Neural network models of language acquisition and processing

Stefan L. Frank[1], Padraic Monaghan[2], and Chara Tsoukala[1]

[1]Centre for Language Studies, Radboud University Nijmegen
[2]Department of Psychology, Lancaster University

1. Neural networks in cognitive science

Artificial neural network models (also known as *Parallel Distributed Processing* or *Connectionist* models) have been highly influential in cognitive science since the mid-1980s. The original inspiration for these systems comes from information processing in the brain, which emerges from a large number of (nearly) identical, simple processing units (neurons) that are interconnected into a network. Each unit receives activation from other units or by stimulation from the external world, and generates an output activation that is a function of the total input activation received. The unit then feeds the output activation onward to the units to which it is connected. Information processing is thus implemented in terms of activation flowing through this network.

Each connection between two units has a weight that determines how strongly the first unit affects the second. These weights can be adapted, which constitutes learning, or "training" as it is commonly known in the neural network literature. Algorithms for network training can be roughly divided into *supervised* and *unsupervised* methods. Supervised training is applied when a specific and known input-to-output mapping is required (e.g., learning to transform orthographic to phonological representations). To accomplish this, the network is provided with a representative set of "training examples" of inputs and the corresponding target outputs. It then processes each example and the difference between the networks' actual output and the target output leads to an update of the connection weights such that, next time, the output error will be smaller. By far the best known and most used method for supervised training is the Backpropagation algorithm (Rumelhart, Hinton, & Williams, 1986) that makes the network's output activations for the training examples gradually converge toward the target outputs. Unsupervised training, in contrast, makes the network adapt to (aspects of) the statistical structure of input examples without mapping to target outputs (e.g., discovery of regularities in the phonological structure of language). These networks are well-suited to uncovering statistical structure present in the environment without requiring the modeller being aware what the structure is. One well-known example of an unsupervised training method is the learning rule proposed by Hebb (1949): Strengthen connections between units that are simultaneously active and weaken the connections between two units if only one is active.

In spite of the superficial similarities between artificial and biological neural networks (i.e., interconnectivity and stimulation passing between neurons to determine their activation, and learning by adaptation of connection strengths), these cognitive models are not usually claimed to simulate processing at the level of biological neurons. Rather, neural network models form a description at Marr's (1982) algorithmic level, that is, they specify cognitive representations and operations while ignoring the biological implementation.

Neural networks underwent a surge of popularity in the 1990s, but from the early 21st century they were somewhat overshadowed by symbolic probabilistic models. However, neural networks have enjoyed a recent revival partly due to the success of "deep learning" models, which display state-of-the-art performance on a wide range of artificial intelligence tasks (LeCun, Bengio, & Hinton, 2015). For the most part, the field of cognitive modelling is still to catch up with these novel

developments. Consequently, the currently most influential connectionist cognitive models are of the more traditional variety. We return to this issue in the Conclusion.

1.1. Feedforward and recurrent networks

Connectionist models are not amorphous networks in which everything is connected to everything else. Rather, a particular structure is imposed, for example by grouping units into a number of layers and allowing activation to flow only from each layer to the next. The first layer receives inputs from the environment, the final layer produces the corresponding output, and any intermediate layer is known as "hidden". Although this so-called "feedforward" architecture can (at least in theory) approximate any computable input-to-output function, it is unable to handle input that comes in over time. This is because the network has no working memory: Each input is immediately overwritten by the next. Hence, the feedforward network is not the most appropriate model for simulating language processing, which is a fundamentally temporal phenomenon.

Elman (1990), in his seminal paper "Discovering structure in time", proposed a solution: Include a set of *recurrent* connections with trainable weights that link each unit of the single hidden layer to all hidden-layer units. Consequently, the hidden layer receives both the current environmental input and its own previous activation state which, in turn, depends on the state before that, etc. In this manner, the model is equipped with a working memory and can therefore encode sequential information, or "structure in time", making it well-suited to processing language as it unfolds over time. This particular architecture became known as the Simple Recurrent Network (SRN) but forms part of a larger class of Recurrent Neural Networks (RNNs) that have connections through which (part of) the network's current activation feeds back to the network itself.

1.2. Neural network models and linguistic theory

Connectionist models of language acquisition and processing offer a view of the human language system that is very different from traditional, symbolic models in cognition. For one, neural networks do not distinguish competence (i.e., language knowledge) from performance (i.e., language behaviour). Instead, knowledge becomes instantiated in network connection weights in order for the network to display particular performance. In a sense, it forms procedural rather than declarative knowledge: it is know-*how*, not know-*that*. Hence, there is no way for the network to assess its own knowledge. As Clark and Karmiloff-Smith (1993, p. 495) put it: "it is knowledge *in* the system, but it is not yet knowledge *to* the system."

Second, language researchers from the nativist tradition have famously argued that infants must possess innate, language-specific knowledge or learning mechanisms, because otherwise language acquisition in the absence of negative evidence would be impossible (e.g., Chomsky, 1965; Gold, 1967; Pinker, 1989; among many others). In contrast, empiricists claim that language acquisition requires only domain-general mechanisms. Connectionism falls squarely into the empiricist camp because the representations and learning mechanisms built into neural networks are not specific to language and the networks receive no negative evidence during training. Hence, successful neural network learning of (relevant aspects of) syntax would undermine the nativist position.

A third major difference with traditional linguistic thinking is that neural networks do not represent discrete categories (be it phonemes, words, parts-of-speech, or any other category) unless these are explicitly assigned to the network's units a priori. However, in most (and, arguably, the most insightful) models, representations are learned in the hidden layer(s) rather than assigned,

which results in “fuzzy” categories. To the extent these learned representations explain (psycho)linguistic phenomena, neural network models reject a causal role for explicit representation of typical linguistics constructs. The alternative proposed by connectionism is that apparently symbolic behaviour emerges from statistical regularities present in the language and in the mapping between form and meaning. By capturing these regularities, neural networks are able to account for a range of phenomena in human language acquisition, comprehension, and production.

Neural network models and their ability to discover structure in the environment have enabled researchers to test the mechanisms required for many aspects of language behaviour. For instance, such models have been applied to determine whether language structures need to be pre-specified before exposure to language, or whether general-purpose statistical processes, embedded in neural networks, are sufficient for reflecting human performance. Further, as language corpora have become more and more representative of language learning environments, neural network models have provided insight into the sources of information present within the communicative environment that can contribute to learning and processing. In the following sections, we review key areas of language behaviour where neural networks have been applied to explore environmental features that are useful for language learning, comprehension, and production.

2. Word learning

In order to learn a spoken word, the listener has to be able to identify its phonological form, isolate it from continuous speech, and relate the word’s form to its meaning in the environment. Each of these tasks is hugely challenging for the learner. Neural network models have been used to illustrate the inherent difficulties of these tasks, as well as to test potential solutions that the learner may bring to the situation to support acquisition, and subsequent online processing of words in speech.

2.1 Identifying words from continuous speech

Neural network models have been useful in uncovering the statistical properties of speech input that can contribute to speech segmentation. Elman (1990) investigated how statistical transitional information from sequences of phonemes that are present in child-directed speech could provide clues to word boundaries. His SRN model had as input a sequence of phonemes from an artificial corpus designed to mimic child-directed speech, and was trained to predict the next phoneme in the sequence. This SRN demonstrated that transitions between phonemes within words tend to be much more predictable than transitions across word boundaries. Thus, the model was able to detect the unpredictable nature of phoneme transitions at word boundaries, and predicted future behavioural studies on infants’ sensitivity to transitional probabilities between syllables in detecting potential words in continuous speech (Saffran, Aslin, & Newport, 1996). Cairns, Shillcock, Chater, and Levy (1997) demonstrated that a similar model was able to scale up to detect word boundaries, as well as learn phonotactic regularities of syllable structure within words, from a transcribed corpus of natural language child-directed speech.

These neural network models of segmentation also indicated limitations of characterising the environment merely in terms of transitions between phonemes in speech, providing further insight into the process of early speech segmentation in language acquisition. One limitation was that transitional probabilities alone provide a glass ceiling for overall accuracy of segmentation. Gambell and Yang (2003) showed that transitional probabilities between syllables in child-directed speech, though able to account for segmentation at a rate significantly greater than chance, still missed very

many word boundaries. So, some aspect of processing or of the environment was lacking from this modelling approach. To address this, Christiansen, Allen, and Seidenberg (1998) employed an SRN with natural language input, adding additional information about utterance boundaries and stress position within words. They found that these combined cues outperformed a model that used only one of these individual cues, indicating that a language learning system benefits from sensitivity to combined cues for segmentation that are present in the speech environment. EEG studies of speech segmentation have shown relations between these proposed prosodic and statistical cues as predicted by computational models, demonstrating human sensitivity to these cues consistent with neural processing of sequence prediction (e.g., Cunillera, Toro, Sebastián-Gallés, & Rodríguez-Fornells, 2006).

These models have been extremely successful in illuminating the information that is present in the speech environment and that is also extractable using simple statistical processes, potentially similar to those employed by the human language learner. But the models are also chronically limited by the fact that they do not construct a vocabulary. This means that they are unable to process ambiguities in speech, limited as they are by very local statistical structure. A classic example is the pair “recognise speech” and “wreck a nice beach” (Fosler-Lussier, Amdal, & Kuo, 2005), which would be indistinguishable in these simple, local-information approaches. A model with a vocabulary, with expectations (or priors) about individual and combined probabilities of combinations of words, would be able to more accurately segment continuous speech (Kamper, Wang, & Livescu, 2016; Monaghan & Christiansen, 2010; Philips & Pearl, 2015).

The TRACE model (McClelland & Elman, 1986) of auditory word recognition provided such a link between top-down lexical information with bottom-up statistical information about relations between acoustic features of speech and phonemes (Allopenna, Magnuson, & Tanenhaus, 1998). TRACE has recently been adapted to incorporate additional information from the visual environment to simulate multiple cue combination in speech processing (Smith, Monaghan, & Huettig, 2017). However, the TRACE model was not designed to process sequences of words embedded within continuous speech. To address this requirement to acquire lexical items online and simultaneously apply them to segment speech, French, Addyman, and Mareschal (2011) constructed a neural network model (TRACX) that learned to construct and identify chunked sequences of phonemes from continuous speech, using backward transitional probabilities to initially detect the word candidates from the language. These were then used for recognition of sequences of encountered speech, rather than, as with the SRN models, for prediction of the next phoneme in the sequence. The model was not only shown to outperform SRN models, but also provided a framework for how a single statistical processing system can both generate and utilise word candidates from continuous speech.

These modelling approaches have thus shown the value of individual and combined cues present in natural language contributing to speech segmentation, turning a notoriously difficult problem into a task substantially more tractable. Furthermore, they highlight the kinds of computations that the learner must develop – and the processing interface between online, speech processing and the statistical probabilities that have to be encoded within a growing vocabulary, to move toward a highly accurate and fast speech segmentation system. Though these approaches have tended to take an abstracted input, ignoring phonetic variability and noise in the acoustic signal, adding such information provides challenges but also opportunities for additional cues to be discovered and derived from the input. Bayesian approaches to discovery of word-level acoustic forms using unsupervised training demonstrate the possibility of this in artificial computational systems (Kamper, Jansen, & Goldwater, 2016). One such example is increased word-final phoneme

duration that is a further useful, and useable, cue for speech segmentation (e.g., Scharenborg, 2010) and that could be incorporated into a neural model of speech processing.

2.2 Mapping words to meaning

A second task for word learning in language acquisition is to determine the referent of a word, once it has been isolated from continuous speech. This task is famously difficult to accomplish, because an utterance tends to contain several words, only one of which is referring to a particular aspect of the environment, and because the environment contains an infinite number of potential referents for each word (Quine, 1960) – the word "pig" could refer to the small cuddly toy that a parent is waving in a child's face, but it could also refer to the action of waving, to the softness of the material of the toy, the colour, its general shape, or to the communicative act itself. Hence, the task can be conceived as, at the very least, a many-to-many mapping problem.

Similar to the models of speech segmentation, there are numerous models of word-meaning acquisition which provide abstract characterisations of the task, with the intention to highlight the environmental properties that enable this mapping task to be accomplished. These models explore the information present around the child acquiring her language in resolving this link between word and world. More recently, the models have become more concretely related to neural processing, in order to incorporate known perceptual and attentional constraints that, paradoxically perhaps, demonstrate even greater tractability of the problem (Regier, 2003; Yurovsky, Smith, & Yu, 2013).

Associations between occurrences of words and occurrences of their intended referent have been proposed as a mechanism driving the formation of word to meaning mappings (Plunkett, 1997). Yu and Ballard (2007) implemented a model of associations between words and objects that was exposed to small corpora of child-directed speech, where objects in the environment had also been encoded in the corpus transcription. When a particular word was spoken in the presence of a particular object, the strength of association between the word and the object was increased. They found that the strongest associations were often between words and the target object, such as the words "oinko" and "pig" being strongly associated with a pig soft toy in the child's environment.

Yu and Ballard's (2007) model was further tested by adding in additional cues that constrained either the referring word, or the referent in the environment. They included prosodic information about words that were emphasised by being prosodically distinct from the rest of the utterance, and increased the possibility of strengthening associations for those words. They also included information about objects in the environment for which the speaker and the child held joint attention during the utterance, again increasing the association strength for the attended object. Each of these information cues improved the model's performance further. The model was able to exploit cross-situational correspondences between words and objects in the environment, such that from a single exposure to a word and an object, the relation between the word and referent is not evident (because of the presence of other words and environmental stimuli at the same time), but over multiple instances the likelihood of the object being named when present increases. Such cross-situational information was shown to be used by infants in acquiring word-meaning mappings (Smith & Yu, 2008), supporting the computational models suggesting that associationism may underlie the acquisition of word meanings.

This associationist modelling approach, however, has been criticised by researchers who propose instead that children do not learn by association between words and environmental properties, but rather that children propose and self-test hypotheses for a possible word-meaning link. Yu and Smith (2012) compared both of these computational approaches and found that an

associationist model provided an adequate fit to the data, and that even apparent complexities that seem to suggest hypothesis testing (e.g., children applying strategies such as mutual exclusivity: using a novel word for an unknown object appearing among known objects) can be accounted for within an associationist model. Similarly, apparent developmental differences between children and adults in terms of whether they employ an associative approach, or a strategic, referential approach can also be accounted for by a model with a single associative mechanism (Fazly, Alishahi, & Stevenson, 2010).

However, other computational models have suggested that combinations of associative and strategic word learning mechanisms are required to adequately describe performance. McMurray, Horst, and Samuelson (2012) proposed a hybrid network model with two components. The first was an associative network that learned mappings between words in an utterance and objects in the environment with a very slow learning rate, and was able to acquire word-object mappings from cross-situational statistics. The second component was a quick-responding referent selection mechanism that was able to strategically select a potential referent for a word, under circumstances where children can demonstrate fast mapping between words and potential referents, such as conditions of mutual exclusivity. The model predicted that word learning would be more robust from multiple presentations of a word, but less stable when learning is inaugurated under conditions of fast-mapping. Twomey, Ranson, and Horst (2014) subsequently confirmed these predictions of the model in children learning novel words: fast mapping and associative learning resulted in immediate word-object mappings, but only associative learning resulted in long-term learning.

Additional attentional and perceptual constraints have recently been included in these models of associative word learning, to align the referent selection mechanism with the cognitive processes known to impact on visual object processing. Samuelson, Kucker, and Spencer (2017) showed how this can further improve performance on word learning, consistent with approaches that closely describe the actual multimodal input that children experience (Yurovsky et al., 2013). Furthermore, such associative models are able to generalise properties of the environment to learn category-level information, accounting for children's distinctions in learning names for shapes and names for materials (Colunga & Smith, 2005), and learning individual and category labels (Mayor & Plunkett, 2010). Such approaches provide a valuable bridge between a long tradition of neural network modelling in perception and attention (e.g., Mozer, 1991) embedding our understanding of language learning models within broader, domain-general processes.

3. Syntactic development

3.1. Lexical category learning

How children acquire category knowledge about words has been a hotly-debated issue, at the heart of the nature-nurture debate in language acquisition. Nativists assume that the lexical categories, or semantic features to which these categories relate, are innately specified in the learner, and individual words' membership to these categories is then determined from exposure to the language. Alternatively, empiricist approaches propose that the categories themselves are acquired as a result of experience, without the need for innately specified structure in advance of exposure to the language. Neural network modelling approaches have been key to determining the extent to which such innate knowledge is required in advance to simulate the acquisition of lexical categories within the language, and also, relatedly, have explored the type of computational mechanisms that might be able to drive the discovery of these categories.

Elman (1990) applied the same SRN approach as used for the speech segmentation task to prediction of words in utterances, again using a small artificial corpus. The model's input was a sequence of words in the utterance and the model had to predict (i.e., generate as output) the next word in the sequence. Utterances were constructed from one of four sentence frames, and there were 29 different words that were either animate or inanimate nouns, or transitive or intransitive verbs. After training, the hidden unit states of the model were investigated to determine whether it had learned to reflect the lexical categories in the language in its internal structure, resulting only from the statistical sequential information from the sentences. A hierarchical cluster analysis revealed distinctions in the model's representations of nouns versus verbs. Furthermore, clusters within the noun category were also evident, with animate and inanimate nouns forming distinct groups. More recently, Mikolov, Chen, Corrado, and Dean (2013) showed that recurrent and feedforward networks can be trained on very large natural language corpora and that the resulting connection weights from input units display semantic/syntactic word clustering, similar to what was demonstrated by (Elman, 1990) but on a much more realistic scale. Thus, apparent semantic features that relate to syntactic properties (such as nouns versus verbs, as well as features within a syntactic category, such as animacy) were detectable from applying generic distributional statistical learning to sequences of words.

The extent to which general-purpose learning of lexical categories scaled up to a more realistic natural language environment was tested computationally by Redington, Chater and Finch (1998). Using a large corpus of English child-directed speech they measured co-occurrences between pairs of words that were either adjacent or separated by one word. Cluster analysis was then applied to these counts of associations, such that words with similar patterns of co-occurrences should be clustered together. The results were spectacular, demonstrating for the first time, with large-scale corpora, that lexical categories were extractable with accuracy using simple statistical processes applied to natural language exposure (see Kiss, 1973, for a smaller-scale version, and Schütze, 1993, for large-scale use of distributional information for lexical clustering in a language engineering context). As with Elman's (1990) computational investigation, the clusters in the co-occurrence model respected broad syntactic category distinctions, such as between nouns and verbs, but also reflected nuanced distinctions within those categories, such as fruits and animals forming individual clusters. Hence, the syntactic categories were indeed just one level of a hierarchy of lexical categories, consistent with constructionist approaches to describing grammatical structure (Goldberg, 2009).

The Redington et al. (1998) approach utilised generic, simple statistical processing mechanisms, consistent with the associative learning approaches that have been tested for effectiveness for speech segmentation (see Section 2.1) and word-meaning (Section 2.2) mappings. However, the extent to which these computations are tractable to a small child acquiring their language remained an open question. Redington et al. (1998) counted co-occurrences between each word in the language and the 150 most frequently occurring words in the same corpus, but in four different positions (two words before, one word before, one word after, two words after). Thus, for 1000 target words, the system is required to store 600,000 separate counts. This is likely to exceed somewhat the child's working memory capacity. Mintz (2003) proposed an alternative method by which co-occurrence information could give rise to syntactic categories based on input alone, that was closer to respecting known limitations about working memory. He identified the 35 most frequent pairs of words separated by one other word, e.g., "*put __ in*") and determined the set of words that occurred within the pair (e.g., *it, them, him, things, teddy, dolly*, …). Then, the extent to

which words occurring within these frames were of the same syntactic category was measured. The results were highly accurate, with more than 90% of words belonging to the same category within each frame. Hence, using a very simple statistical mechanism, based on slightly more specific contextual information than used in Redington et al. (1998), accurate lexical categorisation could proceed.

Though accurate, the frequent frames approach was unable to classify the majority of words in the language, as most words did not occur within these frequent frames. St Clair, Monaghan and Christiansen (2010) applied a neural network to test the optimal statistical information available from child-directed speech in order to generate knowledge about syntactic categories. Specifically, they tested whether sparse, but accurate, frequent frames or whether less specific information, broader in coverage, about just the preceding word and the following word resulted in more effective classification of child-directed speech into syntactic categories. St Clair et al.'s (2010) model was a feedforward network, with distributional information from the corpora of child-directed speech as input. The model was required to produce at the output the syntactic category of each word. The distributional information presented as input represented either the specific frames of Mintz's (2003) approach, or the preceding and the succeeding word. The modelling demonstrated that training on Mintz's specific frames resulted in substantially lower accuracy of syntactic category detection than training with separate preceding and succeeding contextual information. This model indicated that distributional information was sufficient to generate hypotheses about the category of words in child-directed speech, and furthermore that this distributional information was most effective when processed in terms of very local, flexible contextual information about only the preceding and the succeeding word.

Neural network models of lexical categorisation have largely followed the empiricist approach of determining how general purpose statistical mechanisms can apply to the case of language acquisition. However, this work has proceeded somewhat independently of advances in neural network studies of perceptual and attentional development. Exceptions, such as Samuelson et al. (2017) show the benefit of combining these approaches for word learning. For studies of categorisation, there are numerous related neural networks of perceptual category learning, consistent with domain-general accounts of language learning, which demonstrate more closely how computational modelling can relate to neural processing in the brain (see Schultz, 2012, for a review). For instance, young infants are able to categorise stimuli into separate categories, based on distinctions in the visual form of those stimuli (Younger, 1985). Westermann and Mareschal (2012) tested these behaviours using two computational models. The first model simulated perceptual processing of stimuli from different categories using a model with changing receptive field sizes processing the visual input. Visual development has been proposed to adjust from larger to smaller receptive field sizes during infancy (Spencer, Simmering, Schutte, & Schöner, 2007). Implementing this property into the hidden layer of an auto-associator model, which learned to reproduce a visual input at output, was able to simulate visual category processing development in children's early years.

An additional simulation by Westermann and Mareschal (2012) investigated the processing of different memory systems during learning of visual categories. In this model, two auto-associator neural network models mapped the same visual input onto representations of the visual form at two output layers, via two hidden layers. These networks were yoked by interconnecting the networks' hidden layers. One hidden layer was trained with a fast learning rate, representing hippocampal learning, whereas the other was trained with a slow learning rate, representing cortical learning

(McClelland, McNaughton, & O'Reilly, 1995). After training on a set of perceptual category stimuli, the model was presented repeatedly with a test trial, to simulate infant testing in habituation studies. How long the model took to adapt to the stimulus was taken as a measure of the model's processing of the category stimuli. The model was able to simulate developmental shifts in category learning, and, importantly, the role of hippocampal and cortical systems could be appraised separately for the extent to which recent, and long-term memory of stimuli accounted for the effects.

Such closer links between models of perceptual system performance and models that have previously investigated more abstract treatments of linguistic input, are areas where cognitive models of language could be further informed by neural processing. As we shall see in the next sections, neural network models of syntactic processing have made substantial strides in drawing closer together neural processing of the brain's language neural network with computational approaches to studying language.

3.2. Learning syntactic structure

Most, if not all, connectionist models of the acquisition of syntactic structure have been based on models of sentence comprehension or production, which will be discussed in Section 4 below. Here, we focus on the application of RNN sentence processing models to account for aspects of acquisition.

Most often, neural networks that learn sentence processing are trained to generate as output a prediction of the next word at each point of the utterance. Elman (1993) demonstrated that such a next-word prediction SRN can learn a semi-realistic, miniature language with multiple embedded structures. He claimed that the network, somewhat like a child, needs to "start small", that is, it has to be trained on simple sentences first or must start out with limited short-term memory capacity. However, Rohde and Plaut (1999) later showed that the inclusion of semantic (in addition to syntactic) dependencies in the language allows an SRN to learn the language without starting small. These results suggest that a notion of embedded syntactic structure can be learned by systems without hierarchical structure built in, that is, it does not need to be an innate capacity.

Nevertheless, the acquisition of relative (embedded) structures is challenging for children. Diessel and Tomasello (2005) asked English and German four-year-olds to repeat six types of embedded clauses. Even though English and German have very different syntactic structures, both groups of children displayed the same ordering range of production difficulties, with intransitive subject relatives ("That's the woman who played the piano this morning") being the easiest and genitive relatives ("That's the girl whose dog was chasing a cat yesterday") the most difficult to repeat. Fitz and Chang (2008; see also Fitz, Chang, & Christiansen, 2011) simulated this task in a sophisticated RNN model: The Dual-path model of sentence production (discussed in more detail in Section 4.2 below), which assumes separate processing streams for syntactic and semantic information. The model was able to produce embedded structures with more than 90% accuracy, and to approximate most of Diessel and Tomasello's experimental results. Furthermore, Fitz and Chang found that how well and fast a structure is acquired depends on its frequency and its similarities to previously learned structures.

Naturally, children who are acquiring their first language are exposed to more linguistic cues than merely the word-order patterns available to standard SRNs. Reali, Christiansen, and Monaghan (2003) included additional lexical cues during SRN training so that the network also had access to information about word length, phonology, and prosody. Another noticeable thing about this simulation is that the model was not trained on a toy language but on over ten thousand utterances

taken from a corpus of child-directed speech. However, the words were replaced by their lexical categories to reduce the required size of the network. Christiansen, Dale, and Reali (2010) took this realistic approach one step further by representing each of the 1,371 word types in the corpus by a unique vector encoding 16 realistic phonetic cues. RNNs were then trained on sequences of these vectors, corresponding to the child-directed utterances, where the network's task was to predict the upcoming lexical category at each point. As a control condition, other RNNs were trained with a random assignment of the phonetic vectors to words. The networks with the non-random assignment of vectors showed improved generalization, suggesting that infants can indeed make use of phonetic cues to learn grammar. An analysis of the networks' hidden representations revealed that the use of phonetic cues improves the representation of abstract lexical categories (at least for nouns and verbs). Similarly, natural language processing systems, whose primary aim is not to reflect cognitive performance but rather achieve highest possible accuracy, have utilized "character language models", whereby individual letters rather than words are provided as input to the model, and the model must discover the categories of the words to which the letters belong. Such models can outperform models that take as input word-level information (e.g., Kim, Jernite, Sontag, & Rush, 2016).

The Reali and Christiansen studies discussed above demonstrate neural networks' ability to learn about the general word-order patterns in child-directed speech. However, the argument for innate language-specific knowledge often centres on particular constructions that children are claimed to learn without sufficient input and without producing errors. One hallmark example is the formation of polar interrogatives: Turning "the man who is eating is hungry" into a yes/no question requires fronting the correct auxiliary verb (i.e., "is the man who is eating hungry?" and not "is the man who eating is hungry?"). Reali and Christiansen (2005) demonstrated that the Reali et al. (2003) model is in fact able to perform correct auxiliary fronting, which, therefore, may not require innate, language-specific knowledge.

The nativism versus empiricism debate has also been tackled by the Dual-path model of sentence production. Chang et al. (2006) used the model to investigate the syntax learning mechanism early in development. Elicited production tasks (e.g., Tomasello, 2000) tend to support empiricism (where abstractions are learned through experience only, so there is no such thing as innate language competence), whereas some preferential-looking studies (e.g., Naigles, 1990) support nativism. The model simulated both tasks and was evaluated against empirical data (i.e., the patterns found in Tomasello, 2000). Using the same model and input it was shown that preferential-looking preceded production, which led to the conclusion that children learn to abstract syntactic features through experience but that the abstraction mechanism is possible in part because of a pre-existing separation in the brain between "neurons that learn sequences and neurons that encode concepts" (Chang et al., 2006, p. 264), which correspond to the model's syntactic and semantic processing streams, respectively.

Moving beyond the nativism-empiricism debate, connectionist models have provided explanations for several phenomena in first language acquisition. For example, children tend to make overgeneralization errors in the use of locative verbs (e.g. "fill", "pour", "spray") where they seem to understand the verbs' meaning and the sentence structures, but sometimes produce ungrammaticalities like "I filled water into the glass" instead of "I filled the glass with water". Twomey, Chang, and Ambridge (2014) investigated whether the Dual-path model shows the same behavior when it learns syntactic structures. Indeed, the model showed an initial preference for producing first the theme ("water") and then the location ("glass"), which leads to overgeneralization

errors on verbs that have a location-first bias, like "fill". The output of the model was consistent with a corpus analysis on child-directed speech as children seemed to master location-first structures faster than theme-first structures. Over development, the bias towards this structure was reduced and gradually disappeared.

One influential view of language development is that knowledge of syntax helps children learn semantics, a phenomenon known as syntactic bootstrapping (Gleitman, 1990). Conversely, semantic bootstrapping (Pinker, 1984) occurs when knowledge of word meaning supports the acquisition of syntax. In both cases, correlations between syntax and semantics need to be discovered, which is precisely what RNNs do when they learn to map sentence forms to semantic representations. Indeed, such models display syntactic and semantic bootstrapping as an emergent side-effect of learning (Desai, 2002, 2007; Frank & Vigliocco, 2011).

4. Sentence processing

4.1. Sentence comprehension

With only very few exceptions (e.g., Sturt, Costa, Lombardo, & Frasconi, 2003), neural network models of sentence comprehension apply the RNN architecture. If a network is trained on next-word prediction (the standard task in SRNs), it extracts statistical knowledge of the language's word-order patterns (i.e., a form of syntactic knowledge) without explicitly implementing traditional syntactic structures. Early SRN studies (Elman, 1991; Rodriguez, Wiles, & Elman, 1999; Servan-Schreiber, Cleeremans, & McClelland, 1991) aimed to show that these models can nevertheless represent hierarchical syntactic structure implicitly in the organization of their internal state space. It wasn't until several years later that SRNs (and RNNs in general) were used to explain particular phenomena in human sentence processing.

One aspect of sentence comprehension where RNN models have been particularly influential is in simulations of embedded structure processing. Here, a direct advantage of connectionist over traditional, grammar-based accounts is that the former provide an explanation for the difficulty of processing multiple embeddings (as in: "the cook$_i$ who the thief$_j$ who the wife$_k$ loved$_k$ robbed$_j$ served$_i$ food"). In a recurrent network, each next level of embedding needs to be represented in the same network units, resulting in interference that grows stronger as the number of embedded phrases increases. This leads to increased next-word prediction error, which is generally taken to be indicative of processing difficulty in the network (Christiansen & Chater, 1999). Hence, increased processing difficulty for deeper embedding is inherent to RNNs, as it is to humans. In addition, Christiansen and MacDonald (2009) demonstrated that RNNs, like humans, suffer more difficulty with embedded structures (e.g., "the cook$_i$ who the thief$_j$ saw$_j$ serving$_i$ food…") than with cross-dependent recursive structures (which exist in Dutch and Swiss German and correspond to "the cook$_i$ who the thief$_j$ food serving$_i$ saw$_j$…"), while right-branching structures ("the thief$_i$ saw$_i$ the cook$_j$ serving$_j$ food") are the easiest. They also showed that RNNs experience "grammaticality illusion": the phenomenon that removing a required verb from English sentences with double-embedded relative clauses (e.g., "the cook$_i$ who the thief$_j$ who the wife$_k$ loved$_k$ served$_i$ food") makes these (now ungrammatical) sentences appear more acceptable to human readers, who also read them faster. Interestingly, the reversed effect occurs in German and Dutch (i.e., the correct double-embedded structures are read faster than those with a missing verb), a phenomenon that has also been demonstrated in RNNs (Engelmann & Vasishth, 2009; Frank, Trompenaars, & Vasishth, 2016).

English object-relative clauses (e.g., "the cook who the thief saw…") are more difficult to process than subject-relative clauses ("the cook who saw the thief…"), albeit less so for people with

higher exposure to object-relative structures. RNNs behave similarly (MacDonald & Christiansen, 2002). The situation is more complex in Chinese, where the reversed pattern holds when the relative clause modifies the sentence's grammatical subject but not when it modifies the object. Hsiao and MacDonald (2013) replicate this pattern, as well as interactions with noun animacy, in an RNN simulation.

One criticism that can be raised against all simulations referred to above (with the exception of Frank et al., 2016) is that the models are trained and tested on hand-crafted, miniature languages. Hence, they have no knowledge of the true language and are unable to simulate the processing of actual experimental stimuli. Comparisons between model simulations and human behaviour will therefore have to remain qualitative rather than quantitative. However, recent technical developments make it more feasible to train RNNs on large-scale realistic corpora and to evaluate them on the same items used in experiments. The size of next-word prediction errors by these more realistic RNNs explain general patterns of reading times (Frank, 2013; Hahn & Keller, 2016; Monsalve, Frank, & Vigliocco, 2012) and brain activity as measured by EEG (Frank, Otten, Galli, & Vigliocco, 2015) or MEG (Wehbe, Vaswani, Knight, & Mitchell, 2014) from participants reading naturalistic materials. When RNNs are directly compared to symbolic grammars, the RNNs often fit the human data better and the grammars do not account for additional variance (Frank & Bod, 2011; Frank et al., 2015; but see Fossum & Levy, 2012, for conflicting results) suggesting that RNNs form more adequate models of cognitive processing difficulty than traditional grammars.

The ability of RNN prediction error to account for these general reading-time and brain-activity patterns may seem to suggest that many psycholinguistic phenomena can be simulated in this manner. However, this approach has not been very successful in explaining much beyond embedded clause processing. For example, Tabor, Juliano, and Tanenhaus (1997; Tabor & Tanenhaus, 1999) simulated garden-path effects (Frazier & Rayner, 1982) in an RNN but, rather than using prediction error as the relevant measure, they included an additional mechanism that steers the network's internal state towards one of several state-space clusters that correspond to the possible structures of the syntactic ambiguity. The time required to reach a cluster was taken as a predictor of reading time. These simulation results suggest that garden-path effects do not rely (solely) on prediction but also require sentence interpretation, something that is (arguably) not reflected in RNN prediction error. Indeed, the equivalent of prediction error in incremental symbolic probabilistic parsers, which do generate syntactic interpretations, has successfully been used to account for human garden-path effects (Brouwer, Fitz, & Hoeks, 2010; Hale, 2001; Levy, 2008). Again, this suggests that next-word prediction does not suffice to simulate the garden-path phenomena.

However, sentence-processing RNNs are not restricted to next-word prediction. An alternative is to train these networks to transform input sentences into representations of their meaning. As was the case for next-word prediction models, early form-to-meaning RNNs mostly formed proofs of concept that connectionist models can simulate the incremental mapping from sentence input to semantic output, and were much less concerned with explaining human performance (McClelland, St.John, & Taraban, 1989; Miikkulainen, 1996; Miikkulainen & Dyer, 1991; St.John & McClelland, 1990). More recent RNN models of semantic interpretation have been used to explain findings from EEG studies. Hinaut & Dominey (2013) equate the recurrent part of their model with the human brain's frontal cortical network, while the semantic output units are claimed to correspond to the striatum. The amount of activation change in the simulated striatum (i.e., the amount of semantic reinterpretation required to integrate the current word) is taken to simulate the size of the P600 ERP component. In this manner, the model reproduces the finding that object-relative clauses result in

greater P600 than subject-relatives, at the point where the relative clause type is disambiguated. According to the model, this effect is caused not by structural differences between the two sentence types but simply by the fact that object relatives are less frequent and, consequently, less preferred as an initial interpretation.

All models discussed so far receive only linguistic input for training and evaluation. However, as observed in models of word learning and comprehension (e.g., Samuelson et al., 2017; Smith et al., 2017), real-life language use rarely takes place without non-linguistic context. The CIANET model (Mayberry, Crocker, & Knoeferle, 2009) simulates the comprehension of a sentence in visual context. It receives as input not only a sentence but also a simplified representation of a visual scene, and outputs at each point of the sentence both a semantic interpretation and a simulated eye gaze towards one of two relevant parts of the visual scene. This model managed to capture effects of word order, of stereotypicality of the described action, and of conflict between the stereotypical actions of depicted agents and the action described in the sentence. Crocker, Knoeferle, & Mayberry (2010) applied the same model to account for ERP effects. They took the amount of change in recurrent layer activation when processing a word as an index of Left Anterior Negativity (LAN) and P600 size. In this manner, the model could account for effects of word order, disambiguation, and the presence of a visual scene. Interestingly, different recurrent layer units turned out to be responsible for the simulated LAN versus P600 effects, indicating that the network functionally separates the two, as does the brain.

Rabovsky, Hansen, and McClelland (2016) took the amount of activation change in hidden units of the McClelland et al. (1989) model as an index of N400 size, and thereby explained several well-known findings. For one, the simulated N400 was stronger for less expected words and for words earlier in the sentence. Also, semantically incongruent words resulted in stronger N400, but much less so if it they were from the same semantic category as the expected, congruent word. Finally, the model accounted for the so-called *semantic illusion*, where in a sentence like "for breakfast, the eggs eat" the strong association between "breakfast", "eggs" and "eat" results in a relatively weak N400 in spite of the semantic incongruency.

Note that Rabovsky et al. (2016) relate hidden-unit update to N400 size whereas Crocker et al. (2010) take basically the same quantity to reflect the P600. A recent RNN model of semantic illusions by Brouwer, Crocker, Venhuizen and Hoeks (in press) has multiple hidden layers, where one is devoted to lexical retrieval and another to the integration of concept into a semantic representation of the sentence. The N400 is hypothesized to index lexical retrieval difficulty (modeled by the amount of activation change in the lexical retrieval layer) and the P600 corresponds to the update of the sentence interpretation (reflected in the amount of activation change in the integration layer). Indeed, this division of labour accounted for many (if not all) ERP effects observed in the comprehension of semantic illusion sentences.

4.2. Sentence production

There has been substantially less work on models of sentence production than sentence comprehension. It may seem straightforward to construct a production model by running a sentence comprehension model backwards, and this is indeed how two recent connectionist models of production were developed (Calvillo, Brouwer, & Crocker, 2016; Hinaut et al., 2015). However, the most successful and empirically validated sentence production models were specifically designed to simulate production.

The first neural network model of sentence production, the so-called *structural priming* model (Dell, Chang, & Griffin, 1999; see also Chang, Dell, Bock, & Griffin, 2000), was developed to simulate syntactic priming: the tendency of the speakers to repeat the structure of recently spoken or heard sentences (Bock, 1986; Bock & Loebell, 1990). The model assumes a close link between sentence comprehension and production; comprehension of what has been said or heard so far influences the production of a sentence. The model encodes the intended meaning (or "message") by units that represent role-concept pairs (e.g. "agent-CHILD" or "patient-MAN") which form input to the hidden layer during production of the whole sentence. The output layer units represent words, where the most active unit is taken to be the produced word and is fed back into the network, which thereby receives information about what has been produced so far. This model was able to successfully account for several structural priming phenomena, for instance, if "Boys chase dogs" was used as prime (i.e., active rather than passive voice), the model would produce the message "agent-GIRL; action-FEED; patient-CAT" as "girl feeds cat" instead of "cat is fed by girl" (Dell et al., 1999). However, it failed to show priming between transitive locatives ("Boys chase dogs near car") and prepositional datives ("Boys give dog to girl") which is empirically shown by Bock and Loebell (1990). Another limitation of the model was that, because each concept-role pair is represented in a single unit, the agent-concept MAN is different from the patient-concept MAN. Consequently, the model is unable to generalize its ability to produce "the man is chasing a dog" to the ability to produce "the dog is chasing a man". This violates the property of systematicity, which Fodor and Pylyshyn (1988) argued is a fundamental feature of human cognition that neural networks do not possess.

Chang (2002) proposed and compared two neural network models of sentence production, Prod-SRN and Dual-path. Prod-SRN is a simple extension of the structural priming model, tested on a more advanced morphology and closer to a typical SRN, but still lacking systematicity. Dual-path, which is still the most influential neural network model of sentence production, was the first to overcome the limitation of generalization. It does so by creating temporary bindings between a layer for roles and a layer for concepts, so that there is only one unit for MAN, irrespective of its semantic role. These bindings, along with the event semantics (information about tense and aspect, e.g., PRESENT SIMPLE), form the model's *semantic stream*. The model has a second stream (hence its name), the *syntactic stream*, which is an SRN that allows the model to learn syntactic categories. This way, the model was not only able to generalize words to new positions, but also to generalize a noun as a verb; this is something that speakers usually do with proper nouns, e.g. "Skype" becomes "skyping". Chang (2002) compared Prod-SRN to the Dual-path model, and the latter was able to generalize 82% of the time whereas Prod-SRN only reached 6%. The models were also tested on unseen adjective-noun pairs and identity construction (e.g., "a cat is a cat"); Dual-path outperformed Prod-SRN in all tests. The model also expanded on Gordon and Dell's (2003) simple model of aphasic production, offering a natural explanation of two different types of aphasia, agrammatism and anomia (see also Dell & Chang, 2014)

Chang, Dell, and Bock (2006) applied the Dual-path model to a wider range of structural priming phenomena. The model displayed similar priming whether the prime had been produced or only comprehended. It was also able to account for long-term priming, as the extent of structural priming was not dependent on the number of fillers between the sentences. Furthermore, Chang et al. (2006) showed that the strong but short-lived tendency to repeat previously heard or said words (the so-called lexical boost) is due to a different mechanism from structural priming. This prediction was confirmed experimentally two years later (Hartsuiker, Bernolet, Schoonbaert, Speybroeck, & Vanderelst, 2008).

To test whether the model could handle a language that is typologically very different from English, Chang (2009) tested Dual-path on Japanese. Despite the different word orders between these two languages, the model was able to exhibit similar levels of grammaticality (93% for English and 95% for Japanese). Furthermore, the model was able to explain differences in production preferences between speakers of these two languages. For instance, in English long phrases are usually placed after short ones (e.g., "The woman sent a book to the man that she met while traveling" is preferred over "The woman sent the man that she met while traveling a book"). This phenomenon is called heavy NP shift (Ross, 1967), and English exhibits a short-before-long bias whereas Japanese has the opposite. The model was able to account for this cross-linguistic difference. Chang (2009) showed that the phenomenon is caused by a difference in the relative importance of the meaning in the positions ("choice points") where the word orders differed. For English, the choice point was right after the verb, whereas in Japanese the choice point is at the beginning of the sentence as verbs tend to occur at the end of the sentence.

The Dual-path model was further able to account for cross-linguistic differences in lexical/conceptual accessibility between English and Japanese (Chang, 2009). English speakers tend to prefer using animate elements early in the sentence (McDonald, Bock, & Kelly, 1993), which can lead to the usage of less common structures like passives (e.g., "the man was almost hit by a car"). At the same time, they do not have animacy preference in conjunctions: speakers of English find "the man and the car" as acceptable as "the car and the man". Therefore, it was hypothesized that animacy can influence the functional level but not the positional level from Garrett's (1988) theory of sentence production. However, if this were the case, animacy wouldn't affect word order in Japanese as this language uses case-markers to indicate roles and repositioning words doesn't affect the meaning (e.g., in the passive sentence above, "man" would be marked as the subject that receives the action "hit", regardless of word order). Nevertheless, it has been shown that animacy affects word position (Branigan, Pickering, & Tanaka, 2008). Using Dual-path, Chang (2009) noticed that these preferences were related to the frequency of the input; by giving it sentences where animate words were used early in the sentence, the model learned stronger connections between animate concepts and words than for inanimate ones.

One criticism against models of sentence production is that all current models use a miniature (artificial) language instead of natural stimuli. Another limitation of the current models of sentence production is that none of them consider the phonological level of sentence production (Garrett, 1988). Rather, they all focus on message planning (conceptualization) and sentence formation. Furthermore, even though in most models the hypothesis that comprehension influences production is supported, the exact connection between these has yet to be established.

5. Conclusion

Connectionist models have been instrumental in explaining a range of human language behaviours, from word segmentation and word-meaning mapping to sentence processing and syntactic development. These models' successes have demonstrated the richness of the environment for language learning, contributing constructively to debates over empiricist versus nativist positions of language acquisition. They have also been useful in determining the knowledge required by an information processing system for simulating human behaviour, addressing questions of the extent to which language processing is hierarchical or sequential, the interface between syntax and semantics, and the role of prediction and richer interpretation in sentence processing.

In the machine learning and natural language processing (NLP) literature, “deep learning” neural networks currently outperform other systems on a range of NLP tasks. For example, very large corpora of natural texts can now be used to train RNNs on next-word prediction (Mikolov, Deoras, Povey, Burget, & Černocký, 2011) and to let feedforward networks extract lexical semantics from distributional patterns (Mikolov et al., 2013). RNNs have been applied to select which image is described by a sentence (Chrupała, Kádár, & Alishahi, 2015) and, conversely, for generating sentences that describe a given image (Chen & Zitnick, 2015), suggesting that the problem of representing semantics in production models (see Section 4.2) can be solved by using images as meaning representations instead of pairs of semantic roles and fillers.

With a few exceptions (e.g., Hahn & Keller, 2016; Kamper, Jansen, & Goldwater, 2016), such NLP models are not generally intended to simulate human cognitive processing and, consequently, their evaluation against human data is still in its infancy. As a case in point, the RNN model by Mikolov et al. (2011) shows excellent next-word prediction performance on natural language, but unlike the models by Elman (1993) and Rohde and Plaut (1999) discussed in Section 3.2., its behaviour has not been related to human performance or acquisition, for example, regarding embedded clause processing. Nevertheless, the remarkable performance of current neural networks may well suggest that they embody relevant aspects of the human language system, affording great potential for further applications in the psycholinguistic community.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Of Memory and Language*, *38*, 419–439.

Bock, J. K. (1986). Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*, 575–586.

Bock, K., & Loebell, H. (1990). Framing sentences. *Cognition*, *35*, 1–39.

Branigan, H. P., Pickering, M. J., & Tanaka, M. (2008). Contributions of animacy to grammatical function assignment and word order during production. *Lingua*, *118*, 172–189.

Brouwer, H., Crocker, M. W., Venhuizen, N. J., & Hoeks, J. C. J. (in press). A neurocomputational model of the N400 and the P600 in semantic processing. *Cognitive Science*.

Brouwer, H., Fitz, H., & Hoeks, J. (2010). Modeling the noun phrase versus sentence coordination ambiguity in Dutch: evidence from surprisal theory. In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics* (pp. 72–80). Uppsala, Sweden: Association for Computational Linguistics.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, *33*, 111–153.

Calvillo, J., Brouwer, H., & Crocker, M. W. (2016). Connectionist semantic systematicity in language production. In: Proceedings of CogSci 2016. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

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

Chang, F. (2009). Learning to order words: a connectionist model of heavy NP shift and accessibility effects in Japanese and English. *Journal of Memory and Language*, *61*, 374–397.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*, 234–272.

Chang, F., Dell, G. S., Bock, K., & Griffin, Z. M. (2000). Structural priming as implicit learning: A comparison of models of sentence production. *Journal of Psycholinguistic Research*, *29*, 217–230.

Chen, X., & Zitnick, C. L. (2015). Mind’s eye: A recurrent visual representation for image caption generation. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2422–2431).

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: The MIT Press.

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.

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*, 157–205.

Christiansen, M. H., Dale, R., & Reali, F. (2010). Connectionist explorations of multiple-cue integration in syntax acquisition. In S. P. Johnson (Ed.), *Neoconstructivism: The new science of cognitive development* (pp. 87–108). New York: Oxford University Press.

Christiansen, M. H., & MacDonald, M. C. (2009). A usage-based approach to recursion in sentence processing. *Language Learning*, *59*, 126–161.

Chrupała, G., Kádár, À., & Alishahi, A. (2015). Learning language through pictures. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics* (pp. 112–118). Beijing, China: Association for Computational Linguistics.

Clark, A., & Karmiloff-Smith, A. (1993). The cognizer's innards: a psychological and philosophical perspective on the development of thought. *Mind & Language*, *8*, 487–519.

Colunga, E., & Smith, L. B. (2005). From the lexicon to expectations about kinds: a role for associative learning. *Psychological Review*, *112*, 347–382.

Crocker, M. W., Knoeferle, P., & Mayberry, M. R. (2010). Situated sentence processing: the coordinated interplay account and a neurobehavioral model. *Brain and Language*, *112*, 189–201.

Cunillera, T., Toro, J. M., Sebastián-Gallés, N., & Rodríguez-Fornells, A. (2006). The effects of stress and statistical cues on continuous speech segmentation: An event-related brain potential study. *Brain Research*, *1123*, 168–178.

Dell, G. S., & Chang, F. (2014). Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B*, *369*, 20120394.

Dell, G. S., Chang, F., & Griffin, Z. M. (1999). Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science*, *23*, 517–542.

Desai, R. (2002). Bootstrapping in miniature language acquisition. *Cognitive Systems Research*, *3*, 15–23.

Desai, R. (2007). A model of frame and verb compliance in language acquisition. *Neurocomputing*, *70*, 2273–2287.

Diessel, H., & Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language*, *81*, 882–906.

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

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–225.

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, *48*, 71–99.

Engelmann, F., & Vasishth, S. (2009). Processing grammatical and ungrammatical center embeddings in English and German: A computational model. In A. Howes, D. Peebles, & R. Cooper (Eds.), *Proceedings of 9th International Conference on Cognitive Modeling* (pp. 240–245). Manchester, UK.

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*, 1017–1063.

Fitz, H., & Chang, F. (2008). The role of the input in a connectionist model of the accessibility hierarchy in development. In *32nd Annual Boston University Conference on Language Development* (pp. 120–131). Cascadilla Press.

Fitz, H., Chang, F., & Christiansen, M. H. (2011). A connectionist account of the acquisition and processing of relative clauses. In E. Kidd (Ed.), *The acquisition of relative clauses. Processing, typology and function* (pp. 39–60). Amsterdam: John Benjamins.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition*, *28*, 3–71.

Fosler-Lussier, E., Amdal, I., & Kuo, H. K. J. (2005). A framework for predicting speech recognition errors. *Speech Communication*, *46*, 153–170.

Fossum, V., & Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics* (pp. 61–69). Montréal, Canada: Association for Computational Linguistics.

Frank, S. L. (2013). Uncertainty reduction as a measure of cognitive effort in sentence comprehension. *Topics in Cognitive Science*, *5*, 475–494.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*, 829–834.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, *140*, 1–11.

Frank, S. L., Trompenaars, T., & Vasishth, S. (2016). Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, *40*, 554–578.

Frank, S. L., & Vigliocco, G. (2011). Sentence comprehension as mental simulation: an information-theoretic perspective. *Information*, *2*, 672–696.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: eye-movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*, 178–210.

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A Recognition-Based Connectionist Framework for Sequence Segmentation and Chunk Extraction. *Psychological Review*, *118*(4), 614–636.

Gambell, T., & Yang, C. D. (2003). Scope and limits of statistical learning in word segmentation. In *Proceedings of the 34th Northeastern Linguistic Society Meeting* (pp. 29–30). New York: Stony Brooks University.

Garrett, M. F. (1988). Processes in language production. In F. J. Newmeyer (Ed.), *Linguistics: The Cambridge survey. Language: Psychological and biological aspects* (Vol. 3, pp. 69–96). Cambridge, UK: Cambridge University Press.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*, 3–55.

Goldberg, A. E. (2009). The nature of generalization in language. *Cognitive Linguistics*, *20*, 93–127.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *16*, 447–474.

Gordon, J. K., & Dell, G. S. (2003). Learning to divide the labor: an account of deficits in light and heavy verb production. *Cognitive Science*, *27*, 1–40.

Hahn, M., & Keller, F. (2016). Modeling Human Reading with Neural Attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 85–95). Austin, TX: Association for Computational Linguistics.

Hale, J. T. (2001). A probabilistic Early parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics* (Vol. 2, pp. 159–166). Pittsburgh, PA: Association for Computational Linguistics.

Hartsuiker, R. J., Bernolet, S., Schoonbaert, S., Speybroeck, S., & Vanderelst, D. (2008). Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. *Journal of Memory and Language*, *58*, 214–238.

Hebb, D. O. (1949). *The Organization of Behavior*. New York: Wiley & Sons.

Hinaut, X., & Dominey, P. F. (2013). Real-time parallel processing of grammatical structure in the fronto-striatal system: a recurrent network simulation study using reservoir computing. *PLoS ONE*, *8*, e52946.

Hinaut, X., Lance, F., Droin, C., Petit, M., Pointeau, G., & Dominey, P. F. (2015). Corticostriatal response selection in sentence production: Insights from neural network simulation with reservoir computing. *Brain and Language*, *150*, 54–68.

Hsiao, Y., & MacDonald, M. C. (2013). Experience and generalization in a connectionist model of Mandarin Chinese relative clause processing. *Frontiers in Psychology*, *4*, 767.

Kamper, H., Jansen, A., & Goldwater, S. (2016). Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. In *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* (pp. 669–679).

Kamper, H., Wang, W., & Livescu, K. (2016). Deep convolutional acoustic word embeddings using word-pair side information. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4950–4954).

Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (2016). Character-aware neural language models. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence* (pp. 2741–2749). AAAI Press.

Kiss, G. R. (1973). Grammatical word classes: A learning process and its simulation. *Psychology of Learning and Motivation*, *7*, 1–41.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: a comment on Just & Carpenter (1992) and Waters & Caplan (1996). *Psychological Review*, *109*, 35–54.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman and Company.

Mayberry, M. R., Crocker, M. W., & Knoeferle, P. (2009). Learning to attend: a connectionist model of situated language comprehension. *Cognitive Science*, *33*, 449–496.

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

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457.

McClelland, J. L., St.John, M. F., & Taraban, R. (1989). Sentence comprehension: a parallel distributed processing approach. *Language and Cognitive Processes*, *4*, 287–335.

McDonald, J. L., Bock, K., & Kelly, M. H. (1993). Word and world order: Semantic, phonological, and metrical determinants of serial position. *Cognitive Psychology*, *25*, 188–230.

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*, 831–877.

Miikkulainen, R. (1996). Subsymbolic case-role analysis of sentences with embedded clauses. *Cognitive Science*, *20*, 47–73.

Miikkulainen, R., & Dyer, M. G. (1991). Natural language processing with modular PDP networks and distributed lexicon. *Cognitive Science*, *15*, 343–399.

Mikolov, T., Chen, K., Corrado, C., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the ICLR Workshop*.

Mikolov, T., Deoras, A., Povey, D., Burget, L., & Černocký, J. (2011). Strategies for training large scale neural network language models. In *IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 196–201).

Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91–117.

Monaghan, P., & Christiansen, M. H. (2010). Words in puddles of sound: modelling psycholinguistic effects in speech segmentation. *Journal of Child Language*, *37*, 545–564.

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In W. Daelemans (Ed.), *Proceedings of the 13th conference of the European chapter of the Association for Computational Linguistics* (pp. 398–408). Avignon, France: Association for Computational Linguistics.

Mozer, M. C. (1991). *The perception of multiple objects: A connectionist approach*. Cambridge, MA: The MIT Press.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, *17*, 357–374.

Philips, L., & Pearl, L. (2015). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science*, *39*, 1824–1854.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA.

Plunkett, K. (1997). Theories of early language acquisition. *Trends in Cognitive Sciences*, *1*, 146–153.

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

Rabovsky, M., Hansen, S., & McClelland, J. L. (2016). N400 amplitudes reflect change in a probabilistic representation of meaning: evidence from a connectionist model. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structural dependence and indirect statistical evidence. *Cognitive Science*, *29*, 1007–1028.

Reali, F., Christiansen, M. H., & Monaghan, P. (2003). Phonological and distributional cues in syntax acquisition: scaling up the connectionist approach to multiple-cue Integration. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 970–975). Mahwah, NJ: Cognitive Science Society.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic structures. *Cognitive Science*, *22*, 425–469.

Regier, T. (2003). Emergent constraints on word-learning: A computational review. *Trends in Cognitive Sciences*, *7*, 263–268.

Rodriguez, P., Wiles, J., & Elman, J. L. (1999). A recurrent neural network that learns to count. *Connection Science*, *11*, 5–40.

Rohde, D. L. T., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition*, *72*, 67–109.

Ross, J. R. (1967). *Constraints on variables in syntax*. Massachusetts Institute of Technology.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & The PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the microstructure of cognition*. The MIT Press.

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

Samuelson, L. K., Kucker, S. C., & Spencer, J. P. (2017). Moving word learning to a novel space: a dynamic systems view of referent selection and retention. *Cognitive Science*, *41*, 52–72.

Scharenborg, O. (2010). Modeling the use of durational information in human spoken-word recognition. *The Journal of the Acoustical Society of America*, *127*, 3758–3770.

Schultz, T. R. (2012). A constructive neural-network approach to modeling psychological development. *Cognitive Development*, *27*, 383–400.

Schütze, H. (1993). Part-of-speech induction from scratch. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Servan-Schreiber, D., Cleeremans, A., & McClelland, J. L. (1991). Graded state machines: the representation of temporal contingencies in simple recurrent networks. *Machine Learning*, *7*, 161–193.

Smith, A. C., Monaghan, P., & Huettig, F. (2017). The multimodal nature of spoken word processing in the visual world: Testing the predictions of a multimodal parallel integration model. *Journal of Memory and Language*, *93*, 276–303.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 1558–1568.

Spencer, J. P., Simmering, V. R., Schutte, A. R., & Schöner, G. (2007). What does theoretical neuroscience have to offer the study of behavioral development? Insights from a dynamic field theory of spatial cognition. In J. M. Plumert & J. P. Spencer (Eds.), *The emerging spatial mind* (pp. 320–361). New York: Oxford University Press.

St Clair, M. C., Monaghan, P., & Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*, *116*, 341–360.

St.John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, *46*, 217–257.

Sturt, P., Costa, F., Lombardo, V., & Frasconi, P. (2003). Learning first-pass structural attachment preferences with dynamic grammars and recursive neural networks. *Cognition*, *88*, 133–169.

Tabor, W., Juliano, C., & Tanenhaus, M. K. (1997). Parsing in a dynamical system: an attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes*, *12*, 211–271.

Tabor, W., & Tanenhaus, M. K. (1999). Dynamical models of sentence processing. *Cognitive Science*, *23*, 491–515.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, *74*, 209–253.

Twomey, K. E., Chang, F., & Ambridge, B. (2014). Do as I say, not as I do: A lexical distributional account of English locative verb class acquisition. *Cognitive Psychology*, *73*, 41–71.

Twomey, K. E., Ranson, S. L., & Horst, J. S. (2014). That's more like it: multiple exemplars facilitate word learning. *Infant and Child Development*, *23*, 105–122.

Wehbe, L., Vaswani, A., Knight, K., & Mitchell, T. (2014). Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 233–243). Doha, Qatar: Association for Computational Linguistics.

Westermann, G., & Mareschal, D. (2012). Mechanisms of developmental change in infant categorization. *Cognitive Development*, *27*, 367–382.

Younger, B. A. (1985). The segregation of items into categories by ten-month-old infants. *Child Development*, *56*, 1574–1583.

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. B. (2012). Modeling cross-situational word–referent learning: Prior questions. *Psychological Review*, *119*, 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.