Degeneracy results in canalisation of language structure: A computational model of word learning

Padraic Monaghan (p.monaghan@lancaster.ac.uk)
Department of Psychology, Lancaster University
Lancaster LA1 4YF, United Kingdom

Abstract

There is substantial variation in language experience between learners, yet there is surprising similarity in the language structure they eventually acquire. While it is possible that this canalisation of language structure may be due to constraints imposed by modulators, such as an innate language system, it may instead derive from the broader, communicative environment in which language is acquired. In this paper, the latter perspective is tested for its adequacy in explaining the robustness of language learning to environmental variation. A computational model of word learning from cross-situational, multimodal information was constructed and tested. Key to the model's robustness was the presence of multiple, individually unreliable information sources that could support learning when combined. This "degeneracy" in the language system had a detrimental effect on learning when compared to a noise-free environment, but was critically important for acquiring a canalised system that is resistant to environmental noise in communication.

Keywords: canalisation; degeneracy; language acquisition; multiple cues; word learning

Introduction

A key question in the cognitive sciences is how, despite the enormous variation in linguistic experience, each language learner acquires broadly the same language structure, "within a fairly narrow range" (Chomsky 2005). This issue has led to proposals for mechanisms that ensure this "canalisation" of language structure. Traditionally, these mechanisms have been conceived as constraints that apply to structure the language exposure, such as innately specified syntactic or semantic properties. But there is growing realisation that multiple, rich sources of information within the communicative environment may offer substantial, perhaps sufficient constraints to learning.

A similar change in perspective was observed in canalisation in biological evolution. Initial proposals were that canalisation was a consequence of the natural selection of mechanisms that operate to minimise phenotypic variation (Waddington, 1942). However, a more recent explanation is that minimal phenotypic variation is stably achieved as a consequence of interaction between multiple regulators (despite substantial environmental variation) as part of the developmental process of the organism (Siegal & Bergman, 2002). Simulations of the developmental operation of multiple transcriptional regulators found that the greater the interactivity between these sources, the smaller the phenotypic variation resulting from environmental variation.

An analogous perspective can be taken on canalisation of social or cultural systems, such as language, whereby increasing levels of interaction may increase the stability and optimise performance of an information processing system (Bettencourt, 2009). Canalisation of language, long conceived as being a consequence of mechanisms that implement resistance to environmental variation, could instead be the outcome of interacting, multiple sources of information.

Recently, there has been reconsideration of the potential for language learning to be supported by the richness of the language environment. For instance, grammatical category acquisition is not only supported by information from word co-occurrences – the traditional information source for linguistics studies of language acquisition (Redington, Chater, & Finch, 1998) – but also from substantial information in phonotactic and prosodic structure, such as distinct stress patterns on nouns compared to verbs (Monaghan, Christiansen, & Chater, 2007). Furthermore, information about objects and actions within the child's purview may further constrain potential referents for words (Yurovsky, Smith, & Yu, 2013), providing restrictive information about the semantic features associated with particular categories.

There have been several accounts for how such multiple cues may be combined to support learning. The redundancy of different information sources may assist the learner by increasing the saliency of particularly important information present in their environment (Bahrick, Lickliter, & Flom, 2004). Alternatively, the cues may operate summatively (Christiansen, Allen, & Seidenberg, 1998), or they may operate in a hierarchy, such that if one cue is available then it is used in preference to other cues, which are relied upon only if the preferred cues are unavailable (Mattys, White, & Melhorn, 2005).

An alternative possibility, consistent with models of canalisation in biology, is that multiple cues for language learning interact, resulting in a system that is stable in the face of variation in the environment. This property of language is its "degeneracy", defined as "the ability of elements that are structurally different to perform the same function or yield the same output" (Edelman & Garry, 2001). Degeneracy affects not only acquisition – where presence or absence of particular cues will not adversely affect the structure acquired – but also the robustness of the system once the language is acquired, due to reduced dependency on any one information source. Computational models of degeneracy in language and other complex systems have shown that it is important for robustness of

learning (Whitacre, 2010), permitting, for instance, effective processing of speech sounds against background noise (Winter, 2014).

In this paper, a computational model of multiple interacting information sources is presented as a proof of concept that degeneracy can result in canalisation of language structure. The domain of study is word learning, where forms and meanings of words have to be mapped. This task is difficult, because there are numerous possibilities for the target candidate word in multi-word utterances, and multiple possible referents in the environment to which the target word may map (Quine, 1960). However, multiple cues are present both in the spoken language and in the environment that surrounds the learner to assist in this task. This perspective requires extending the notion of degeneracy from examining the redundant, overlapping cues within language structure to examine cues more broadly within the communicative situation.

Within spoken language, information about the grammatical roles for words can be ascertained from distributional information, consequently reducing the number of possible target words that need to be considered. For instance, nouns are frequently preceded by articles (the, a) and these also tend to succeed verbs. Use of such simple distributional information has been shown to assist in determining word referent mappings (Monaghan & Mattock, 2012). Further information for identifying the critical information in an utterance is also available from prosodic information. When teaching a child a new word, the speaker tends to increase the pitch variation, intensity, and duration of the target word within the utterance (Fernald, 1991).

In addition, constraints within the environment help to reduce uncertainty about the potential referents. One of these information cues is derived from cross-situational statistical information, where over multiple situations the learner can increase their association between the target word and target object (McMurray, Horst, & Samuelson, 2012) even when several possible words and referents are present. Such cross-situational learning (Yu & Smith, 2012) can further be supplemented by information that the speaker uses to indicate the field of reference. For instance, speakers tend to use deictic gestures (finger pointing or eye gaze) toward a referent which is being described (Iverson, Capirci, Longobardi, & Caselli, 1999).

However, in isolation, each of these cues is insufficient to perfectly constrain learning: The word succeeding an article is not always a noun – in English adjectives might intervene, and spontaneous language is replete with false starts, and word sequencing errors. Similarly, the loudest word in speech is not always the target word, or a novel word being learned by the listener, and gestural cues are not always reliable. In Iverson et al.'s (1999) study they found that 15% of utterances were accompanied by gestures indicating aspects of the immediate environment to direct children's attention. Yet, such unreliability has profound value for learning. Consider if the child always learned from a speaker who reliably pointed to the intended referent. Then, if ever a situation arose where a referent was not gestured towards, this could impair effective communication, because the cue may be relied upon for effective mapping from word to referent.

There are costs to including multiple cues in the learning situation, because this increases the amount of information needed to be processed in each instance of learning. So, the trade-off between the increased strain on the cognitive systems required by processing of multiple as opposed to single, or no, cues and the potential advantages of interacting information sources for learning must be examined. Specifically, we tested the value of multiple information sources for learning, and we examined the importance of interaction among information sources for linguistic canalisation, i.e., the robustness of learning in the face of environmental variation.

A computational model was constructed to test integration of information received from multiple sources to assist the learning of relations between words and their referents. Two sets of simulations testing the model were conducted. The first assessed the contribution of single cues to word learning. The hypothesis was that adding cues to the input would assist in acquisition of the mapping, with gestural cues assisting in defining the referent, prosodic cues promoting identification of the target word, and distributional cues supporting acquisition of both. However, the reliable presence of cues was hypothesised to result in impaired ability to identify the form-meaning mapping when the cue is no longer present.

The second set of simulations explored the role of multiple cues for learning. The prediction was that multiple cues would further promote learning, but that noisy cues would be most effective for supporting not only effective acquisition but also robustness in the learning, immune from effects of variability in the environment. Thus, a model trained with a degenerate environment should result in a canalised system, being able to effectively map between words and referents even when environmental cues that support this mapping are no longer available.

A Multimodal Model of Word Learning

The starting point for the current model used the hub-and-spoke architecture (Plaut, 2002), where information from different modalities is inputted to a central processing resource, and is thus unconstrained in its integration. These models then determine the optimal way in which information sources can cohere to support learning. The model implementation is closely based on a previous model of multimodal information integration in sentence processing, which was created to simulate behaviour in the visual world paradigm (Smith, Monaghan, & Huettig, 2014). This modeling approach has been effective in demonstrating how and when different information modalities interact in language processing, and how the influence of different modalities on language processing

derive from the nature of the representations themselves, rather than requiring architectural assumptions to be imposed on the system.

The model used here is a subsystem of this larger modeling enterprise, addressing the special case of acquiring word-referent mappings. The model is compatible with previous associative models of word learning (McMurray et al., 2012), as well as being broadly consistent with the principles of statistical models of cross-situational word learning (Yu & Smith, 2012). The model therefore applies these general modeling principles to explore the role of multiple information sources in facilitating, and constraining word learning.

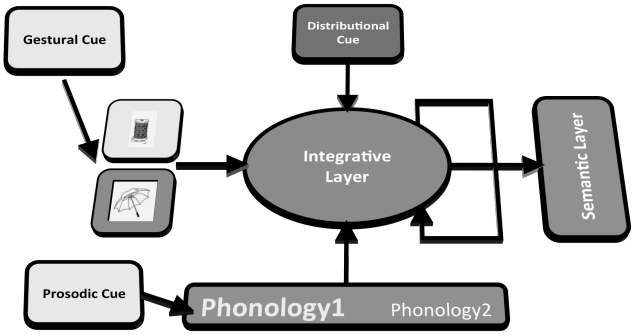

Figure 1: The multimodal integration model of word learning.

Architecture

The model architecture is illustrated in Figure 1. The model is implemented as a recurrent backpropagation neural network. It comprises a central hidden layer of 100 units which received connections from various input modalities, and projected to a semantic layer output.

The phonological input represented two word slots, each of which contained 20 units. The visual input contained two locations each comprising 20 units, where object representations were presented. The semantic layer was composed of 100 units. For some simulations that included a distributional cue, the model also received input from a distributional cue layer, which was composed of 2 units. The integrative layer was also fully self-connected.

Representations

The model was trained to learn 100 words. Representations of each modality of a word was encoded as a pseudopattern so that the properties of the relations between representations could be controlled. The phonological representation of each word was composed of four phonemes, randomly drawn from a set of 10 different phonemes. Each phoneme comprised 5 units, with 2 units active. The visual representation of the word's referent was constructed from 20 units with 8 units active for each representation. The semantic representations were localist, such that one of the 100 units was active for each of the words.

Fifty of the words were randomly assigned to one category, and the remaining fifty were assigned to the other category, such that these categories could be defined by a distributional cue.

Table 1: Proportion of training trials with each cue according to condition.

Condition	Dist Cue	Prosodic Cue	Gestural Cue
No Cue	0	0	0
Single Cues			
Dist Cue	1	0	0
Prosodic Cue	0	1	0
Gestural Cue	0	0	1
Combined Cues			
25% reliability	.25	.25	.25
50% reliability	.50	.50	.50
75% reliability	.75	.75	.75
100% reliability	1	1	1

Training

The model was trained to identify the meaning of the word from phonological and visual representation inputs, for all 100 words. Each trial was a simulation of a cross-situational learning task, where two words and two objects were presented, but only one of the objects was named by one of the words (Monaghan & Mattock, 2012). The model had to learn to solve the task by generating the correct semantic representation for the named object.

For each training trial, a word was randomly selected. Its phonological form was presented at one of the two word slots in the phonological input (position was randomly chosen), and another randomly selected word's phonological form was presented at the other word slot. The object depicting the word's referent was presented at one of the two visual input positions (randomly chosen) and another randomly selected visual representation was presented at the other visual input position.

For the simulations with cues, gesture and prosody were implemented as intrinsic properties of the visual and phonological input representations, respectively, by doubling the activation at the input of the target visual object or the target phonological form. This had the effect of increasing the contribution of the target representation within each representational modality to affect the activation state of the integrative layer, and was a simulation of increased saliency of that representation (i.e., that a gestural cue increases saliency of the target object, and prosodic cue is implemented as an increase in intensity, duration, and pitch of the target spoken word). This is illustrated in Figure 1 as a highlighting of the uppermost object and the first phonological representation as a consequence of gestural and prosodic cues, respectively.

The distributional cue was implemented as an extrinsic cue. If the word was from the first (randomly assigned) category then the first unit in the distributional layer was

active, and if the word was from the second category the second unit was active. This cue could therefore assist the model in determining which was the target object and spoken word, but the cue did not operate within either of these modalities.

The simulations of single cues presented each learning trial with the cue present with 100% reliability (see Table 1). The simulations of multiple cues varied the extent to which the cues were reliably present in each learning situation, from 25%, through to 100% reliability.

Activation cycled in the model for 6 time steps. At time step one, the visual and phonological inputs were presented. For two time steps activation passed from the input to the integrative layer and from the integrative layer to the semantic layer, and from the integrative layer to itself. At time steps 3 to 6 the target semantic representation was presented at the semantic output layer, and activation continued to cycle around the model. The model was trained with continuous recurrent backpropagation through time (Pearlmutter, 1989) with error determined by sum squared error of the difference between the actual and target semantic representations. In one epoch of training, each of the 100 words occurred once as the target. The model was trained up to 100,000 epochs at which point performance for each condition had asymptoted.

Twenty versions of the model with different pseudopattern representations, different randomised starting weights, and different randomised ordering of training patterns were run.

Testing

The model's performance was assessed during training on its ability to produce the target semantic representation for each phonological and visual input. If the activity of the semantic unit corresponding to the target word was more active than any other unit in the semantic layer, then the model was determined to be accurate.

Accuracy during training was assessed, and also the point in training at which the model was able to accurately detect all 100 words for five consecutive epochs.

At the end of training, the robustness of the model's learning was assessed by measuring its accuracy when no cues were present during testing.

Results

Single Cues

The model's accuracy during training when no cues and single cues were present is shown in Figure 2.

An ANOVA with epoch at which the simulation reached the accuracy criterion as the dependent variable, and cue condition (no cue, distributional cue, prosodic cue, gestural cue) as within subjects factor was conducted to test whether the model learned differently according to the presence of cues. The result was significant, $F(3, 57) = 70722$, $p < .001$, $\eta_p^2 = 1.00$. Post hoc tests revealed that the model reached criterion more quickly for the prosodic cue (mean epochs = 35,800, SD = 1,005), and gestural cue (mean = 35,650, SD = 745) conditions than the no cue condition which had not reached criterion by 100,000 epochs (mean proportion correct was .96), both $p < .001$. Though the trajectory of learning was distinct, as shown in Figure 2, the effect of distributional cues was smaller, and not significantly different in time to criterion compared to the no cue condition (mean proportion correct after 100,000 epochs was .99). The prosodic and gestural cues supported learning more than the distributional cue, both $p < .001$, but there was no statistical difference in speed of learning from the prosodic and gestural cues, $p = 1$.

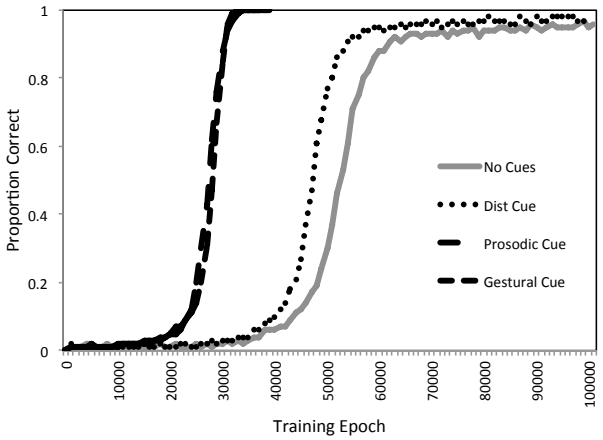

Figure 2: Accuracy during training for the single cues conditions, compared to no cue condition.

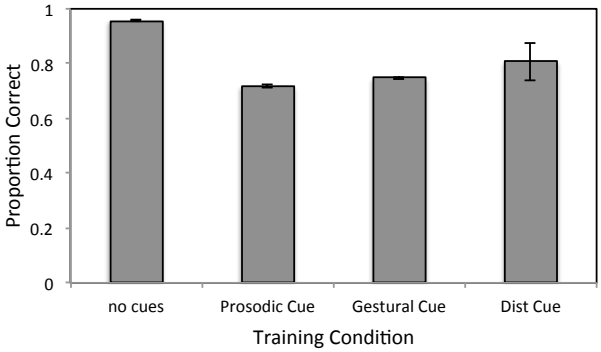

Figure 3: Accuracy after training for the single cues conditions, when no cues are present during testing (Dist = Distributional).

The robustness of the model's learning to omission of cues during testing is shown in Figure 3. An ANOVA on accuracy in the post-learning test with no cues present, and cue condition as within subjects factor was significant, $F(3, 57) = 8.982$, $p < .001$, $\eta_p^2 = .321$. Post hoc tests showed that the distributional cue did not significantly affect robustness of learning compared to the no cue condition, $p = .284$, however, the prosodic and gestural cue both resulted in poorer performance than the no cue condition, both $p < .001$. The gestural cue resulted in more robust learning than

the prosodic cue, p = .001, but these conditions did not differ significantly from the distributional cue condition, both p = 1.

We tested whether the difference between the intrinsic cue conditions (prosodic and gestural cues) was due to their quicker acquisition. We trained every model to the same number of training trials (100,000) then tested robustness of learning. The results were similar. Even with more training, the effect of a single, reliable intrinsic cue was detrimental to the model's ability to map between form and meaning when the cue was not present, $F(3, 48) = 45.62$, $p < .001$, $\eta_p^2 = .740$. Prosodic and gestural cues were now not significantly different than one another, $p = .423$, but were both significantly different than the no cue and the distributional cue conditions, all $p < .001$.

Multiple Cues

The model's accuracy during training for combined cues with different levels of reliability is shown in Figure 4. For epoch taken to reach training criterion, an ANOVA indicated that combined cues with different reliability significantly affected speed of learning, $F(4, 76) = 3855$, $p < .001$, $\eta_p^2 = .99$. Post hoc tests indicated that learning in the no cue and the 25% cue reliability condition were significantly slower than the 50% condition, both $p < .001$, which was in turn slower than the 75% condition, $p < .001$, which was in slower than the 100% perfect reliability multiple cue condition, $p < .001$. Thus, as anticipated, the greater the reliability of information present during learning, the faster the model learned to map between forms and meanings.

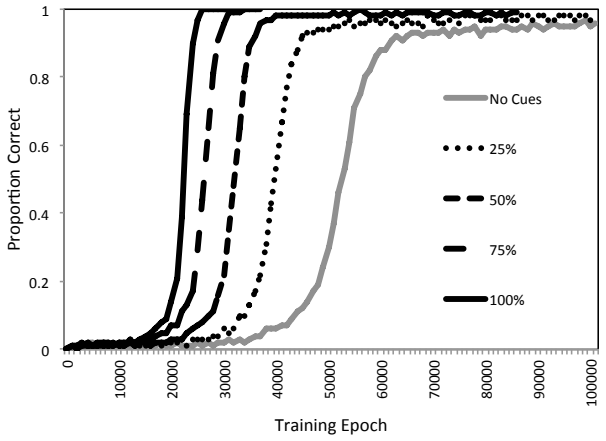

Figure 4: Accuracy during training for the multiple cue conditions, compared to no cue condition.

The robustness of learning was also compared between these conditions. The results are shown in Figure 5. An ANOVA demonstrated that the robustness of performance at testing was affected by the cues present during training, $F(4, 76) = 2.953$, $p = .025$, $\eta_p^2 = .135$. Post hoc tests revealed that the no cue and 50%, 75%, and 100% cue conditions were significantly different, all $p < .001$. The 25% cue condition was not significantly different than any other condition, all $p \geq .718$. As reliability increased from 50% to 75%, the robustness of the model declined, $p < .001$, and similarly declined from 75% to 100% reliability, $p < .001$. Thus, low reliability of cues did not seem to assist in learning quickly or robustly, but once individual cues appeared at least half the time, further increasing the reliability of the cues began to reduce the resistance of the model to the absence of cues after training. 50% reliability appears to be close to the optimal trade-off for speed and robustness of learning.

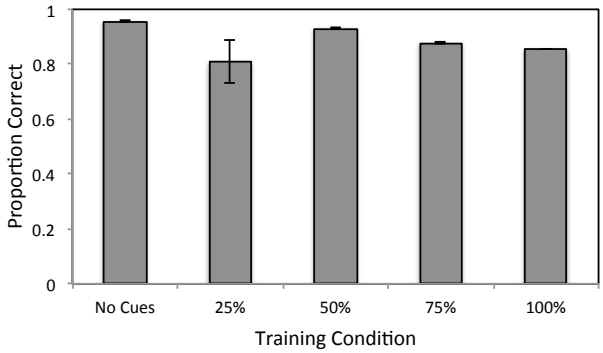

Figure 5: Accuracy after training for the multiple cue conditions, when no cues are present during testing.

Discussion

Language learning occurs in situations where multiple, interacting sources of information are available to support acquisition. Although attending to multiple cues increases the processing load on the individual, this degeneracy in language results in two important advantages for the language learning system.

First, adding a combination of cues to the model's input improves the speed and accuracy of learning to map between representations. Providing some guiding information about the intended object in a scene containing more than one referent, and emphasis of the target word in a multiword utterance, along with additional information about the general category of the target, improves performance. Even when the individual cues occurred only 50% of the time, learning of form-meaning mappings was still significantly enhanced compared to learning in the absence of cues.

This observation that speed and accuracy of language learning is promoted by multiple cues has been explored extensively, and is consistent with several current accounts of multiple cue integration in learning (Bahrick et al., 2004; Christiansen et al., 1998; Mattys et al., 2005; Monaghan et al., 2007). All these theories would predict the growing advantage of learning as cues increase in reliability, as observed in the current simulations.

However, the degeneracy of language also has a second advantage. The learning that is acquired from a degenerate

environment is highly robust (Ay, Flack, & Krakauer, 2007), and the model was able to make use of cues even when they were variably present across communicative situations. However, this multiple cue advantage for robustness was only observed when there was noise in the environment: When the cues occurred with perfect reliability then, even though learning was at its fastest, the acquired system was fragile and prone to error under suboptimal subsequent conditions. Thus, canalisation of language structure in a word learning task can be conceived of as a consequence of the interaction of multiple information sources for learning.

There is therefore a trade-off between speed of initial learning, and the robustness of that learning. The former is supported by perfectly reliable information (see, e.g., Onnis, Edelman, & Waterfall, 2013), and more information resulted in better and better learning. The latter is supported by multiple information sources, but with each individual source being somewhat noisy. The precise point of this trade-off is an issue for further exploration in computational systems, in order to determine the extent to which natural language environments are optimally designed for acquisition.

The simulations presented here suggest that, rather than canalisation being a challenge in the face of environmental variation, it is instead a primary consequence of this variation in a system that is able to integrate multiple information sources.

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 Rebecca Frost for comments on this work.

References

Ay, N., Flack, J., & Krakauer, D. (2007). Robustness and complexity co-constructed in multimodal signalling networks *Philosophical Transactions of the Royal Society B: Biological Sciences, 362* (1479), 441-447.

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.

Bettencourt, L. M. A. (2009). The rules of information aggregation and emergence of collective intelligent behavior. *Topics in Cognitive Science, 1*, 598–620.

Christiansen, M.H., Allen, J. & Seidenberg, M.S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*, 221-268.

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

Edelman, G., & Gally, J. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences, 98* (24), 13763-13768.

Fernald, A. (1991). Prosody in speech to children: Prelinguistic and linguistic functions. *Annals of Child Development, 8*, 43-80.

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

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General, 134,* 477–500.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*(4), 831-877.

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.

Onnis, L., Edelman, S., & Waterfall, H. (2011). Local statistical learning under cross-situational uncertainty. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society.*

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation*, *1*, 263-269.

Plaut, D. C. (2002). Graded modality-specific specialization in semantics: A computational account of optic aphasia. *Cognitive Neuropsychology*, 19, pp 603-639.

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

Siegal, M. L., & Bergman, A. (2002). Waddington's canalisation revisited: developmental stability and evolution. *Proceedings of the National Academy of Sciences*, *99*(16), 10528-10532.

Smith, A.C., Monaghan, P., & Huettig, F. (2014). Literacy effects on language and vision: Emergent effects from an amodal shared resource (ASR) computational model. *Cognitive Psychology, 75,* 28-54.

Waddington, C. H. (1942). Canalisation of development and the inheritance of acquired characters. *Nature*, *150*(3811), 563-565.

Whitacre, J. (2010). Degeneracy: a link between evolvability, robustness and complexity in biological systems *Theoretical Biology and Medical Modelling, 7,* 6.

Winter, B. (2014). Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. *BioEssays*, *36*(10), 960-967.

Yu, C., & Smith, L. B. (2012). Modeling cross-situational word–referent learning: Prior questions. *Psychological Review*, *119*(1), 21-39.

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