, Monaghan et al. (2007) also found that phonological coherence also applied to these grammatical categories – though there is substantial variation, nouns tend to sound like other nouns and verbs tend to sound like other verbs, in terms of a range of phonological and prosodic properties.

Yet, it was the interplay of these cues that was striking: when distributional information was a weak indicator of grammatical category, Monaghan et al. (2007) found that phonological cues were more reliable, and vice versa. Thus, there was not so much a redundant overlap of cues, but rather a serendipitous arrangement of cues across situations to provide useful information (Christiansen & Monaghan, 2016).

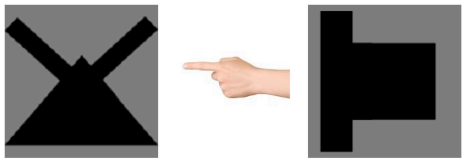
An alternative perspective, then, is that multiple cues for language structure enable robust learning, but not due to intersensory redundancy, but rather due to providing a safety net that is resistant to variation of their presence in the environment. In Monaghan (2017) this idea of degeneracy was implemented in a connectionist model that took as input multiple information sources to support learning of cross-situational statistical regularities between an object in vision and a word in auditory input, when both the object and the word occurred alongside others. The model was able to learn the cross-situational statistical regularities, but this learning was boosted when additional cues were added to the model's learning environment. One was distributional information (where the referring word was preceded by a marker word, such as "the" preceding a noun). Another was a prosodic cue, where the referring word in the utterance was emphasised in the auditory input. The final cue was a gestural cue, where attention was drawn to the object that was being referred to in the utterance. In each case, adding the cue improved the model's learning. Furthermore, adding all the cues improved performance still further.

The second set of simulations in Monaghan (2017) tested what effect individually unreliable cues would have on learning. The presence of each of the three cues varied between 33% and 100% of the time, but note that in most learning situations, at least one of the cues was likely to be present. The reduction of reliability of multiple cues reduced the speed of learning, however, following training, the ability of the model to respond correctly to word-object mappings when they were presented with no additional cues in the environment was more robust when cues were individually unreliable. The presence of noise in the environment, when that environment provides an unreliable constellation of individual cues, meant that the model was better able to recognise words when the environment was momentarily impoverished. Consider a language instructor who always pointed to the object to which they were referring. That is likely to be helpful for constraining the potential referents for words that the learner hears. But what would happen when the instructor is distracted – or a new instructor with different habits arrives – and does not provide the gestural cue? If the cue was previously 100% reliable, then this would become a crutch that was relied upon for determining the speaker's intention, and the referent would not be identifiable if not gestured towards.

A computational approach with a similar outcome is Srivastava et al.'s (2014) dropout model, where hidden units in a model are stochastically deactivated to prevent the model overlearning one aspect of the input – to resist relying only on the most reliable information stream in the environment, and consequently preventing effective generalisation. This switching off meant that the model maximised use of information from the environment. However, critically for our purposes, the learning system does not selectively prevent attention to environmental information. The noise in the language environment provides this function. Far from being a problem for learning, environmental noise enabled effective, reliable, and robust learning to take place, providing a positive perspective on poverty of the stimulus (Chomsky, 2005). Indeed, stimulus poverty resulted in rich learning.

However, the benefit of multiple, noisy cues is a prediction of the degeneracy model (Monaghan, 2017) but has not yet been tested empirically. Here, we provide a behavioural test of whether the presence of multiple, variable cues promotes robust word-referent learning. We constructed a cross-situational learning task, with each situation presenting learners with two objects and a set of words (see Monaghan & Mattock, 2012, for similar outline of the cross-situational word learning design). One of the words always referred to one of the objects, but the other object and the other words varied. Over multiple trials, participants may come to recognise that certain words and objects always co-occurred. We measured the extent to

which additional cues in the environment assisted in learning – implementing gestural, distributional, and prosodic cues to support learning, but we varied the extent to which these cues were present. The degeneracy model predicted that (very) noisy cues should slow learning, but that there may be an optimal level of variability at which learning is more accurate than perfect information conditions when all cues are present. We examined three levels of variability as well as no variability, where cues were present 25%, 50%, 75%, or 100% of the time. We measured performance during training exposure, and we also measured whether learning was robust to omission of cues – by testing participants after learning on trials where no cues were present. Based on the predictions of the degeneracy model (Monaghan, 2017), we anticipated that learning would be resistant to omission of cues in all conditions, but that omission of cues may be least affected when those cues were variable during exposure.



"tha FINTOOM noo chatten"

Figure 1. Example of a learning trial, containing distributional, prosodic (i.e., fintoom is emphasised in the speech), and gestural cues.

Testing the effect of multiple, variable cues for word learning

Method

Participants

Participants were 72 native English speaking adults, mean age = 19.8 years (SD = 2.46), who were students at Lancaster University. Participants were paid £3.50 for participating, or received course credit. Participants were assigned to one of four conditions (N = 18 per condition) which varied the extent to which cues were reliably present during training (25%, 50%, 75%, or 100% of the time).

Materials

The materials comprised a set of abstract objects and a set of novel words with which the objects were paired during learning. We took 10 arbitrary shape pictures from Fiser and Aslin (2002) (see Figure 1 for examples). For the speech, we generated 22 nonsense words. Ten of the words each referred to one of the object shapes. An additional 10 words did not refer to any shape. A final two words were also generated to act as distributional marker words. Words were read by a female native English speaker in monotone, and were also read in emphasized form, with the speaker imagining they were speaking the word to a child. Emphasised words had higher mean pitch, greater pitch variation, longer duration, and greater intensity than monotone words (all $t(19) > 8.98$, $p < .001$).

Each learning trial comprised an utterance containing a referring word and a non-referring word. When the distributional cue was present, the two words were preceded by marker words that distinguished the referring and non-referring word. When the prosodic cue was present the referring word was emphasised, otherwise both words were monotonic. When the gestural cue was present, a finger pointed to the intended referent. In the example trial shown in Figure 1, "tha" indicates the following word is the referring word and "fintoom" refers to one of the pictures (in this case, the picture on the left). Cues were randomly selected individually according to the variability condition (e.g., for the 25% cue, there was a ¼ chance that each cue was present or absent, such that there were trials where 3, 2, 1, or no cues were present).

An additional training block was constructed from 6 novel shapes and 12 novel words, but these new training data are not reported further here.

Procedure

Participants were instructed to try to learn which object was referred to by the speech. There were 6 blocks of training, each of which contained 30 trials, where for each trial an utterance was played through headphones and two objects were presented on a computer screen simultaneously. One of the objects was the target and always co-occurred with the referring word, the other object was selected from the remaining nine objects. Within each block of training, objects appeared an equal number of times as target and as foil, and were counterbalanced for appearing on the left or the right of the screen. Presence or absence of cues was manipulated between conditions by randomly selecting whether each cue was present or absent in 25%, 50%, 75%, or 100% of trials.

Participants responded by pressing "1" or "2" for left object or right object, respectively, on a computer keyboard. No feedback was provided on accuracy of performance.

After training, participants were tested for their knowledge of word-referent mappings when all cues were absent, to determine whether learning was robust, or required presence of cues for accurate performance.

Results

We conducted four separate analyses exploring how learning was affected by the variability of cues. In each analysis, a series of generalized linear mixed-effects models (GLMER) were performed, predicting the dependent variable of response accuracy (correct or incorrect). The models were built up incrementally, adding in fixed effects and performing likelihood ratio tests after the addition of each new fixed effect term (following Barr, Levy, Scheepers & Tily, 2013). Random effects of participant and experiment version were included in all reported analyses.

First, we analysed learning during training. The effect of block (1-6) significantly improved model fit ($\chi^2(1) = 314.1$, $p < .001$), indicating that over the course of training, there was a significant increase in participant's response accuracy. Including variability condition (25%, 50%, 75% and 100%) also significantly improved model fit ($\chi^2(3) = 21.259$, $p < .001$). Crucially, there was a significant improvement to model fit when the interaction term of block x condition was added ($\chi^2(3) = 71.113$, $p < .001$), indicating that performance over the course of training varied by reliability condition. See Table 1 for the final model summary, which indicates that the 75% condition resulted in more rapid learning than the other conditions (see Figure 2).

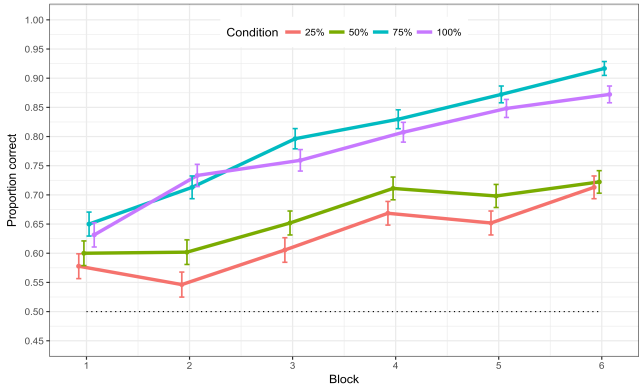


Figure 2. Learning trajectories for the word-object mapping cross-situational learning task with multiple cues of different reliabilities.

Table 1. GLMER model summary predicting accuracy from training data.

Fixed effects	est.	*SE*	*z*	*p*
(Intercept)	**.47**	**.21**	**2.24**	**.025**
Block	**.30**	**.03**	**10.96**	**<.001**
Condition (25%-100%)	-.39	.29	-1.36	.175
Condition (50%-100%)	-.17	.29	-0.58	.559
Condition (75%-100%)	-.10	.29	-0.33	.739
Block*Condition(25%-100%)	**-.17**	**.04**	**-4.81**	**<.001**
Block*Condition(50%-100%)	**-.17**	**.04**	**-4.65**	**<.001**
Block*Condition(75%-100%)	**.09**	**.04**	**2.11**	**.035**

For the second analysis, we investigated the effect that variability of cues had on sensitivity to the individual cues during training, by measuring the effect of presence of individual cues on learning. In this analysis, only trials where at least one cue was present were included (see Figure 3).

The addition of variability condition significantly improved model fit ($\chi^2(3) = 16.199$, $p = .001$), indicating that there was a difference in overall accuracy across conditions, with performance in the 100% condition being significantly greater than the 25% and 50% conditions (both $p < .01$), but not the 75% condition ($p > .05$). Next, the addition of cue type also significantly improved model fit ($\chi^2(2) = 32.083$, $p < .001$). This result indicates that there was a significant increase in accuracy when gesture cues were present, compared with when distributional and prosodic cues were present, both $p < .001$. Importantly, there was a significant improvement to model fit when the interaction term of variability condition x cue type was added ($\chi^2(6) = 23.665$, $p < .001$). See Table 2 for the final model summary, which indicates that when variability is at 75%, the salience of gesture cues was increased compared to the 100% condition, when cues were always present. Variability had the effect of emphasising the contribution of gesture. The benefit of gesture over the other cues was also present for 25% and 50% cues, but only when variability was at 75% was accuracy greater than the 100% condition.

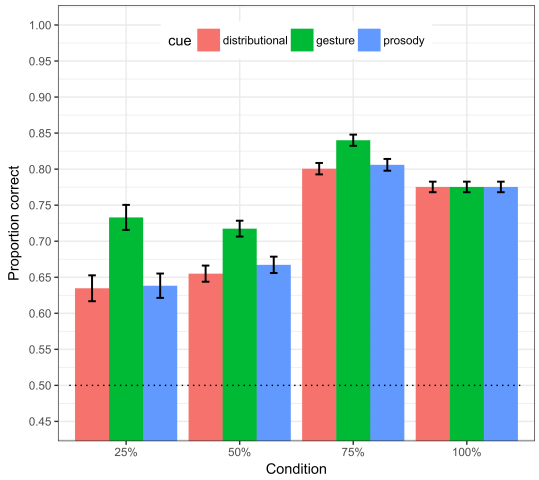


Figure 3. Performance during training trials by variability condition and cue type.

Table 2. GLMER model summary predicting accuracy from trials when at least one cue was present.

Fixed effects	est.	*SE*	*z*	*p*
(Intercept)	**.01**	**.01**	**7.08**	**<.001**
Condition (25%-100%)	**.01**	**.01**	**-2.99**	**.003**
Condition (50%-100%)	**.01**	**.01**	**-2.52**	**.012**
Condition (75%-100%)	.01	.01	0.69	.491
Cue(dist-gesture)	.01	.01	0.00	.99
Cue(dist-prosody)	.01	.01	0.00	.99
Cue(dist-gesture)*Condition(25-100%)	**.01**	**.01**	**3.53**	**<.001**
Cue(dist-gesture)*Condition(50-100%)	**.01**	**.01**	**3.23**	**.001**
Cue(dist-gesture)*Condition(75-100%)	**.01**	**.01**	**2.98**	**.003**
Cue(dist-prosody)*Condition(25%-100%)	.01	.01	0.01	.898
Cue(dist-prosody)*Condition(50%-100%)	.01	.01	0.62	.535
Cue(dist-prosody)*Condition(75%-100%)	.01	.01	0.39	.696

During the training trials, the number of cues available to the learner varied from 0 to 3 in the variability conditions. In order to determine the effect of number of cues present, we tested the number of cues present in terms of improvement to model fit. We found that they did ($\chi^2(1) = 66.342$, $p < .001$), indicating that as the number of cues present increased, accuracy improved (see Figure 4). Further, the interaction of number of cues x variability condition also improved model fit ($\chi^2(2) = 14.309$, $p < .001$).

In order to determine how variability affected use of cues, we examined accuracy when all cues were present,

comparing across variability conditions. Importantly, when all three cues were present, accuracy in the 75% condition was significantly greater than the 100% condition (estimate = .86, *SE* = .41, z = 2.11, p = .035). Thus, 75% variability improved the accuracy of performance when all cues were present.

Finally, we determined whether learning was robust under conditions of cue variability, and how variability affected performance during the test trials when none of the cues were present. The addition of variability condition significantly improved model fit ($\chi^2(3) = 11.357$, p = .010), with significant differences between the 100% condition when compared to the 25% and 50% conditions (both p < .05), but no significant difference between the 100% and 75% conditions (p = .556). Importantly, this reflects the pattern of results found in the final block of training (see Figure 2), where performance improved as reliability of cues increased. See Figure 5 for results. Thus, in all conditions learning was robust to absence of cues.

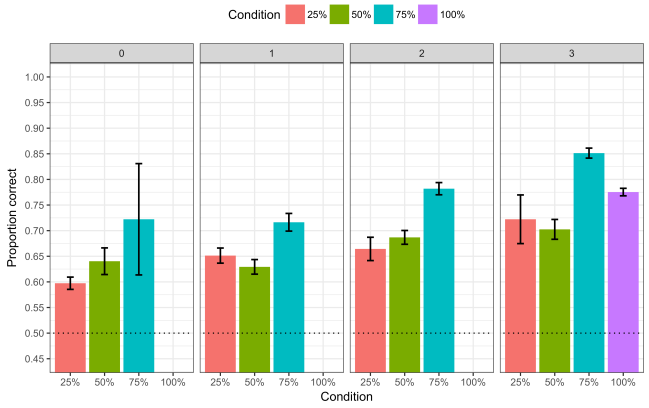


Figure 4. Test of performance for different number of cues present during training.

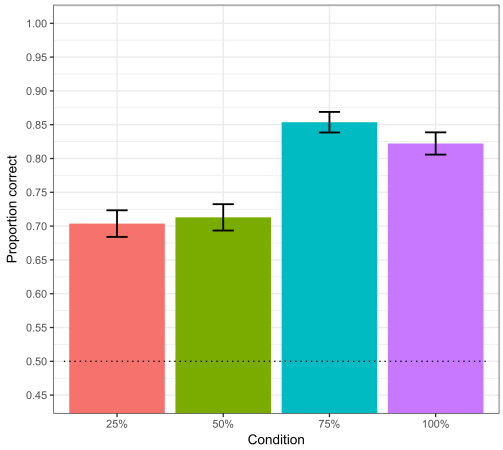


Figure 5. Performance on all words after training for test trials, when no cue was present.

Discussion

The main aim of this study was to test the effect of variation of multiple cues in the language environment for supporting word learning. We predicted, based on the degeneracy model of learning (Monaghan, 2017), that optimal performance would be a consequence of variable presence of multiple cues that aid learning. This was because the learner can exploit multiple information sources, without relying on any one cue, or coming to ignore the contribution of other highly correlated cues.

The results of the behavioural study of learning word-referent mappings supported the degeneracy model, in that learning was faster and more accurate when distributional, prosodic, and gestural cues occurred in 75% of trials during training, than when cues were present 100% of the time.

However, greater variability – 25% and 50% occurrence of individual cues – reduced accuracy compared to the 75% condition, indicating that, for learning a small number of words, the optimal conditions were with cues present more than half the time, but not all the time. In natural language learning situations, reliability of individual cues to support word learning seems to be substantially lower. For instance, the prosodic cue of highest amplitude as an indicator of the referring word occurs in 50% of learning situations (Messer, 1981), and explicit gestural cues occur substantially less often – even as low as 15% of learning situations (Iverson et al, 1999). However, these are situations where the vocabulary is much greater than the 10 word-object mappings of the current learning situation, and additional cues to word-referent mappings when the possibilities for those mappings are exponentially higher may have a greater effect even when they occur more rarely. For instance, the model of Monaghan (2017) was trained on 100 words, and under those circumstances 50% variability was found to be optimal for learning. Scaling up the current language to larger vocabularies will be an important further test of the principles of variation in multiple environmental cues.

Analysis of the trials where individual cues were present or absent indicated that the benefit of variability in presence of cues was greatest for the gestural cue, with variability enhancing the use made of this cue when it occurred (Figure 3). Such a result is consistent not only with the degeneracy model of multiple cues, but also with the intersensory redundancy hypothesis (Bahrick et al, 2004), such that correlated cues increase in salience, but with the exception that the redundancy should not be absolute: if cues are perfectly correlated then their salience does not increase, as in the 100% condition.

The results from analyses of different numbers of cues present showed that combining cues boosted learning (Figure 4), indicating that the learner was exploiting information present from each of the individual cues. It was not the case, for instance, that participants learned to only attend to particular cues, as their confluence resulted in greater improvement. Indeed, when those cues were variable but all present, performance was best of all – again, the 75% variability condition outperformed the 100% condition when

all three cues were available in the trial.

In all variability conditions, learning was shown to be robust to absence of individual cues. This is an important result, because it demonstrates that though cues can support learning, they do not over-shadow the cross-situational statistical relations between particular words and objects co-occurring. This was the case even when cues were always present, thus, even if multiple cues are always present they do not result in brittle learning of statistical relations. It may be that individual cues, if occurring with high reliability could interfere with robust learning (e.g., Srivastava, 2014), and this is a topic for future investigation.

We know that the language environment is noisy, but replete with numerous multimodal cues that point in different ways to the same language structures (Whitacre, 2010; Winter, 2014; Yurovsky, Smith, & Yu, 2013). We have shown that learners are able to exploit these multiple cues, and also their variability, to support word learning.

Acknowledgments

This work was supported by the International Centre for Language and Communicative Development (LuCiD) at Lancaster University, funded by the Economic and Social Research Council (UK) [ES/L008955/1]. Thanks to Katy Rudd for conducting pilot testing and contributing to the construction of materials.

References

Bahrick, L. E., Lickliter, R., & Flom, R. (2004). Intersensory redundancy guides the development of selective attention, perception, and cognition in infancy. *Current Directions in Psychological Science, 13,* 99-102.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255-278.

Childers, J. B. (2011). Attention to multiple events helps two-1/2-year-olds extend new verbs. *First Language, 31*, 3–22.

Christiansen, M.H., & Monaghan, P. (2016). Division of labor in the vocabulary. *Topics in Cognitive Science*, *8*, 610-624.

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry, 36*, 1-22.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, USA, 99,* 15822-15826.

Fitneva, S., Christiansen, M.H., & Monaghan, P. (2009). From sound to syntax: Phonological constraints on children's lexical categorization of new words. *Journal of Child Language, 36,* 967-997.

Iverson, J. M., Capirci, O., Longobardi, E., & Caselli, M. C. (1999). Gesturing in mother-child interactions. *Cognitive Development, 14,* 57–75.

Macnamara, J. T. (1982). *Names for things: A study of human learning* (p. 4). Cambridge, MA: Mit Press.

MacWhinney, B. (1991). A reply to Woodward and Markman. *Developmental Review, 11*, 192-194.

Markman, E. M. (1994). Constraints on word meaning in early language acquisition. *Lingua, 92*, 199-227.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology, 20*, 121-157.

Messer, D. J. (1981). The identification of names in maternal speech to infants. *Journal of Psycholinguistic Research, 10*, 69-77.

Monaghan, P. (2017). Canalization of language structure from environmental constraints: A computational model of word learning from multiple cues. *Topics in Cognitive Science, 9*, 21-34.

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The Phonological Distributional coherence Hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology, 55*, 259-305.

Monaghan, P. & Mattock, K. (2012). Integrating constraints for learning word- referent mappings. *Cognition, 123,* 133-143.

Quine, W.V.O. (1960). *Word and object.* Cambridge, MA: MIT Press.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929-1958.

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

Yu, C. & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*, 414-420.

Yurovsky, D., Smith, L. B. & Yu, C. (2013). Statistical word learning at scale: The baby's view is better. *Developmental Science, 16*, 959-966.