In: Westermann, G., & Mani, N. (Eds.) (2018). *Early Word Learning*, pp. 138-154.
New York: Routledge

Chapter 11: Computational models of word learning

Gert Westermann & Katherine Twomey

Department of Psychology, Lancaster University, UK

g.westermann@lancaster.ac.uk, k.twomey@lancaster.ac.uk

Abstract

Computational models are a means to develop explanations for the mechanisms underlying human behavior and behavioral change. Specifically, one type of computational models, artificial neural networks, has been used widely in modeling children's language and cognitive development. These models can learn from their experience with an environment and are sensitive to the environment's statistical structure, making them ideally suited to investigating how statistical learning can account for aspects of word learning across all levels, from the earliest phoneme acquisition to the development of the bilingual lexicon. Here we first describe the general principles underlying artificial neural network models with a goal of making them accessible to readers without experience with computational modeling. We then review the most common model architectures that have been used in simulating children's word learning in the broad context established in the other chapters of this volume. Finally, we review a number of specific models of word learning and discuss their contributions to our understanding of the mechanisms underlying early word learning, and the factors that shape this process in infants and toddlers.

Introduction

Computational models of word learning have made important contributions to understanding the mechanisms underlying this process. Different models have addressed virtually all the aspects of word learning discussed in the other chapters of this book, from learning speech sounds and segmenting words from speech, to mapping words to objects and acquiring a lexicon. In this chapter we first motivate the computational modeling approach and discuss what it can contribute to our understanding of cognitive development in general, and to word learning in particular. We then explain the basic principles of one widely used type of computational model, artificial neural networks. Finally, we review and evaluate several word learning models and their contributions to our understanding of this process.

Why computational modeling?

Computational models are computer programs that mimic some aspect of psychological processing. If their performance matches human behavior on a set of defined criteria, the mechanisms implemented in the model can serve as an explanation for the simulated human behavior. One way to look at a model is as a restricted artificial organism with a very limited set of specific behaviors. Such a model can be exposed to data similar to that experienced by humans in an experimental task (such as a speech segmentation task) or in their natural environment (such as hearing parental language), and the performance of the model (e.g., looking times in response to a set of stimuli) can then be compared with human data. It is then possible to manipulate the model, for example, by changing its internal processing mechanisms or the data to which it is exposed, to examine changes in performance. This approach can lead to predictions about human behavior in new situations which can be tested in experiments with human participants. In developmental psychology, computational models are often used to account for the change in cognitive abilities across age, allowing researchers to examine the effects of accumulating experience and changes in learning processes on the observed developmental trajectories.

One class of computational models that have been particularly powerful in furthering our understanding of cognitive development, and on which we focus in this chapter, are artificial neural network models, also called connectionist models (Mareschal & Thomas, 2007; Munakata & McClelland, 2003; Quinlan, 2003; Westermann & Plunkett, 2007). On a relatively abstract level these models are inspired by the functioning of neural networks in the brain. The basic idea here is that cognition arises from the complex interactions of many simple processing units (neurons in the brain). Consequently, connectionist models aim to show precisely how network structure, processing mechanisms and experience with the environment can give rise to such high level processes. The most important property of connectionist models is that they can learn from experience with the environment (as detailed in the next section), making them ideally suited to model children's cognitive development as an interaction between internal learning processes and experience with an environment.

A large number of connectionist models have been applied to various aspects of word learning (e.g., Althaus & Mareschal, 2013; Aslin, Woodward, LaMendola, & Bever, 2006; Li, Altmann, Hare, McRae, & Plunkett, 2007; Mayor & Plunkett, 2010; McMurray, Horst, & Samuelson, 2012; Räsänen, 2011; Samuelson, Schutte, & Horst, 2009; for an overview see Westermann, Ruh & Plunkett, 2009). These models suggest how such diverse empirical findings as a vocabulary spurt, overextensions of meaning, the effect of labels on object categorization and many others can arise from general learning mechanisms. This chapter will review such computational approaches to word learning with a focus on artificial neural networks that link cognitive development to processes in the brain.

We first describe the principles of artificial neural networks and specifically the three most common types used in models of word learning. Instead of focusing on the minutiae of the models' functioning we hope to achieve two things: first, we discuss the overall design decisions that modelers face when developing a model. This point is often neglected but we think it might be interesting to non-modelers in helping to assess the usefulness of a model. Second, we discuss the general principles and contributions made by models to our understanding of numerous aspects of word learning discussed in other chapters in this book.

Principles of artificial neural networks

Artificial neural networks (ANNs) consist of often large numbers of simple processing units with weighted connections between these units. Although a variety of specific modeling paradigms exist, they share some common principles. In all models the units can be activated, and activation then flows through the connections to other units. How much activation flows through a connection depends on the strength (weight) of this connection, which can be positive or negative (or indeed, zero). Each unit typically sums up the activation it receives through these connections (or directly from the environment), and if this activation is greater than a certain threshold, or if it falls within a certain range, the unit becomes active itself and in turn sends activation through its outgoing connections. ANNs are loosely inspired by the basic principles of the functioning of biological neurons in the brain. These neurons receive activation through synaptic connections with other neurons, and if this incoming activation exceeds the neuron's firing threshold it creates a spike that then travels through its axon to the synaptic connections with further neurons.

Despite these superficial similarities between artificial and biological neural networks, ANNs should not be seen as attempts to implement the specific biological networks underlying cognitive development and processing – the function of biological neurons is of course far more complex than the described principles, and the number of neurons involved in a specific function in the brain are by several orders of magnitude larger than the number of units in even the largest ANNs. Nevertheless, ANNs show how even complex cognitive functions can emerge from the interactions of large numbers of simple nonlinear associative processors.

The most important property of ANNs for modeling cognitive development is their ability to learn from experience. Learning occurs through changes to the weights of the connections between neurons, resulting in changes in the activation patterns across the network. Different types of models vary in the specific way in which weights are updated, and they will be discussed below.

How a model experiences the world

Information is presented to neural network models in the form of strings of numbers that translate into the activation values of input units. For example, phonemes are often encoded by the presence or absence of phonetic features such as voiced, labial, plosive for consonants, and frontal and low for vowels. An input will include all possible features, set to 1 when the feature is present, and to 0 otherwise. Similarities between inputs can therefore easily be represented. For example, the only difference between the representations for /p/ and /b/ is that the 'voiced' feature is set to 0 for /p/ and to 1 for /b/, while both phonemes have 1s for 'bilabial' and 'plosive' and 0 for all other features (depending on the specific feature description). Representations such as these, where similarities are reflected, are called *distributed* representations and they are chosen when similarity is assumed to play a role in processing (e.g., in modeling mispronunciations, priming, or categorization). In contrast, *localist* representations allocate a separate input for each item, for example, a word in a model of word-object mapping. This encoding scheme assumes that similarities between different inputs are irrelevant for the simulated process (such as mapping two distinct words to two distinct objects).

Another important aspect of computational models is the statistical structure of the environment. As discussed, neural network models learn from experience, and the more often a specific input or a class of inputs is experienced the more the model learns from it. Therefore, many models of word learning aim to reflect the statistical structure of a child's experience. For example, a model of vocabulary development could present words to the model according to the frequency with which these words are uttered to children (gleaned from corpora of child directed speech).

Together, these factors show that the modeler has to make specific assumptions at each step of the modeling process. A computational model thus not only comprises the processing architecture but also the representation scheme of environmental information and the statistical structure of the environment of the system – an aspect that is sometimes forgotten when discussing and evaluating models.

We will now discuss the most common types of ANN that form the basis for many of the models of word learning.

Supervised learning

In supervised learning a model receives an input and has to learn to generate a specific output as a response. Supervised models are usually arranged in a layered structure with an input layer that receives information from the environment, an output layer that generates a response, and a variable number of intermediate ('hidden') layers (often just one; Figure 1). Often all units in one layer send outputs to all units in the next layer through weighted connections: strong connections send more information than weak ones. There can also be negative connection values so that one unit can inhibit the activation of a downstream unit. Recent 'deep learning' models contain many more hidden layers but work on the same basic principles as these simpler models (Zorzi, Testolin, & Stoianov, 2013).

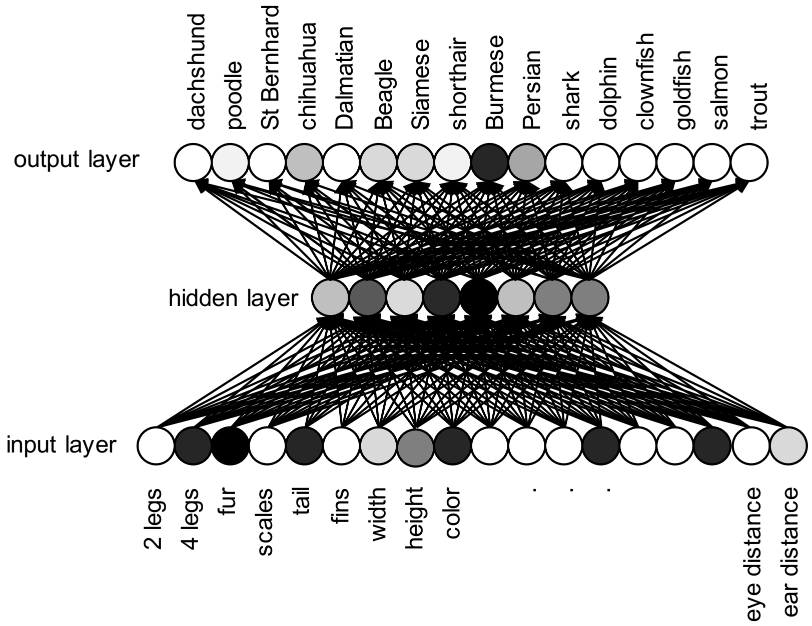

Figure 1: A three-layer neural network. This illustrative model learns the mapping from a set of features to an animal. Activation values are indicated by grayscale. Whole some activations on the input layer are binary (e.g., 4 legs: yes/no), other are continuous (e.g., width). On the output layer several units can be activated to different degrees, reflecting the uncertainty of the model.

When an input is presented to a supervised model, the respective units on the input layer are activated and send activation to the hidden layer. Each unit in the hidden layer sums up the incoming activation and computes its own activation value as a function of this incoming activation. The hidden units then send their activation through outgoing connections to the output layer. There again, units become active as a function of their incoming activation. The pattern of output unit activations is then interpreted by the modeler as a response.

In a supervised model the pattern of output activations is compared with a desired (target) pattern. Then, for each unit the incoming weights are adjusted so that the actual output will become closer to the target output. In effect, when a unit's output is lower than the target value, its incoming connections from active units are strengthened. Conversely, when the output is higher than the target, connection weights are weakened. The most popular weight change algorithm that enables this process for multiple network layers is the backpropagation algorithm (Rumelhart, Hinton, & Williams, 1986). In effect, this algorithm sends error signals backwards through the network to generate target values for the internal (hidden) units that do not have an explicit target from the environment.

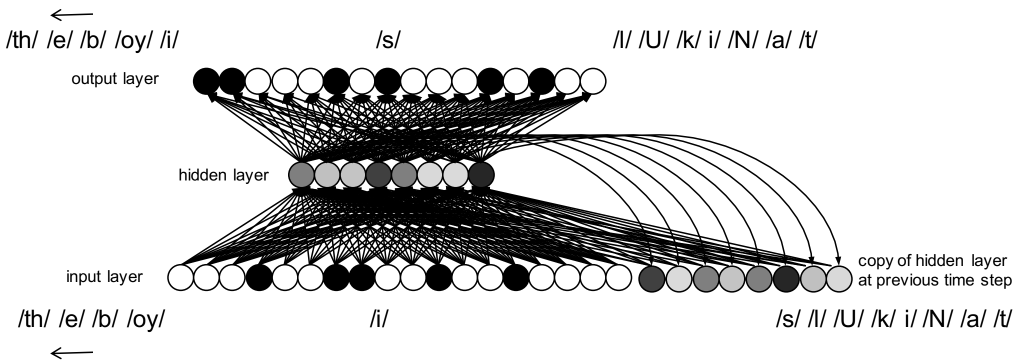
/th/ /e/ /b/ /oy/ /i/
/s/
/l/ /U/ /k/ i/ /N/ /a/ /t/
output layer
hidden layer
input layer
copy of hidden layer at previous time step
/th/ /e/ /b/ /oy/
/i/
/s/ /l/ /U/ /k/ i/ /N/ /a/ /t/

Figure 2: A simple recurrent network. This model is here presented with an unsegmented sequence of phonemes and has the task to predict the next phoneme in the stream.

An interesting extension to ‘feed-forward’ supervised models are those with recurrent connections from higher to lower layers (Figure 2). This apparently simple modification profoundly affects the processing characteristics of a model: now it can represent time and thereby process sequences. In the most common model, the Simple Recurrent Network (*SRN*; Elman, 1990), the hidden activation pattern for one input is presented to the input layer alongside the next input. Therefore, the way in which a model processes a specific input is affected by the previous input and becomes context dependent: the same input in two different contexts (i.e., with two different previous inputs and therefore two different hidden unit activation patterns) will lead to different activation patterns across the network. Recurrent models are presented with sequences of inputs and often the task of such a model is to predict the next item in the sequence (i.e., produce the next upcoming input on the output layer, before this input is actually presented to the model) – a task that is impossible to get correct all the time, but since predictability of the next item in a sequence often varies (such as in sentences: compare ‘She __’ and ‘She switched on the __’), the model’s prediction accuracy at each step is informative.

Unsupervised learning

In unsupervised learning there is no target for learning. Instead the model learns independently, from the environmental information. One way in which this can happen is through *Hebbian learning*. Here, a connection between two units is strengthened when both units are active at the same time. Hebbian learning is thus well suited for learning associations between stimuli. Variations of this process exist: for example, connections can be weakened when the two units are active at different times, or they can decay when no unit activation occurs.

A second type of unsupervised learning that is often used in models of word learning is the *self-organizing feature map* (Kohonen, 1998). Here, units are arranged on a two-dimensional map (Figure 3). Inputs are presented to an input layer and activation flows through connections to all units on the map. The unit on the map that is maximally responsive to a specific input then changes its incoming connection weights so that at the next presentation of the same input, it will respond even more strongly. Importantly, all units in a predefined radius or *neighborhood* around this winning unit also change their weights in a similar way. Thus, regions on the map become responsive to similar inputs. The update radius is gradually reduced during learning so that learning becomes progressively more fine-grained.

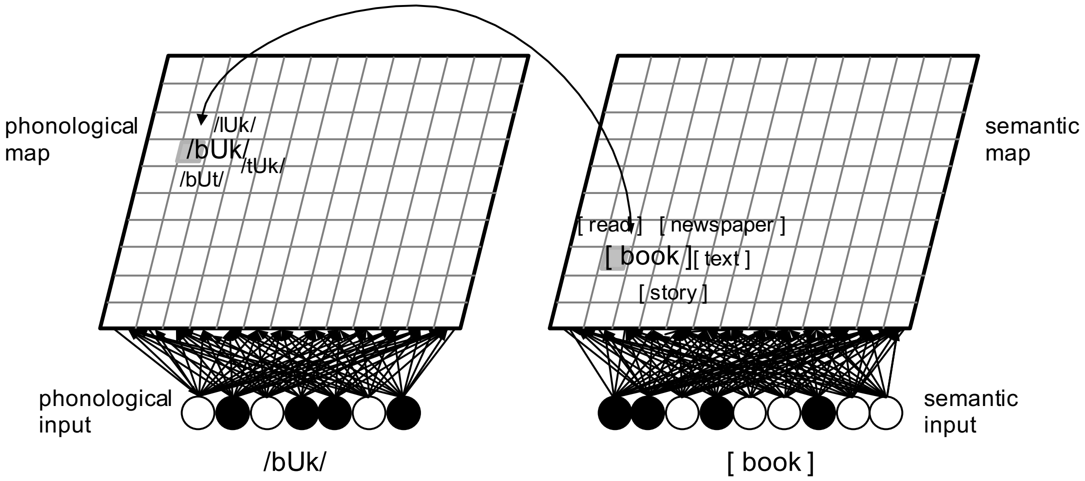

Figure 3: A model that consists of two self-organizing feature maps that are linked with Hebbian connections. In this illustrative example the model learns, through co-occurrence, the link between a word and its meaning. On the phonological map, similar-sounding words cluster together, whereas on the semantic map, concepts with overlapping meanings are clustered.

The final structure of a SOM reflects the statistical structure of the environment and similarity relationships between inputs. For example, a map that is trained on phonemes would usually develop one region for consonants and one for vowels (because of overlapping and distinct feature-based representations of both classes). Within each region, similarities between phonemes would be reflected, with similar phonemes located close together. This type of learning therefore only makes sense with distributed representations where similar inputs have overlapping activation patterns.

Self-supervised learning

Self-supervised learning falls between supervised and unsupervised learning. Here, the model is trained as in supervised learning, but it extracts the target without need for an external teacher. For example, in *auto-encoder* models the target is the same as the input: the model has to learn to reproduce its input on the output layer. This is a useful task because in order to do so, the model has to extract and represent regularities in the input. As a consequence the model also learns to generalize what it has encountered to new information in meaningful ways. For example, when familiarized with a category of items, the model can generalize this learned category when tested with a novel object (Mareschal & French, 2000). The activation patterns of the hidden layer in the network can then be examined to understand how the model internally represents such information. Typically, as in SOMs, similar inputs lead to similar activation patterns of the hidden units.

Several other types of ANN exist, but those described here capture the majority of models of the various stages of word learning.

We now turn to a description of the contributions that specific models have made to our understanding of various aspects of word learning. While we focus on ANNs, we also discuss related models which employ different frameworks to illustrate ways in which different modeling techniques can be used to simulate the same phenomenon, and consequently make different predictions about the mechanisms driving infants' learning.

Learning speech sounds

As described by Benders and Altvater-Mackensen (this volume), the first task of the language learner is to develop a phonemic repertoire of the native language. Computational models have simulated this process to investigate the links between perception and production and the role of parental reinforcement in this process. Westermann & Miranda (2004) directly addressed the interactions between the auditory and motor system in shaping a speech sound repertoire. The model consisted of two neural maps. A motor map contained neurons that activated a number of articulatory "muscles" in a speech synthesizer simulating human speech production. An auditory map was activated by heard vowel sounds. Both maps were linked with Hebbian connections that were strengthened when motor and auditory units were co-active. The model babbled by randomly generating motor settings and producing the resulting sound, and as a consequence the links between motor settings and their resulting sounds were reinforced. Importantly, activation flowing through the between-map links affected the representations on each map so that articulatory-auditory pairs that were produced with high reliability became prototypical. In this way, non-linearities in the articulation-sound mappings biased the model to preferentially produce and perceive certain sounds. The model also learned to adapt to an external language environment: speech sounds that were experienced in the environment selectively strengthened connections from the relevant auditory units to their associated articulatory settings. As a consequence, over time speech sounds produced by the model came to reflect to the speech sounds in the environment. The process implemented in this model provided a mechanistic explanation of the *articulatory filter hypothesis* (Vihman, 1993) according to which the sounds an infant itself produces are more salient to that infant than sounds not in its speech sound repertoire.

A shortcoming of the Westermann & Miranda model was that it did not account for the speaker normalization effect, in which different speakers' phonemes are perceived equivalently despite between-speaker variability (e.g., Bladon, Henton, & Pickering, 1984); rather, the model assumed that self-generated sounds were perceived in exactly the same way as sounds produced by external speakers. Other models which have aimed to account for this effect by integrating a parent's reinforcement demonstrate that characteristics of parental input play an important role in shaping infants' early vocalizations (Warlaumont, Westermann, Buder, & Oller, 2013, Yoshikawa, Asada, Hosoda & Koga, 2003).

Segmenting words from a continuous stream

As described by Junge (this volume), words do not occur in isolation in the child's environment; rather, they have to be segmented from a continuous auditory stream. One way in which this complex ability can be achieved is by exploiting the statistical regularities of language at different levels: differences in the probability of one phoneme following another can be reliable cues for word boundaries. On the level of whole words, transitional probabilities between different words can enable the detection of the grammatical class of a word.

The idea that phonotactic probabilities are cues to word boundaries was explored by Elman (1990) using the first SRN model. The model was trained on a simple artificial corpus of phoneme strings with no indication of word boundaries. The model saw the current phoneme as input and had the task of predicting the next phoneme in the sequence. Elman found that the model's prediction error (that is, the uncertainty about the next phoneme) was usually high at the beginning of words and then decreased within a word. Thus, as the model learned the phonotactics of language, peaks in network error coincided with word boundaries because they formed the least predictable instance within the language stream.

The seminal Elman (1990) model was subsequently improved upon, most notably by Christiansen and colleagues (Christiansen, Allen, & Seidenberg, 1998). Their model was also an SRN but used as input a real corpus of child directed speech which not only provided phonemic information but also relative lexical stress and utterance boundaries. While these three cues individually were not reliable indicators, when they were learned together the model could accurately predict word boundaries. Importantly, this model showed how the combination of accessible but unreliable statistical cues can together provide reliable cues for aspects of language for which there is no direct evidence in the input, and that an associative learner can extract this information from the language input. The focus on using actual language data on the statistical cues inherent in language – and on the powerful ability of statistical learners to extract and use this information – have been important drivers in the move away from the Chomskyan argument of the 'poverty of the stimulus' and the inevitability of domain-specific innate language abilities (Chomsky, 1957).

Mapping words to objects

An intuitive approach to learning word-object mappings is to imagine a pool of words and a pool of objects. Word learning then consists in establishing links between an element in the word pool and an element in the object pool (see Figure 3). Pioneered by Miikkulainen (1993, 1997), this approach has been directly instantiated in a number of models based on self-organizing feature maps where units on one map become linked to units on the other through Hebbian learning and has since been adopted by others (e.g., Mayor & Plunkett, 2010). The most advanced developmental word learning model based on linked feature maps so far is DEVLEX (Li, Farkas, & MacWhinney, 2004) and its extension DEVLEX II (Li, Zhao, & Mac Whinney, 2007). The DEVLEX models explored the effects of the detailed statistical properties of the input heard by children on their lexical development. DEVLEX II consisted of three linked SOMs. A phonological map received word forms that were based on phonetic feature vector representations. A semantic map contained semantic concepts derived from large corpora of language. Finally, an output sequence map learned to generate sequences of phonemes to produce words. The model was trained on word-object pairings so that representations formed on the respective maps. In parallel, links between the maps were trained with Hebbian learning to strengthen for co-occurring words, semantic concepts and phoneme sequences.

Word comprehension in the model was simulated by presenting a word to the phonological map. The maximally active unit on this map then propagated activation through the Hebbian links to the semantic map, activating a unit that represented a semantic concept. Production was modeled by propagating activation from the semantic to the phonological output map. Training data was based on real input to children, using 591 words from the MacArthur-Bates CDI (Fenson et al., 1993), including verbs, adjectives, nouns from different categories, and closed class words. Word meanings were represented by the distributional co-occurrence statistics of the target word in parental input to children (from CHILDES transcripts). These representations were learned and enriched gradually as learning progressed. During learning, words were presented to the model depending on their frequency in parental input.

DEVLEX II demonstrated the feasibility of simulating word comprehension and production in an associative model, modeling a range of phenomena observed in children's lexical development. First, it displayed a vocabulary spurt in the form of a phase of rapidly increasing word-meaning mappings after initial slow learning. This emerged from the simultaneous learning of organization within each map and connections between them: once a basic organization on the maps had been achieved, inter-map connections could be learned rapidly and accurately. Furthermore, the model showed a lag in word production relative to comprehension and individual differences between iterations of the model in the onset of the vocabulary spurt. These differences were linked to the specific learning experiences of the model, with exposure to short and frequent words leading to an earlier onset of the spurt.

During word production the model generated errors that are commonly found in word-learning children, such as leaving out final consonants (e.g., *ca* for *cat*) or consonants from consonant clusters (*mile* for *smile*), and substitution of consonants (*birb* for *bird*). These errors arose from incomplete sequence learning on the output map and incomplete links from meaning to words. Importantly, then, as well as capturing children's word learning trajectories DEVLEX II also offered a mechanism for the errors they make during this process.

DEVLEX II nicely illustrated how a model can be seen as an artificial learner embedded in the same environment as a developing child: by simulating a realistic learning environment (within the confines of a disembodied learner), it provided insights into how the precise structure of the child's experience can shape the learning process. Subsequently DEVLEX II has also been applied to bilingual word learning (Zhao & Li, 2010) to investigate how differential onset of the two languages affects structure and interaction of phonological and semantic representations on the respective maps.

A different approach to learning word-object mappings was taken in Westermann & Mareschal, 2014 (for an earlier related model see Plunkett, Sinha et al., 1992). In feature map based models the representations that develop on each map are unaffected by the links between the maps. Nevertheless it is possible that the different aspects of an object representation – visual appearance, auditory and functional features, the object name etc., become integrated so that different features can affect each other. For example, research with adults suggests that objects that share a name are perceived as more similar than the same objects if they do not share a name (Lupyan, Rakison, & McClelland, 2007). In development it has been found that labels affect how infants represent visual objects: they group together objects that share a common label and separate similar objects that have different names (Althaus & Westermann, 2016; Gliozzi, Mayor, Hu, & Plunkett, 2009; Plunkett, Hu, & Cohen, 2008). Therefore, Westermann and Mareschal (2014) modeled how developing mental representations can be affected by common labels using an auto-encoder neural network. The way in which representations of different objects relate to each other can be assessed by examining the activation profiles of the hidden units: activation patterns for objects perceived as similar will cluster together (see also Rogers & McClelland, 2004).

Westermann & Mareschal (2012, 2014) provided their model with feature-based representations of 26 different object categories from four superordinate categories. When trained without language, the model developed object representations that were based on the visual similarity between objects. But when the model was enhanced by units encoding the category name for an object, the representational space in the model became warped to that objects with different labels became more dissimilar.

This and other models also address the issue of the status of labels in early word learning. Two contrasting theoretical standpoints have been put forward. According to one, labels are qualitatively separate from the perceptual representation of objects and refer to these objects (Waxman & Gelman, 2009). This viewpoint is expressed in models that have separate maps for labels (e.g., Mayor & Plunkett, 2010; Li et al, 2007). According to another theory, labels, at least in very early word learning, are mere features of objects at the same representational level as other perceptual features. This viewpoint was instantiated in a model by Gliozzi et al (2009) in which visual features and object labels fed into a single map that developed holistic object representations. The Westermann & Mareschal (2014) model implemented a third view: here, labels were separate from visual object descriptions but through learning became closely integrated with them, leading to an object representation that took account of visual similarity modulated by the label. The status of labels in object representations remains a topic of ongoing research (e.g., Deng & Sloutsky, 2015) and predictions made by different models will likely be able to advance our understanding about this aspect of word learning.

Word-object mapping: hypothesis testing or association?

The discussed models of learning word-object mappings all have assumed that the mapping to be learned is unambiguous: at each time, there is only one object and one word present from which to learn. While this simplification has been useful to further our understanding of lexical structure and the learning mechanisms involved, in the real world a word learning situation is often more ambiguous with several possible referents for a heard word (see Monaghan, Kalashnikova & Mattock, this volume). While it has become clear that infants can track the co-occurrence probabilities between words and objects across learning situations and therefore resolve this ambiguity there has been controversy about the mechanism underlying this ability.

One view argues that infants have implicit, relatively sophisticated *a priori* hypotheses about co-occurrence statistics and the probability that a word refers to a specific object, and that they test and confirm or reject these hypotheses based on some inference procedure, making a probabilistically optimal word-object mapping. This mechanism has been implemented in probabilistic, Bayesian models. These models are based on prespecified probability distributions that determine the model's "decision-making" process; in this case, probabilities of words mapping to a particular referent. For example, Xu & Tenenbaum (2007) presented a Bayesian model which captured 4-year-old children's novel category label generalization in an empirical study which manipulated whether children themselves or the experimenter selected exemplar objects during training. The authors argued that the behavior seen in their empirical study cannot be accounted for by associative learning since their manipulation did not alter the statistical structure of the learning environment – and if all an associative learner does is to extract the statistics from the environment this would not lead to different learning outcomes for these two conditions. Instead, they argued, their empirical results depended on a process of hypothesis testing. Specifically, the difference in training experience between the two groups (i.e., experimenter demonstration vs. independent choosing) had shaped children's hypotheses about the possible referents of the label, producing the contrasting generalization patterns seen at test. More generally, since the model was equipped with predetermined prior probabilities, this reflects a learning situation in which children have substantial prior knowledge, without, however, addressing where this prior knowledge may have come from.

An opposing view argues that correct word-object mappings (and the prior knowledge necessary to learn them) can be formed through associative learning. Yu (2008) presented a simple statistical associative learner that counted the co-occurrences of labels and their referents across learning events and calculated probabilities from these. The model received real linguistic input, consisting of a corpus of transcribed speech elicited from parents in a storybook narration task, and visual input, consisting of a list of the objects visible in the storybook at the time a given utterance was made. For each of these "scenes", the mapping between words and objects was ambiguous, and only around 5% of the co-occurrences were "correct". The model was tested by interrogating word-referent association probabilities calculated across the entire corpus. Like children, the model learned to associate words with their correct referents with a high degree of accuracy. Further, a second model, which could use its existing lexical knowledge to support the acquisition of new words, exhibited a vocabulary spurt. These models made important predictions about infants' word learning; in particular, that word learning is an incremental process, with word-object mappings starting out weak, but becoming stronger with experience. This "partial knowledge" account of word learning has recently been empirically tested and supported (Yurovsky, Fricker, Yu, & Smith, 2014), illustrating how models, with their explicit specification of mechanism, can shed light on real-world learning.

Taken together these two models illustrate how modeling can force us to be absolutely clear what pre-existing knowledge and cognitive structures are necessary for learning. For example, on any given learning event both Yu's (2008) model and that of Xu and Tenenbaum (2007) make word-object mappings based on prior association probabilities. The critical difference is that in Bayesian models, those prior probabilities are determined *a priori* by the modeler, while in the associative model, prior probabilities are learned from naturalistic input. Thus, Yu's model relied on a probabilistic mechanism, but did not require the complex inferential processes or built-in knowledge central to Bayesian methods. More generally such differences between Bayesian and associative models once again speak to the core-versus-learned knowledge debate, with the former assuming native knowledge and the latter assuming knowledge can be learned from the rich statistics of the environment.

Models are therefore uniquely positioned to make important contributions to theory: modeling forces us to specify assumptions not only about the processing mechanisms, but also about the representation and availability of data to the learner.

Timescales of language development

As described by Horst (this volume), from around 18 months, when infants are shown an array of two known and one novel object and are asked for a novel label (e.g., “Which one is the dax?”) they often select the correct referent, i.e., the novel object. Traditionally this ability has been explained by intrinsic constraints such as mutual exclusivity (the knowledge that an object only has one name; e.g., Markman, 1991). McMurray, Horst & Samuelson (2012) described a computational model in which referent selection arises from real-time competition between referents instead of such higher-level inferential processes. This model also addressed the finding that even when children select the correct referent, they often do not show long-term retention of the label-referent mapping (Horst & Samuelson, 2008). Specifically, the model demonstrated that while the mapping problem can be solved in-the-moment, this online association leads to only minimal strengthening of the word-object connection; to learn a robust word-object association takes many encounters of the same mapping. Thus, this model emphasizes the importance of cross-situational associative learning to capturing the real behavior demonstrated by infants in empirical studies of word learning.

Overall, by providing a mechanistic, low-level explanation of referent selection, McMurray et al. (2012) showed that observed behavior in children does not have to rely on high-level inferential processes (see also Twomey, Morse, Cangelosi, & Horst, 2016). As in Yu (2007) in this work an apparently complex behavior can emerge from the interaction between two timescales of associative word learning: in-the moment referent selection, and cross-situational learning.

Summary

In this brief chapter it has been impossible to provide an exhaustive overview of computational models of word learning. First, a number of other neural network models not covered here have addressed different aspects of word learning (e.g., Colunga & Smith, 2000; Regier, 2005). Second, while we have mentioned two non-connectionist modeling approaches to word learning in Xu and Tenenbaum’s (2007) Bayesian model and Yu’s (2008) associative learner, there remain a range of other informative formal approaches to understanding word learning. One such approach consists in pure mathematical modeling. Notable here is McMurray’s (2007) work on the vocabulary spurt, which demonstrates that the patterns of vocabulary acquisition commonly observed in children can be captured by a simple learning system situated in a structured learning environment without internal changes to the system that accelerate learning. Further, semantic network approaches have demonstrated that the age at which a word is acquired is influenced by the density of a word's semantic network and the diversity in the linguistic contexts in which this word typically occurs (Hills, Maouene, Riordan, & Smith, 2010). Equally, Dynamic Neural Field models, a type of model related to the connectionist approach that focuses on how learning can occur on the basis of interactions between neural excitation and inhibition in large networks, have successfully captured a number of phenomena in early word learning (e.g., Samuelson, Kucker, & Spencer, 2016; Samuelson et al., 2009; Samuelson, Smith, Perry, & Spencer, 2011).

Despite progress in these modeling approaches, currently neural network models have arguably made the strongest contribution to our understanding of the mechanisms of the development of word learning, providing explicit mechanistic accounts of often surprising phenomena, and generating predictions that are subsequently captured in empirical studies with infants and children. Their strength lies in the ability to form complex associations (between words and objects, motor and auditory representations and so on) and to learn these multimodal representations from experience, thereby showing sensitivity to the statistical structure of their learning environment and the specific experiences to which they are exposed. In this way these models have shown how the richness of the stimulus can overcome the need for innate learning biases and how specific learning trajectories can be explained by interactions between domain-general learning mechanisms and the precise structure of the learner's environment.

Acknowledgements

The writing of this chapter was supported by the International Centre for Language and Communicative Development (LuCiD; [ES/L008955/1]). The support of the Economic and Social Research Council is gratefully acknowledged. GW was further supported by a British Academy/Leverhulme Trust Senior Research Fellowship (SF150163), and KT by an ESRC Future Research Leaders fellowship [ES/N01703X/1].

References

Althaus, N., & Mareschal, D. (2013). Modeling cross-modal interactions in early word learning. *IEEE Transactions on Autonomous Mental Development,*, *5*(4), 288–297.

Althaus, N., & Westermann, G. (2016). Labels constructively shape object categories in 10-month-old infants. *Journal of Experimental Child Psychology*. https://doi.org/10.1016/j.jecp.2015.11.013

Aslin, R. N., Woodward, J. Z., LaMendola, N. P., & Bever, T. G. (2006). Models of word segmentation in fluent maternal speech to infants. In Morgan & Demuth (Eds.). In *Signal to Syntax: Bootstrapping From Speech To Grammar in Early Acquisition.* New York: Taylor & Francis.

Bladon, R. A. W., Henton, C. G., & Pickering, J. B. (1984). Towards an auditory theory of speaker normalization. *Language & Communication*, *4*(1), 59–69. https://doi.org/10.1016/0271-5309(84)90019-3

Chomsky, N. (1957). *Syntactic structures*. Mouton: The Hague.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*(2–3), 221–268.

Colunga, E., & Smith, L. B. (2000). Committing to an ontology: A connectionist account. *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*, 89–94 1075.

Davies, M. H. (2003). Connectionist modelling of lexical segmentation and vocabulary acquisition. In Philip Quinlan (ed.), *Connectionist models of development: Developmental processes in real and artificial neural networks*, pp. 151–187. Hove: Psychology Press.

Deng, W. S., & Sloutsky, V. M. (2015). Linguistic labels, dynamic visual features, and attention in infant category learning. *Journal of Experimental Child Psychology*, *134*, 62–77. http://doi.org/10.1016/j.jecp.2015.01.012

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., … Reilly, J. S. (1993). *The MacArthur Communicative Development Inventories: User's Guide and Technical Manual.* San Diego: Singular Publishing Group.

Gliozzi, V., Mayor, J., Hu, J. F., & Plunkett, K. (2009). Labels as features (not names) for infant categorization: A neurocomputational approach. *Cognitive Science*, *33*(4), 709–738. https://doi.org/http://dx.doi.org/10.1111/j.1551-6709.2009.01026.x

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of Memory and Language*, *63*(3), 259–273. https://doi.org/10.1016/j.jml.2010.06.002

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, *13*(2), 128–157. https://doi.org/Doi 10.1080/15250000701795598

Kohonen, T. (1998). The Self-Organizing Map, a possible model of brain maps. *Brain and Values*, 207–236 568.

Li, P., Altmann, G., Hare, M., McRae, K., & Plunkett, K. (2007). Rumelhart Symposium: Language as a Dynamical System: In Honor of Jeff Elman. In *Proceedings of the Cognitive Science Society* (Vol. 29). Retrieved from http://escholarship.org/uc/item/4rf599q9.pdf

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks*, *17*(8), 1345–1362.

Li, P., Zhao, X., & Mac Whinney, B. (2007). Dynamic self-organization and early lexical development in children. *Cognitive Science*, *31*(4), 581–612.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking - Redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077–1083. http://dx.doi.org/10.1111/j.1467-9280.2007.02028.x

Mareschal, D., & French, R. (2000). Mechanisms of categorization in infancy. *Infancy*, *1*(1), 59–76. https://doi.org/10.1207/S15327078IN0101_06

Mareschal, D., & Thomas, M. S. C. (2007). Computational modeling in developmental psychology.*, IEEE Transactions on Evolutionary Computation*, *11*(2), 137–150.

Markman, E. M. (1991). The whole-object, taxonomic, and mutual exclusivity assumptions as initial constraints on word meanings. In S. A. Gelman & J. P. Brynes (Eds.), *Perspectives on language and thought: Interrelations in development.* (pp. 72–106). New York: Cambridge University Press.

Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological Review*, *117*(1), 1–31. https://doi.org/10.1037/a0018130

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, *317*(5838), 631–631. https://doi.org/DOI 10.1126/science.1144073

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), 83877. https://doi.org/doi: 10.1037/a0029872

Miikkulainen, R. (1993). *Subsymbolic natural language processing: An integrated model of scripts, lexicon, and memory*. MIT press.

Miikkulainen, R. (1997). Natural language processing with subsymbolic neural networks. *Neural Network Perspectives on Cognition and Adaptive Robotics*, 120–139.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, *38(11)*, 39-41.

Munakata, Y., & McClelland, J. L. (2003). Connectionist models of development. *Developmental Science*, *6*(4), 413–429.

Plunkett, K., Hu, J. F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*(2), 665–681. https://doi.org/10.1016/j.cognition.2007.04.003

Quinlan, P. T. (2003). *Connectionist models of development: Developmental processes in real and artificial neural networks*. Taylor & Francis.

Räsänen, O. (2011). A computational model of word segmentation from continuous speech using transitional probabilities of atomic acoustic events. *Cognition*, *120*(2), 149–176. https://doi.org/10.1016/j.cognition.2011.04.001

Regier, T. (2005). The Emergence of Words: Attentional Learning in Form and Meaning. *Cognitive Science*, *29*(6), 819–865. https://doi.org/10.1207/s15516709cog0000_31

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Samuelson, L. K., Kucker, S. C., & Spencer, J. P. (2016). Moving word learning to a novel space: A dynamic systems view of referent selection and retention. *Cognitive Science 41*(S1), 52-72.

Samuelson, L. K., Schutte, A. R., & Horst, J. S. (2009). The dynamic nature of knowledge: Insights from a dynamic field model of children's novel noun generalization. *Cognition*, *110*(3), 322–345. https://doi.org/10.1016/j.cognition.2008.10.017

Samuelson, L. K., Smith, L. B., Perry, L. K., & Spencer, J. P. (2011). Grounding word learning in space. *PloS One*, *6*(12), e28095.

Twomey, K. E., Morse, A., Cangelosi, A., & Horst, J. (2016). Children's referent selection and word learning: insights from a developmental robotic system. *Interaction Studies 17*(1), 93 – 119. http://doi 10.1075/is.17.1.05two.

Vihman, M. M. (1993). Variable paths to early word production. *Journal of Phonetics*, *21*(1–2), 61–82.

Warlaumont, A. S., Westermann, G., Buder, E. H., & Oller, D. K. (2013). Prespeech motor learning in a neural network using reinforcement. *Neural Networks*, *38*, 64–75. https://doi.org/10.1016/j.neunet.2012.11.012

Waxman, S. R., & Gelman, S. A. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, *13*(6), 258–263. https://doi.org/10.1016/j.tics.2009.03.006

Westermann, G., & Mareschal, D. (2014). From perceptual to language-mediated categorization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120391.

Westermann, G., & Miranda, E. R. (2004). A new model of sensorimotor coupling in the development of speech. *Brain & Language*, *89*(2), 393–400.

Westermann, G., & Plunkett, K. (2007). Connectionist models of inflection processing. *Lingue E Linguaggio*, (2/2007). https://doi.org/10.1418/25655

Westermann, G., Ruh, N., & Plunkett, K. (2009). Connectionist approaches to language learning. *Linguistics*, *47*(2), 413-452. DOI: 10.1515/LING.2009.015Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Development*, *4*(1), 32–62. https://doi.org/10.1080/15475440701739353

Yoshikawa, Y., Asada, M., Hosoda, K., & Koga, J. (2003). A constructivist approach to infants' vowel acquisition through mother-infant interaction. *Connection Science, 15*, 245–258.

Yurovsky, D., Fricker, D. C., Yu, C., & Smith, L. B. (2014). The role of partial knowledge in statistical word learning. *Psychonomic Bulletin & Review*, *21*(1), 1–22.

Zhao, X., & Li, P. (2010). Bilingual lexical interactions in an unsupervised neural network model. *International Journal of Bilingual Education and Bilingualism, 13:5*, 505-524, DOI: 10.1080/13670050.2010.488284

Zorzi, M., Testolin, A., & Stoianov, I. P. (2013). Modeling language and cognition with deep unsupervised learning: a tutorial overview. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00515