Neurocomputational models of natural language

Alistair Knott

October 27, 2011

Abstract

In this chapter I review computational models of the neural circuitry which implements the human capacity for language. These models cover many different aspects of the language capacity, from representations of phonology and word forms to representations of sentence meanings and syntactic structures. The computational models discussed are neural networks: structures of simple units which are densely interconnected and can be active in parallel. I review the computational properties of the networks which are introduced, and also empirical evidence from different sources (neural imaging, behavioural experiments, patterns of impairment following brain dysfunction) which supports the models described.

Contents

1 Introduction 2

2 Phonological representations and processing 3
   2.1 Phonemes and syllables .............................................. 3
   2.2 Phonological working memory ...................................... 6

3 Lexical representations and processes 9
   3.1 Phonological representations of whole words .................. 9
       3.1.1 Phonological input word representations ................... 10
       3.1.2 Phonological output word representations .................. 11
   3.2 Neural representations of word meanings ....................... 12
   3.3 Mappings between word forms and word meanings ................ 14
   3.4 Morphology ............................................................ 16

4 Sentence-sized semantic representations 18
   4.1 Formal models of sentence and discourse semantics ............ 18
   4.2 Episodic long-term memory representations ..................... 19
   4.3 Working memory episode and context representations .......... 20
   4.4 Binding mechanisms in episode representations ................. 22
1 Introduction

Since the earliest days of neural network research, theorists have attempted to model the neural circuitry which supports the human language faculty. Rosenblatt’s (1958) perceptron was designed to recognise individual letters; Widrow and Hoff’s (1960) Adaline system was deployed in speech recognition among other tasks. Many of the pioneering PDP models of the 1980s had linguistic applications—for instance, McClelland and Rumelhart’s (1981) interactive activation model of word reading, and Rumelhart and McClelland’s (1986) model of past tense learning. These early models paved the way for a veritable industry of research into the neural basis of language processing, combining empirical experimentation with computational simulations. In this chapter I will sketch some of the main results of this research programme, summarising what has been learned experimentally about language processing in the brain, and describing some contemporary neural network models informed by these findings.

As is conventional in reviews of neural language circuitry, I will proceed ‘from the outside inwards’, working from the more peripheral sensory and motor representations of language in the brain to the more abstract, processed representations which are harder to investigate. In Section 2 I will consider the circuitry which supports auditory and articulatory processing. (I will restrict my attention to spoken language in the current chapter.) In Section 3 I will consider the circuitry which supports representation and processing of individual words; in Section 4 I will discuss some ideas about how the meanings of whole sentences are stored, and in Section 5 I will consider the circuitry involved in representing and processing the syntactic structure of sentences.

The neural network models I introduce in this chapter are models which learn to perform their allocated task through exposure to training data in one format or another. As a consequence, they are not just models of how the brain represents and processes language—they are also models of how children acquire language. In various places, therefore, I will also draw on empirical studies of child language acquisition to motivate the models I present.
2 Phonological representations and processing

2.1 Phonemes and syllables

Articulatory movements are continuous in nature, so speech sounds vary along a continuum. However, there is good evidence that the neural circuitry which produces and interprets these sounds represents them discretely. The basic data bearing this out come from the linguistic discipline of phonology. Any given language has a fixed repertoire of discrete speech sounds called phonemes; speakers of a given language parse a speech stream into a sequence of phonemes. Different phonemes are available in different languages: for instance, the speech sounds which in English correspond to the phonemes ‘l’ and ‘r’ map onto a single phoneme in Japanese. Moreover, a given phoneme is produced in different ways in different contexts, due to a phenomenon called coarticulation: for instance, the phoneme /t/ manifests itself differently in the phrases *hot porridge* and *light*. During phonological development, infants become attuned to the phonological distinctions made in their mother tongue (see e.g. Werker and Tees, 1999). What are the neural circuits which learn a repertoire of phonemes?

We know that phonemes must be represented both in the articulatory system and in the auditory system—and of course, the same sets of phonemes must be learned in each system. The neural areas we expect to be involved in articulation are in the inferior areas of motor cortex controlling the vocal apparatus, and also in associated inferior areas of premotor cortex, where complex articulatory movements are prepared (see e.g. Boldland and Guenther, 2006). One influential idea is that phonemes correspond to planned articulatory gestures—i.e. planned movements of the lips and tongue to particular goal locations (see Browman and Goldstein, 1995). A discrete repertoire of phonemes can then be understood as a discrete set of articulatory plans. When an agent is speaking, there is a premium on speed, causing speakers to overlap their articulatory gestures: for instance, when we produce the word *hot*, we do not wait until we finish producing the ‘h’ sound before shaping our mouth for the ‘o’ sound. Often gestures succeed each other so rapidly that some gestures are left incomplete: for instance, in the gesture producing the ‘t’ phoneme in the phrase *hot porridge*, the tongue does not reach the same point it reaches in the word *light*. Nonetheless, the articulatory intention is the same in both cases. The hypothesis that phonemes correspond to planned or intended articulatory gestures is still subject to debate, but it can account for a good deal of the complexity of the mapping between phonemes and acoustic signals.

Of course, speakers can only overlap articulatory gestures to the extent that hearers can recover the intended sequence of gestures. A speaker must be able to transmit a phoneme sequence to a hearer in order for communication to take place. If phonemes are intended articulatory gestures, the hearer must be able to recover the speaker’s articulatory intentions from the speech signal. Coarticulation means that the mapping cannot be expressed at the level of single phonemes: rather, hearers must learn to recognise the acoustic correlates of short sequences of phonemes. We must therefore envisage a neural circuit which learns to map acoustic signals onto planned sequences of articulatory gestures.
This circuit obviously begins in the auditory cortex; there is evidence that it continues in the superior temporal sulcus (STS), by which time input speech signals are already represented in a more discrete phonological code (see e.g. Hickok and Poeppel, 2007).

How does the circuitry in between auditory cortex and STS learn to map from acoustic signals to phonemes? We know that this process happens during infancy (see again Werker and Tees, 1999). The standard proposal is that infants teach themselves the mapping by babbling. According to this proposal, infants produce articulatory gestures at random, generating reafferent acoustic signals which they perceive through audition, and learn to map these auditory signals back onto the motor plans which produced them. This idea requires that efferent copies of articulatory plans are available to the circuitry in STS—i.e. that there are strong connections between the articulatory system and the auditory system. Such connections have indeed been found. There are direct connections between STS and the inferior motor/premotor cortices, via a tract of white matter called the arcuate fasciculus, but also strong indirect connections between these two areas via a way-station in the inferior parietal cortex (Catani et al., 2004). These tracts are better developed in humans than in chimpanzees or macaques (Rilling et al., 2008), especially in the left hemisphere, which is specialised for language. In mature human speakers, there is good experimental evidence that auditory speech signals evoke articulatory representations in premotor cortex. For instance Wilson et al. (2004) found using fMRI that hearing a nonsense syllable activates the same premotor area activated when that syllable is pronounced. Fadiga et al. (2002) found that in a condition when hearers received transcranial magnetic stimulation (TMS) over their motor cortex to amplify prepared motor commands, hearing words whose pronunciation required large tongue movements elicited more activity in tongue muscles than hearing words whose pronunciation required smaller tongue movements. These findings suggest that hearers do map speech sounds onto their associated articulatory movements. Additionally, Ito et al. (2009) found that stretching the skin around a hearer’s mouth to mimic the shape it assumes when pronouncing a particular vowel sound modulated the hearer’s phonological interpretation of an acoustic stimulus presented in synchrony with the motor deformation. This finding is evidence that there is also a route from the articulatory system back to the auditory system. Corroborating evidence for this comes from an fMRI study by Hickok et al. (2000), finding that speakers activate auditory representations in STS even when they cannot hear the sound of their own voice. In summary, we know something about the neural circuitry which allows infants to learn the discrete phonological representations which capture invariances between spoken and heard speech sounds. Interestingly, the circuitry implicated in learning to recognise the articulatory actions of others overlaps very significantly with the circuitry thought to be involved in the perception of other types of motor action, which also runs from primary sensory cortices through the STS, and then through inferior parietal cortex to the premotor cortex (see e.g. Battaglia-Mayer et al., 2003). The prevalent models of this more general action-recognition circuitry is trained also assume a role for something like babbling (Iacoboni et al., 2001; Oztop and Arbib, 2002; Keysers and Perrett, 2004).

Returning to the domain of speech, there are several connectionist models of the circuitry which learns associations between acoustic stimuli and articulatory plans—see e.g.
Westermann and Miranda (2004), Guenther et al. (2006), Garagnani et al. (2008). In these models, there are separate neural media representing articulatory plans and acoustic stimuli, which are linked by synaptic connections. During a simulated babbling phase, associations are learned between articulatory and acoustic representations using some variant of the Hebbian rule which associates units which are temporally correlated. These connections serve slightly different purposes at different points in development: to begin with, connections from articulatory to acoustic representations provide a training signal to allow a mapping to be learned, while in the mature system, these connections also serve to create auditory targets for articulatory gestures, which can be compared against the actual acoustic signals produced by articulations, to provide an error signal which can be corrected if necessary.

As already noted, the mapping between phonemes and acoustic signals must be expressed at the level of short *sequences* of phonemes rather than (or as well as) at the level of individual phonemes. (Indeed, the connectionist models just described operate at the level of phoneme sequences.) We therefore expect a distinct level of phonological representation in the brain, which groups phonemes into short sequences. Again, the linguistic discipline of phonology tells us that there is such a representation: phonemes are organised into *syllables*. The syllable can be defined in several ways, which roughly converge. On one definition it is a rhythmic unit of speech: a stream of speech consists of a stream of syllables, delivered in a roughly periodic manner. On another definition it is an acoustic unit, organised around a high-sonority sound (a vowel). A syllable consists of an *onset* followed by a *rhyme*, each of which is a short phoneme sequence. The onset is a (possibly empty) sequence of consonants; the rhyme is a vowel, followed by another (possibly empty) sequence of consonants. There is some evidence that syllables are the level at which articulatory and acoustic representations communicate with one another; see e.g. Wheeldon and Levelt (1995).

Figure 1 summarises the proposals about phonological representations introduced in this section. The labels ‘input syllable’ and ‘output syllable’ are somewhat misleading: as we have seen, superior temporal areas are active during output, and premotor areas are active during input. But the diagram at least roughs out the basic pathways involved in mapping between input and output representations.

Figure 1: Phonological input and output representations
2.2 Phonological working memory

When hearers apprehend a speech signal, the phonological representations they evoke can be maintained for a short time in working memory. There is good evidence that phonological representations are stored in a special-purpose working memory medium, which only interfaces with other cognitive mechanisms in certain well-delineated ways. The medium is called the phonological buffer, or the phonological loop; see classically Baddeley (1974). The properties of this memory store are typically investigated in experiments where subjects are presented with a phonological stimulus taking the form of a sequence of syllables, which they must reproduce after a short interval. There are several key findings. Firstly, sequences containing syllables which phonologically resemble one another are harder to retain. This means that items in phonological working memory are likely to be active in parallel while they are being retained, in a way which causes similar items to interfere with one another. Secondly, sequences whose syllables form meaningful words are easier to retain. So there must be an interface between representations active in phonological working memory and semantic representations. Thirdly, there is evidence that subjects retain phonological stimuli in working memory by rehearsing these stimuli, either vocally or subvocally. If subjects have to perform a distractor task engaging the articulatory system (for instance counting to ten) while retaining a phonological stimulus, their ability to recall it is considerably impaired.

The dominant computational models of phonological working memory envisage a medium in which multiple phonological representations are active in parallel and compete for selection, with competition being biased by a separate mechanism representing a sequence of temporal contexts. For instance, in the model of Burgess and Hitch (1999), input and output phonological representations are bidirectionally connected to one another through a ‘direct pathway’ analogous to the arcuate fasciculus, but are also indirectly connected via a set of ‘item’ units, which can be temporarily associated with both input and output phonemes, and also with a temporal context signal which can be played forward deterministically to produce a reference sequence of context representations. A phonological sequence is encoded in working memory by playing the context signal forward while the sequence is presented. At each time point, the current phoneme and the current context representation activate the item units. Item units compete so that only one remains active; this stores a temporary association between a temporal context and a phoneme. Winning item units also self-inhibit after a short time, ensuring that subsequent phonemes are stored by different item units. To recall a phonological sequence, the context signal is replayed, reactivating the item units, and biasing competition towards the production of a particular phoneme at each point in time. The direct connections from output phonemes back to input phonemes allow for a sequence of phonemes to be rehearsed: when a phoneme is produced (either vocally or subvocally), it reactivates its corresponding input phoneme, which can be reassociated with the current item and temporal context.

Burgess and Hitch’s model uses single phonemes as its input and output units, but other models feature more structured phonological representations. For instance, in Hartley and Houghton’s (1996) model, the phonological units which are sequenced by being associated
with items represent whole syllables. Each syllable is stored as an onset and a rhyme unit, each of which in turn can represent multiple phonemes. Onset and rhyme units are directly associated with phonemes, but they are also associated with a template encoding general constraints on how onsets and rhymes can be formed, which help impose an order on the activated phonemes. Many of the errors found in phonological short-term recall reflect the syllabic structure of the stimulus: for instance, ordering errors tend to be of either onsets or rhymes, rather than of arbitrary phonemes (thus *dap, kon* is more likely to be misrecalled as *kap, don* than as *dan, kop*). Hartley and Houghton’s model accounts well for these syllable-sized effects. It is also consistent with the suggestion argued for in Section 2.1, that the direct pathway from input to output phonological representations operates at the level of syllables.

Phonological working memory is thought to play a role in language development. It has been found that the capacity of infants’ phonological working memory predicts their vocabulary size (Gathercole and Baddeley, 1990). Computational models of this effect typically assume that phonological working memory increases the temporal window within which associations between phonological word forms and their meanings can be made. An infant rarely hears the word *dog* at exactly the same moment as she is looking at a dog. Models of phonological working memory, with their assumption that multiple phonological units can be active in parallel, allow a current meaning representation to be associated with several recent phonological forms, not just with the most recent one. Phonological working memory also correlates with syntactic development (Adams and Gathercole, 1995); I will discuss models of this effect in Section 5.

While there is a direct mapping from input to output phonological representations at the level of syllables, there also appear to be working memory media in which input and output phonological representations are stored separately. Producing phonemes is ultimately a matter of articulation, i.e. of motor control. The motor system is hierarchically organised; like any high-level movements, complex articulatory movements need to be planned in advance, and there is likely to be a motor medium in which planned complex articulations are represented. Indeed there is evidence for such a medium: for instance Shallice et al. (2000) describe a patient whose working memory for digits is relatively normal, but who nonetheless has difficulty sequencing phonemes when speaking. The patient makes the same kinds of error whether repeating phonologically complex spoken words, reading complex words, or naming objects: since these tasks have different inputs, the problem appears to be in the output modality. At the same time, the patient does not have low-level articulatory difficulties: the problem is with planning complex phonological structures. Shallice et al. call the planning medium which appears to be damaged the phonological output buffer. To account for this pattern of dysfunction, we should probably assume that the phonemes (or syllables) which are produced as output by Burgess and Hitch’s model are not directly pronounced, but provide input to a separate, specifically articulatory, phonological output buffer, which is also accessed by other tasks such as speaking or picture naming. Other patients display a converse pattern of dysfunction, showing serious impairments in phonological short-term memory, but intact abilities to generate speech and name objects (see e.g. Shallice and Butterworth, 1977). These patients are diagnosed with
damage to a **phonological input buffer**, which stores incoming phonological sequences and supports their repetition syllable by syllable. Burgess and Hitch’s original network, or the modification of it proposed by Hartley and Houghton, are good models of the damaged circuitry in these patients.

A modified version of Burgess and Hitch’s (1999) model of phonological working memory, updated to encode whole syllables rather than single phonemes, and extended with a separate phonological output buffer, is shown in Figure 2. Note that the ‘direct’ connection

![Figure 2: A modified version of Burgess and Hitch’s (1999) model of phonological working memory](chart)

between input and output phonological representations hypothesised in Section 2.1 features in this diagram: it is the bidirectional link between input and output syllables. This is also the link through which subvocal repetition happens, which allows pre-articulatory outputs to reactivate phonological inputs. There is also an ‘indirect’ link from phonological inputs to outputs through the ‘item’ units, which could correspond to the indirect link between superior temporal and articulatory areas via the inferior parietal cortex. Whether the direct and indirect links between input and output phonology correspond to the arcuate fasciculus and the indirect pathway through inferior parietal cortex is not yet clear, but the inferior parietal cortex is certainly implicated in many studies of phonological short-term memory; see e.g. Paulesu *et al.* (1993); Henson *et al.* (2000).

A final point to note is that recall of a phonological sequence is sensitive to the way syllables are grouped when they are presented. For instance, if syllables are presented rhythmically in groups, the first and last syllables in each group are better recalled than the middle syllables (see Burgess and Hitch, 2005 for a review). Burgess and Hitch (1999) model this finding by assuming that the context representation is two-dimensional, with one dimension representing position of a syllable within a group and another representing serial position of the whole group. ‘Item’ units now store the position of a syllables in a group, as well as the position of a group within a sequence of groups.
3 Lexical representations and processes

Words have phonological, semantic and syntactic aspects. In this section I will discuss these aspects in turn.

3.1 Phonological representations of whole words

The phonological structure of a word consists—for the main part—of a sequence of syllables. Word-sized phonological representations are basically representations of sequences of syllables which occur with particularly high frequency. In the input modality, word forms are regularities in the exposure language which infants become attuned to very early: 8-month-old infants can identify statistical regularities in syllable sequences after only a short period of ‘training’ (Saffran et al., 1996), long before they can reliably associate word forms with meanings. In the output modality, word forms are akin to motor schemata: sequences of articulations which are frequently performed, and through practice become compiled into encapsulated motor routines.

Whether we consider input or output modalities, word forms are evoked in some medium of working memory. Since we have already proposed separate input and output media for phonological working memory, we expect to find separate input and output phonological representations of words. And indeed this seems to be the case. For instance, a patient described by Auerbach et al. (1982) can produce words normally, and can understand written words, but cannot understand spoken words. Conversely, the patient described by Shallice et al. (2000) can understand spoken words, but has difficulty producing words, especially when they are phonologically complex: this patient lacks word-sized phonological output representations. Figure 3 shows a model of phonological working memory extended with representations of whole word forms. In the remainder of this section, I will discuss

![Figure 3: Phonological input and output word representations](image)

...some of the influential computational models of word forms and how they are accessed, in
both the input and output modalities, and refer back to this figure.

### 3.1.1 Phonological input word representations

One way of modelling phonological word representations in the input modality is to extend the model of the phonological input buffer by adding long-term synaptic connections encoding phonological patterns which occur particularly frequently. For instance, Burgess and Hitch (1999) include long-term synaptic connections throughout their model to capture such common patterns. Note that their use of a two-dimensional context representation also allows the representation of multisyllable words. A multisyllable word has the same form as a rhythmically presented group of syllables in a working memory experiment. If we follow Burgess and Hitch’s (1999) proposal that ‘item’ units can store the position of syllables within a group, then we can think of phonological word representations as patterns of long-term synaptic connection which create a bias in the population of item units towards particular syllable sequences. This would explain why it is easier to retain real words in working memory than nonwords, all other things being equal. In this scheme, an active word representation is a particular pattern of activity in the item layer, denoting a particular syllable sequence. In the context of Figure 3, we can think of these patterns as activating a separate layer of whole word representations, through the link connecting ‘items’ to ‘input word forms’. Other models of word representation in working memory have a similar flavour; see e.g. Grossberg and Myers (2000).

Another way of modelling phonological word representations in the input modality is to envisage a network which directly receives a stream of phonological inputs and learns to recognise sequential patterns in this stream. (This kind of model is provided for in Figure 3 in the links directly connecting phonological inputs to input word forms.) The standard network for this purpose is a **simple recurrent network (SRN)** (SRN; Elman, 1990), which is illustrated in Figure 4. The classical SRN takes a sequence of inputs, and for each of these

![Figure 4: A simple recurrent network](image)

is trained to predict the next input in the sequence. The network takes a copy of its own hidden layer at the previous time point as an additional input, which allows it to condition its predictions on an exponentially decaying representation of past inputs, as well as on the current input. To configure a SRN for representing whole word forms, it is typically trained to generate a static representation as output for each word. Thus for each training word,
we present a particular sequence of phonological input signals, and train the network to predict an unchanging output representation which declaratively represents that sequence. A well-known example of this kind of network is that of Gaskell and Marslen-Wilson (1997). Networks with this recurrent design have the interesting property of incrementality: at each point during the presentation of a phonological sequence, they predictively activate all the whole word form representations which start with the sequence provided so far. This mirrors experimental findings about the priming effects of partially-presented words (see e.g. Zwitserlood and Schiefers, 1995). Gaskell and Marslen-Wilson’s model is actually restricted to monosyllabic words; it takes as input fairly low-level phonological features, and generates as output the kind of static syllable representation proposed by Hartley and Houghton. So it could conceivably be a model of the circuitry which learns to map from acoustic inputs to syllable structures (if syllables are thought of as copies of static articulatory plans passed back from the phonological output system. But SRNs have also been used to model the phonology of multisyllable words. For instance, Sibley et al.’s (2008) recurrent network maps sequences of input phonemes straight onto static wordforms, learning a large vocabulary of mainly polysyllabic words. Whatever their precise inputs and outputs, SRN-based models also capture the idea that a speech signal can activate multiple wordforms in parallel, in a medium in which they compete with one another. Broadly speaking, the recurrent circuitry in SRN models plays the same role as the additional temporal context dimension representing ‘serial position within a chunk’ in Baddeley and Hitch’s model.

A final interesting technique used in representing input phonological word forms is a self-organising map (SOM; Kohonen, 1982). A SOM receives training inputs in some high-dimensional space, and learns to represent them in a lower dimensional (typically 2D) space, in a way which positions similar inputs close to one another, and which devotes more space to inputs which appear frequently. If a SOM’s inputs represent complex phonological sequences associated with whole words, it will learn a representation of word forms which places phonologically similar words close together, and which is biased towards recognition of frequently occurring words. A SOM is like a SRN in that it captures patterns in the input data, but it also has the effect of sharpening these patterns. Li et al. (2004) use a SOM to represent word forms. The inputs to their network are declarative patterns representing sequences of up to three syllables, and encoding the onset, vowel and coda of each syllable separately (Li and MacWhinney, 2002). (The vowel and coda together constitute the ‘rhyme’ component of a syllable.) This encoding effectively bypasses much of the machinery discussed so far, and focusses on word-level encodings of phonology: a practical first step in many neural network applications.

3.1.2 Phonological output word representations

Models of phonological word representations in the output modality are models of word production. The two dominant models in this area are those of Dell et al. (1997) and of Levelt et al. (1999). These models target slightly different levels: Dell et al.’s model only produces monosyllabic words, while Levelt et al.’s produces polysyllabic words as well.
Dell et al.’s model can be thought of as a model of the mapping from output syllables to articulations shown in Figure 3. Monosyllabic word units are directly connected to phonemes, which are arranged in three banks: onsets, vowels and codas. A word like cat is connected to c in the onset units, a in the vowel units and t in the coda units. (Each consonant is represented twice, once as an onset and once as a coda.) Importantly, these connections are bidirectional, so that when word units activate, the phoneme units they activate also activate phonologically similar words. This allows the model to simulate phonological errors made by speakers, where they produce words other than the one intended: these errors tend to take the form of phonologically similar words. (Note that this model can also probably account for lexical effects in working memory tasks, though it was not designed for this purpose.)

Levelt et al.’s model of word forms maps polysyllabic words onto low-level phonological outputs. In relation to Figure 3, this model encompasses all the stages from output word forms to articulatory outputs. Importantly, syllabification occurs at a fairly late point during this process. A word-form is associated directly with a sequence of phonemes, represented as a set of positions: for instance the word escort activates the phoneme e in position 1, the phoneme s in position 2 and so on. Particular subsequences of phonemes in turn activate particular syllables. The rationale for this structure is that syllables do not respect word boundaries: for instance pronunciation of the words escort us creates the syllables es-cor-tus. There are many ways in which phonemes could be associated with particular serial positions in a sequence: for instance, we could use Burgess and Hitch’s idea of a dynamically changing temporal context representation, together with item units associating contexts with phonemes, or we could use a SRN which maps a static word-form onto a sequence of phonemes. (For instance, Dell et al.’s earlier 1993 model of word production takes the form of an SRN.) But the choice of representation has consequences on the kinds of psycholinguistic results the model can simulate.

Aside from its extension to polysyllabic words, there are several other differences between Levelt et al.’s (1999) model and that of Dell et al.’s (1997). One important difference is that Dell et al.’s model is geared towards simulating the kinds of errors which speakers make, while Levelt et al.’s model is geared towards accurately modelling the timecourse of speech production. Another difference—which relates to this difference in emphasis—is that Dell et al.’s model envisages feedback from phonemes to word forms, while Levelt et al.’s model does not. Dell et al (2011) give a detailed discussion of these differences.

Note that while there is a direct pathway from input to output syllables in Figure 3, there is no direct pathway from input to output word forms: the equivalences between phonological input and outputs are assumed to be expressed at a level lower than whole words.

3.2 Neural representations of word meanings

A child learns to associate phonological word representations with meanings. What do these meaning representations look like? Linguistics can tell us something about them. For instance, in logical models of natural language semantics in the tradition of Tarski
(1936/1983), word meanings are defined in relation to the role they play in the meanings of sentences: thus, knowing the meaning of the verb chase involves being able to identify all the episodes in the world in which one thing chases another. In more recent empiricist models, the meaning of a word is defined statistically, by a vector of its co-occurrences with other words (see e.g. Pantel and Lin, 2002). These models can learn surprisingly intuitive and fine-grained semantic classifications of words from naturally-occurring text corpora: words with similar meanings are assigned similar representations. In fact, these statistically-defined meaning representations are learned naturally by the hidden layer of an Elman network, if it is given words from a text corpus as input, and trained to predict the next word (see Elman, 1990 and much subsequent work).

Logical and empiricist conceptions of word meaning share the idea that meanings of words are tightly interconnected: the representation of one word’s meaning makes reference to those of other words. However, word meanings also have to be grounded: at some point, we have to be able to tie word meanings to sensorimotor experience of the world. Researchers investigating neural representations of word meanings tend to focus on the meanings of concrete words. The main result from these studies is that concrete words activate sensorimotor areas of the brain—the same sensorimotor areas which are evoked by direct experience of the concepts they represent. For instance, concrete words with strong visual associations but weak motor associations (e.g. nouns denoting animals) tend to activate visual cortices, while words with the opposite associations (e.g. action verbs) tend to activate motor, premotor and prefrontal cortices (Pulvermüller et al., 1999). This is perhaps not surprising—but it is interesting that these sensorimotor responses are elicited very soon after presentation of a word, as early as 200msec. Another study has shown that action verbs denoting actions of the mouth, arm and leg activate the areas of motor and premotor cortex associated with these different effectors (Hauk et al., 2004; Pulverüller et al., 2005), again within a very short interval of word presentation.

One common theme in all these studies is that each concrete word activates a widespread cortical region, encompassing sensorimotor areas but also many other areas. Word-sized semantic representations are not localised in any given area, but distributed through cortex—for concrete words, there is a bias towards the sensorimotor areas most associated with direct experience of the denoted concept. An interesting study bearing out this idea comes from Mitchell et al. (2008). This study makes use of the vector-based word meaning representations which can be derived from word co-occurrence statistics (discussed earlier in this section). Mitchell et al. found that these highly distributed word representations could be mapped to the distributed patterns of brain activity evoked in subjects reading these words. They recorded the fMRI signals associated with each word in a set of concrete nouns, and used these signals to train a function mapping the vector-based representation of a noun onto its associated fMRI signal. They found that this function could quite accurately predict the fMRI signal of an ‘unseen’ word from its vector representation. These fMRI signals extended across the whole brain, but words with associations to particular sensory or motor modalities had predictable activations in the neural areas representing these modalities. Note that all words were of the same grammatical category (nouns), and that the vector-based input representations abstract completely away from the phonology.
of words, so the predicted fMRI activity really reflects the meanings of words, rather than their phonological form or grammatical class.

In summary: we know a little about the form which neural representations of word meanings take, at least in the case of concrete words. The meaning of a concrete word is represented by a highly distributed pattern of neural activity, which includes activity in the sensorimotor areas activated by direct experience of the denoted concept, but also activity in other cortical areas. The distributed nature of word meanings makes them hard to study—however, the technique of Mitchell et al. (2008), using distributed meaning representations derived from co-occurrence statistics, may provide one way of making progress. In the diagrams in the remainder of this paper, for clarity’s sake, I will depict ‘word meanings’ as if they occupy a well-defined neural area of their own.

3.3 Mappings between word forms and word meanings

We have described the neural representations of word forms, and of word meanings: we must now consider the circuitry which connects these two types of representation, and implements a mapping between word forms and word meanings. Figure 5 provides a diagram to refer to in this discussion.

As the figure shows, models of word interpretation and word production are not entirely symmetrical as regards this mapping. Models of interpretation often assume that word meanings are evoked directly from phonological inputs: for instance, Gaskell and Marslen-Wilson’s (1997) incremental model of word interpretation discussed earlier learns not only to predict whole word forms from a phonological input sequence, but also whole word meanings (see the dotted lines in Figure 5). On the other hand, during language production, there is often claimed to be an intermediate unit called a lemma intervening between word
form and word meaning representations, as shown on the right-hand side of Figure 5 (see especially Levelt et al., 1999). A lemma is a linguistic representation of a word, which abstracts away from its form. Levelt et al.’s model of word production has three stages: first a word-sized semantic concept is activated; this in turn activates a lemma; and the lemma activates the form of the word. The three stages are delineated by the dotted lines on the right of Figure 5.

A lemma is an associative representation—something akin to a ‘convergence zone’ in the parlance of Damasio and Damasio (1994). A lemma represents a particular word by connecting independently to a word meaning, and to a word form (a phonological sequence) and a prosodic template (a specification of which of the word’s syllables are stressed). It also connects to a set of syntactic properties, indicating the word’s part of speech, as well as properties such as gender for nouns, or tense for verbs. These syntactic associations are particularly close, being activated at the same stage of the word production pipeline as the lemma itself. Levelt et al. give several reasons for postulating this intermediary representation. One is the existence of ‘tip-of-the-tongue’ states, in which a speaker has a particular word in mind, but cannot access its form (but can sometimes access its prosody). But it would be possible to obtain this effect even if word-sized semantic representations linked directly to output word forms, so it is not strong evidence. Better evidence for lemmas as an independent stratum comes from van Turennout et al. (1997), who found in an ERP study that speakers can access syntactic properties of words (specifically gender) without activating word forms, but that the reverse is not possible.

Are lemmas specific to word production system, or do they also participate in the process of word interpretation? Levelt et al. propose that lemmas are also activated during word interpretation, in the same kind of incremental fashion proposed by Dell et al. for word forms and word meanings. This claim amounts to a claim that there are three routes from linguistic inputs to linguistic outputs: one using a sub-lexical phonological medium (connecting input to output syllables), one using a lexical medium (lemmas), and one through semantic representations. This idea is somewhat at odds with the neuropsychology and neuroimaging literature, which tends envisage a phonological route and a semantic route, but not a lexical one (see e.g. Hickok and Poeppel, 2007). However, there is some evidence for a route through lemma units. Levelt et al. refer to a classic experiment by Glaser and Düngelhoff (1984), in which subjects were asked to name a pictured object, on which a distractor word is superimposed. It was found that semantically related distractor words have more of an effect than semantically unrelated words—for instance naming a pictured chair is delayed more by the distractor word BED than the distractor word FISH. One might wonder why the effect is inhibitory rather than excitatory, given that the associative relationships between the concepts ‘chair’ and ‘bed’ are presumably of equal strength. Levelt et al.’s proposal is that a distractor word can directly activate an output lemma (BED) to compete with the required output lemma (CHAIR), while its facilitatory effect through associated semantic concepts is indirect. Indeed, if the experiment is set up so that the distractor word is not one of the possible words to be spoken, its net effect is facilitatory. These experiments provide further evidence for an independent layer of lemmas, and moreover, for Levelt et al.’s suggestion that the same lemmas participate in
Considering how neural circuits are formed, it seems likely that lemmas have some existence as a representational medium in their own right. Connections between neural regions are profligate: if associations are learned between several distinct neural assemblies, it is likely that some additional population of neurons comes to participate in the resulting grouping, because it happens to be independently connected to each of these assemblies. In other words, postulating a layer of lemmas does not exclude the possibility of direct links between meanings and word forms.

Where in the brain is the circuitry which connects word forms and word meanings? Given that word meanings are widely distributed, we should expect this circuitry to be similarly widely distributed—however, there is evidence that the temporal cortex is particularly important in implementing these mappings, in both directions. Damage to the left superior temporal gyrus (somewhat informally termed ‘Wernicke’s area’) often causes difficulty in understanding the meaning of words. It can also cause difficulty in producing meaningful words—in the classic (but rather rare) condition termed ‘Wernicke’s aphasia’, patients produce sentences which are syntactically and phonologically well-formed, but whose words lack meaning (see e.g. Brookshire, 1997). Other areas of temporal cortex are also implicated in the mapping between word forms and word meanings. For instance, Damasio et al. (1996) found that damage to different areas of left temporal cortex correlated with deficits in naming different semantic types of object: damage to anterior, inferior and posterior left temporal cortex correlated with impairments in naming people, animals and tools respectively. Lu et al. (2002) found that damage to anterior temporal cortex correlated with impairments in naming actions.

Note that the circuitry which maps input word forms onto meanings must be somewhat distinct from that which maps meanings onto output word forms, according to our model of input and output word forms. We expect some overlap, to the extent that input and output syllable representations connect with one another. But there is still good evidence that input and output meaning-form mappings are somewhat distinct. For instance, Kay and Ellis (1987) describe a patient who can understand concrete words, but has difficulty generating these same words in a picture naming task. This patient seems to have a relatively intact phonological output system, because he can repeat spoken words quite well—even those words he cannot spontaneously generate. This patient’s deficit appears to be quite specific to the circuitry which maps semantic representations onto output word forms. Auerbach et al.’s (1982) patient described earlier has the converse condition: he can produce words normally, but cannot understand spoken words. His deficit could be in the circuitry representing the form of incoming spoken words, but it could equally well be in the circuitry linking these form representations to meaning (or lemma) representations.

3.4 Morphology

Words often carry inflections signalling their syntactic properties. For instance, in the English sentence *He walks*, the verb *walks* is made up of the word stem *walk*, plus the inflection *-s*, which signals syntactic agreement with the subject *he*. *Walk* and *-s* are termed mor-
phemes: they are phonological units in the word which carry their own meanings and/or syntactic properties. The branch of linguistics which investigates how words decompose into morphemes is called morphology.

Some inflections carry rich semantic information—for instance, tense inflections in English, or subject agreement inflections in Romance languages (which carry enough information that pronominal subjects can often be omitted altogether). These inflections are very much like normal words: we can envisage them as having lemmas of their own. But inflections are normally also syntactically obligatory, even when they do not carry much information of their own: the English agreement inflection -s is a case in point. I have already briefly touched on the syntactic role of inflections in the discussion of lemmas in Section 3.3; in the current section I will just consider the nature of the neural circuitry which interprets and generates inflected words.

We must envisage circuitry in the input modality which decomposes an incoming word into its constituent morphemes, and circuitry in the output modality which assembles morphologically complex words. In either case, there is some evidence that frontal circuitry is involved. Patients with damage to left inferior frontal cortex (regions including, but not limited to ‘Broca’s area’ comprising left Brodmann areas BA44–45) often have difficulty generating inflected words, and often fail to notice inflection errors in incoming sentences, especially in cases where inflections do not carry much semantic information. There is interesting evidence that the production of inflected verbs involves a region of left dorsolateral frontal cortex, part of BA9. Shapiro et al. (2000) find that TMS over this area selectively impairs the production of inflected verbs; and Shapiro et al. (2001) describe a patient with damage to this same area who has a permanent selective deficit in producing inflected verbs.

The phonological form of inflected words is governed by a complex system of rules and exceptions. For instance, the morpheme -s is realised as s on the stem walk (to form walks), but as es on the stem catch (to form catches). Some of the most venerable connectionist models of language processing are models of circuitry which learns these rules from a training corpus of examples: for instance, Rumelhart and McClelland (1986) proposed an influential model of the circuitry which learns the past tense form of verbs. Their network learns to map a phonological representation of a verb root onto a phonological representation of the appropriate past tense form, using a supervised training scheme. The network appeared to show a similar developmental profile in the task as real children: early in training it learned past tense forms on a word-by-word basis; later it learned general rules (and over-applied these rules, for instance producing knowed as the past tense of know); and finally it learned which words to treat as exceptions. While there have been many more sophisticated models of this process, the basic idea that a connectionist network can successfully learn a mixture of rules and exceptions is now very widely accepted.
4 Sentence-sized semantic representations

Before we consider how syntactic processing is implemented in the brain, we must first consider how the meanings of sentences are represented. Syntactic processing maps between sentences (represented as sequences of words) and their meanings. We have discussed how the brain represents word sequences in working memory. Before we discuss syntax, we must say something about how the brain represents sentence meanings in working memory, so we can define the computation which syntax performs. (We need to know what syntax is before we can look for it in the brain!) In this section, I will consider what we know about ‘semantic representations’ in the brain. My approach will be to use formal theories of natural language semantics as a guide. In Section 4.1 I will introduce some concepts which have proved useful in formal models of sentence semantics. In Sections 4.2–4.3 I will consider the cognitive representations which best correspond to these concepts, drawing mainly on neural theories of working memory and long-term memory. My proposals about cognitive representations of meaning will be broadly in line with those proposed in ‘cognitive linguistics’ (see e.g. Lakoff and Johnson, 1980; Langacker, 2008), especially recent models which hold that entertaining the meaning of a sentence amounts to simulating the process of experiencing the episode it describes (Gallese and Lakoff, 2005; Barsalou, 2008). However, the arguments I adduce for this conclusion, drawing on theories of memory, are largely my own. In Section 4.4 I will discuss various schemes for representing the relationships between concepts which are expressed in sentence meanings.

4.1 Formal models of sentence and discourse semantics

Formal semanticists sometimes refer to a sentence-sized meaning representation as an episode. The term ‘episode’ is intended to encompass two ontologically distinct types of semantic entity: one the one hand a state, and on the other, an event. A state is a proposition which can be classed as true or false at any given single point in time: for instance, MY_CAT_IS_GREY. An event describes an operation which brings about a change in the world: for instance, MY_CAT_ATE_MY_SANDWICH. In the standard model of semantics devised by Reichenbach (1947), the meaning of any episode must be represented in relation to a reference time, which denotes ‘the time the speaker is currently talking about’. When a hearer assimilates an event, she must update the reference time, to accommodate the fact that events change the world: after an event takes place, certain things stop being true, and others start being true. On the other hand, when assimilating a state, the hearer should not update the reference time: states report facts which obtain at the reference time, and thus update our knowledge about this point in time, but not the temporal reference itself.

There are elegant formal theories about the nature of temporal update operations (e.g. Reiter, 2001), and of the kinds of statements about time which natural language sentences can express (see e.g. Steedman, 1997). There are also elegant theories about how interpreting a sentence updates a hearer’s knowledge of the world (see e.g. Kamp and Reyle, 1993; Heim, 1982). A key suggestion in both these bodies of theory is that the
meaning of a sentence is best thought of not as a direct representation of some small part of the world, but as a function, which takes a representation of the world as input, and delivers a new representation as output. If the sentence denotes a state, the new representation provides new information about the current reference time; if it denotes an event, the new representation updates the reference time and provides information about the situation which obtains at the new time. The representations which are updated are often termed discourse contexts. These representations are rich, containing far more information than single sentences. Formally, a discourse context contains a set of discourse referents which are available to talk about, and a structure of conditions, which represent the sentence-sized things which have been said about these referents in the discourse so far. There may be many referents, and many conditions.

What are the neural equivalents of states, events, reference times, discourse contexts, referents and conditions? There are some very natural neural correlates of these concepts, which draw on models of human memory, as I will describe in the next two sections.

4.2 Episodic long-term memory representations

A good starting point is Tulving’s (1972) well known model of episodic memory. An agent’s episodic memory is her memory for specific episodes she has experienced in her own life, either directly through perception, or indirectly through stories, films and so on. Episodic memories are tied to particular places, and tend to have well-defined temporal structure: they take the form of narratives, in which one event leads to a subsequent event, and they are naturally recalled by recreating this structure. Computational models of episodic memory assume that this form of memory makes use of neural ‘context’ representations which denote particular places and/or moments in time. We know something about these context representations: for instance, there is good evidence that spatial context representations are stored in the parahippocampal cortex, both in rodents (Eichenbaum et al., 2007; Diana et al., 2007) and in humans (Epstein and Kanwisher, 1998; Epstein et al., 2007), and that temporal contexts are also stored in this parahippocampal area (Bar and Aminoff, 2003; Howard and Kahana, 2002). Hippocampal representations of temporal contexts are a natural analogue of Reichenbach’s notion of temporal indeces, and the episodic memories which are linked to these contexts are natural analogues of the sentence-sized states which obtain at particular temporal indeces, or the events which transition between temporal indeces.

When a hearer interprets a narrative delivered in natural language, she ultimately encodes this narrative in episodic memory. Episodic memory is a form of long-term memory: if the hearer is asked about the events in the narrative some time later, she can query her episodic memory to retrieve the answers. In the standard model of episodic memory storage (see e.g. Marr, 1971; McClelland et al., 1995), an episode is initially stored in the hippocampus and neighbouring cortical areas, where it is associated with a spatiotemporal context representation to individuate it. Gradually, this hippocampal episode representation is transferred to (even) longer-term storage in a more widespread cortical network, in a process called ‘consolidation’.

19
4.3 Working memory episode and context representations

Of course, it is also important to consider how the meanings of the sentences in a narrative find their way into hippocampal storage in the first place. The standard proposal here is that in order to store an episode in the hippocampus, it must first be represented in working memory. There is good evidence that if an episode is actively elaborated in working memory, it is recalled better in the long term (see Craik and Lockhart, 1972 and many subsequent studies). We have already discussed working memory for phonological sequences—see Section 2.2. However, the working memory medium in which episode representations are maintained is quite distinct from this phonological storage medium. It is semantic in nature, and associated primarily with activity in the prefrontal cortex (PFC). Evidence for prefrontal involvement in the process of encoding episodic memories comes from several sources. For instance, both ERP and fMRI studies have shown that the level of PFC activity at the time a stimulus is encoded predicts how well it will later be retrieved (Rugg, 1995; Rotte et al., 2000). There is also evidence that material retrieved from long-term memory is re-established in prefrontal working memory (see Buckner, 2003 for a review).

An interesting recent proposal by Baddeley (2000) is that the working memory medium which communicates with hippocampal long-term memory is specialised for holding semantic episode representations. Baddeley calls this working memory medium the episodic buffer, and gives it a role in sentence processing. I will argue that the episodic buffer is the cognitive medium in which sentence meanings (‘events’ and ‘states’, in semanticists’ terms) are maintained in working memory. This idea makes intuitive sense, but it also fits well with a broader model of updates in working memory which I will outline below. Recall that semanticists think of events and states as update operations, which operate on rich, multifaceted ‘discourse context’ representations. If we want to think of the episodic buffer as the medium which holds events and states, we must also propose some analogue of discourse context representations, and suggest how material in the episodic buffer can function to update these representations.

I will start by considering the analogue of discourse context representations. To begin with, note that the episodic buffer is not the only semantic representation in working memory. There is also a more holistic working memory representation known as cognitive set, which Baddeley is at pains to distinguish from the episodic buffer. An agent’s cognitive set specifies the rules which govern her actions at the current moment; formally, it defines a mapping from the stimuli she perceives to the actions she performs. The consensus is that cognitive set is also implemented by activity in the prefrontal cortex—see Miller (2000) for a good articulation of this idea. Cognitive set can be thought of as an action-based representation of the agent’s current situation—in the terminology of Gibson (1950), we can think of it as an encoding of the actions which the situation ‘affords’. If we define actions broadly enough to include actions of attention to objects, cognitive set can also encompass the agent’s working memory for the salient objects in the current situation. This kind of memory is also known to be prefrontal: there is evidence that working memory representations of task-relevant objects are maintained in PFC, both in monkeys (e.g. Rainer et
al., 1998; 1999) and in humans (e.g. Courtney et al., 1998; Haxby et al., 2000). I suggest that an agent’s cognitive set is a good cognitive analogue of semanticists’ notion of ‘the current discourse context’. A cognitive set includes representations of recently encountered objects, and of the relevant properties of these objects. It does not represent these properties as explicitly as the discourse contexts of semanticists: they are not enumerated as conditions, but rather compiled into a rich set of action affordances. But their purpose is ultimately the same: semanticists envisage reasoning procedures which draw the kinds of conclusion about the current situation and the best actions to perform which are compiled into the cognitive set.

Just as semanticists talk about updates to the current discourse context, behavioural psychologists talk about updates to an agent’s cognitive set. One of the interesting properties of cognitive set is that it can be changed suddenly (at least in humans and higher mammals). Updates come from new information about the environment. Often this information makes only minor changes to cognitive set. If we think of an agent’s ‘set’ as representing her current task or focus, many episodes she experiences have little or no impact on her cognitive set: one of the points of a task representation is to help the agent ignore incoming episodes which are not currently relevant. But sometimes, incoming episodes are important enough to update or change the agent’s current goals. The interesting thing about these updates is that they are relatively fast and discrete: humans are very good at switching from one cognitive set to another (see Miller and Cohen, 2000 for a good model of this switching process.) The fact that cognitive set supports update operations emphasises its suitability as the cognitive analogue of semanticists’ notion of discourse context.

Now I will consider the relationship between cognitive set and the episodic buffer. If the episodic buffer holds events or states, which semanticists model as operations updating the current discourse context, and if cognitive set represents ‘the current discourse context’, we expect the contents of the episodic buffer to be implicated in bringing about updates to the cognitive set. More precisely, we expect the episodic buffer to implement a function mapping the agent’s current cognitive set onto an updated set. A function implements a particular mapping from inputs to outputs, but it also has a domain of definition: it creates a new cognitive set given the current set. As regards the first of these roles, there are good accounts of the mechanisms through which current experiences update PFC-based cognitive set; see in particular Braver and Cohen (2000). The mechanisms involve a special kind of reinforcement learning: an agent learns to update PFC in ways which have beneficial consequences some time later. As regards the second role, we must explain how an agent’s current cognitive set influences or determines the material coming into the episodic buffer. If we think of an agent’s cognitive set in Gibsonian terms, as the set of actions which are afforded in the current situation, then one natural way to think about the episodic buffer is as the medium in which these actions compete to be selected. This makes sense of the idea that the episodic buffer holds ‘just one episode’, but also shows how this one episode relates to the agent’s wider representation of the current situation. Note that once an action is selected in the episodic buffer, it is also executed: it stops being a potential action, and becomes an actual one. In an agent interacting with the world,
activating a representation in the episodic buffer actually changes the current situation—or, in the case of a perceptual action, the agent’s knowledge about this situation. In either case, the change triggers a change in the agent’s cognitive set.

In summary, many of the key concepts in a formal model of sentence semantics have natural correlates in a theory of working memory. Entertaining the meaning of a sentence is analogous to evoking an episode representation in the episodic buffer. Activating a particular representation of ‘the current discourse context’ is analogous to adopting a particular cognitive set. If we accept these analogies, the proposal from formal semantics that ‘sentence meanings are context update functions’ fits well with an account of how working memory representations bring about behaviour in the world.

Note that the notion of a context update can also be connected to the account of long-term memory described in Section 4.2. We have to envisage particular points during experience of the world where information in the episodic buffer is transferred to longer-term hippocampal storage. Since there are representations of context in both working memory and episodic long-term memory, these transfer points should coincide with points at which working memory context is updated. If we consider a particular situation, where a particular episode is evoked an agent’s episodic buffer, we should envisage two operations—one which updates the agent’s current cognitive set, and another which encodes the update operation in episodic memory (perhaps updating the hippocampal context representation in the process). This second operation should obviously happen before the episodic buffer is cleared to set the stage for a decision about the next action.

4.4 Binding mechanisms in episode representations

Having given a high-level picture of the semantic memory structures involved in representing a sentence, I will now consider the details of individual episode representations. For concreteness, I will focus on the format of episode representations in the episodic buffer.

An episode representation must combine several different conceptual elements: for instance a state like MY_CAT_IS_GREY combines at least the concepts ME, POSSESSION, CAT, PRESENT, GREY (and possibly others too); an event like JOHN_JUMPED combines at least JOHN, JUMP and PAST. Of course, an episode representation is more than just a bag of concepts: its component concepts are combined together in well-specified ways (which are reflected in the syntax of sentences which express it). A well-established idea in linguistics is that any episode representation takes the form of a set of role-filler pairs, identifying the type of the episode, and the roles played by the objects or entities which participate in it (see e.g. Jackendoff, 2002 for a good presentation of this idea). So, for instance, we might represent JOHN_CHASED_MARY by associating the role ACTION_TYPE with the filler CHASE; actions of this type require AGENT and PATIENT participant roles, which in our case are filled by JOHN and MARY respectively. There is a good consensus that we need a scheme of this kind to represent sentence meanings. A key question for computational neurolinguistics is how such a scheme can be implemented in neural circuitry. In computational terms, what is needed is a mechanism for expressing binding relationships between roles and fillers. This mechanism must be able to flexibly bind fillers to roles, and
to express a fairly large number of binding relationships. Moreover, there must be some
degree of hierarchy in the representation of fillers, which allows fillers to be units which are
themselves specified by role-filler pairs. (For instance, in the episode my cat jumped, the
AGENT role is occupied by an object whose TYPE is cat, and whose POSSESSOR is ME.)
The desiderata for a neural model of binding are well set out by Jackendoff (2002). There
are many proposals about how a binding scheme can be implemented in neural hardware,
and as yet no consensus, but I will briefly outline some of the key proposals.

The binding-by-space model represents a proposition in a set of units divided into
groups, each of which is explicitly associated with a particular role: for instance, the first \( n \)
units might be used to represent the filler of the AGENT role, the next \( n \) units to represent
the filler of the TYPE role and so on. This model is something of a straw man. It requires
each concept to be represented multiple times—for instance, we would need separate rep-
resentations of cat to fill AGENT and PATIENT roles. This is not just representationally
expensive (there are lots of roles to fill, especially in nested propositions)—it also fails to
capture the commonalities between concepts associated with different roles. (A cat par-
ticipating in an episode as an agent has much in common with a cat participating as a
patient.)

The binding-by-synchrony model (see e.g. Shastri and Ajjanagadde, 1993) Makes
reference to the periodic patterns of activity found in neural populations (revealed, for
instance, in EEG signals). Different periodic patterns are found in different brain regions
at different times; binding-by-synchrony typically refers to the theta cycle, which has a
frequency of 4-8Hz. The model is motivated from the observation that populations of
neurons often fire at a particular phase in the brain’s theta cycle. It proposes that roles
and fillers are bound together by their activation at a particular phase of the theta cycle.
This allows a single cat concept to participate both as an agent and as a patient. While
there is now good consensus that precise timing of neural spikes is important (see e.g.
Izhikevich et al., 2004), the idea of binding by synchrony has many problems; some of
these are summarised by Shadlen and Movshon (1999). In any case, there are not enough
phases in a theta cycle to hold the number of bindings required in representing the meaning
of even moderately complex sentences.

The binding-by-connection model envisages that there are synaptic links between
each role and each possible filler for that role, and that associating a filler with a particular
role can be achieved by temporarily potentiating the appropriate connection, so that activ-
vating the filler activates the role and vice versa. This model also allows a single concept
to be associated with different roles. The model has problems expressing the semantics of
nested sentences featuring multiple copies of some roles—for instance, in The boy chased
the girl who ran, the boy and the girl are both represented as agents (of different actions).
But some sentence processing models put the topic of nested sentences to one side, and
employ binding-by-connection models quite effectively (see e.g. Chang, 2002). Other mod-
els (e.g. Hummel and Holyoak, 2003) resort to a binding-by-synchrony account to explain
how multiple propositions can be stored.

The neural blackboard model (van der Velde and de Kamps, 2006) is an extension
of the binding-by-connection idea, which envisages a pool of dedicated binding units, each
of which is linked to every possible role and every possible filler, and can temporarily store an association between a particular role and filler. Dedicated binding units considerably reduce the number of connections which must be envisaged, and also allow for hierarchical objects to feature as fillers. However, it is unclear whether there are neural assemblies in the brain which have the right properties to function as binding units. There must be a relatively small pool of such units, and each of these must be fully connected to a wide range of other neural media. It is important that each binding unit can hold a connection between every possible role-filler pair — otherwise there will be certain bindings which simply cannot be expressed. What is more, there are some indications that the amount of neural circuitry required to implement binding units does not scale up to realistically sized sets of atomic concepts (see e.g. Stewart and Eliasmith, 2012).

The binding-by-serial-position model (Knott, 2012) makes reference to the proposal that experiencing an episode involves a canonical sequence of sensorimotor operations. For instance, experiencing a man grabbing a cup canonically involves attention to the agent first, then attention to the cup, and then a process of classifying or monitoring the grab action (see also Ballard et al., 1997; Goldin-Meadow, 2008). If this is the case, then we can store an episode in working memory as a prepared sensorimotor sequence, in which particular thematic roles are identified by particular serial positions. We know a great deal about how prepared attentional and motor actions are stored in the prefrontal cortex; see e.g. Averbeck (2007); Rhodes et al. (2004): Knott’s model assumes that episodes are held in prefrontal working memory in a similar format. There are limitations of this model: in particular, it does not yet address how nested propositions are stored in working memory. But the idea that entertaining a proposition involves rehearsing a sequence of representations does provide some interesting scope for an account of nested propositions, because it allows that representations of different propositions could be dominant in PFC at different points during this rehearsal process.

Another class of binding models use distributed representations to hold bindings between roles and their fillers. In these models, the representation of a binding relationship is held in a pattern of activation distributed over a whole ensemble of units, in which representations of the individual items being bound are hard to identify. An early example of the distributed approach is that of McClelland et al. (1989), which can be termed binding by query training. In this model, a sentence is fed into a SRN-like network one word at a time, to create a ‘sentence gestalt’ on the output layer. This gestalt then forms the input to a second network, which learns to answer questions about the binding of roles and fillers in the encoded sentence. The second network takes two inputs—the sentence gestalt itself, and a single role or filler which functions as a query—and is trained to generate the item which is bound to this role or filler as output. Training updates the weights in the recurrent network as well as in the query network, so the recurrent network learns to generate sentence gestalts which support correct answers about the binding relationships between roles and fillers. A more complex model using a similar technique of query training is presented by Rohde (2002), which supports the encoding of nested propositions. In this model, queries carry richer information: rather than just naming a role or a filler, queries are complete propositions with one missing element.
The **vector symbolic** model of binding (Gayler, 2003; Plate, 2003) capitalises on the idea that the atomic components of episode representations are expressed in some distributed, high-dimensional feature space, as suggested in Section 3.2. I will describe Plate’s model here. Plate’s proposal is that binding two representations together involves performing an operation which takes their associated \( n \)-dimensional vectors as input, and returns a new vector in the same \( n \)-dimensional space as output. The important property of the operation is that it returns a vector which does not closely resemble either of the representations being bound together: the \( n \)-dimensional space is assumed to be big enough to be able to hold a large number of atomic concepts, and to represent the bindings of arbitrary role-filler pairs as distinguishable vectors. One attractive property of this scheme is that it naturally supports hierarchical binding operations: since the vector representing the result of a binding operation is in the same \( n \)-dimensional space as the concepts being bound, we can apply the same sort of operation to bind this derived representation with another concept. This binding model scales more naturally to large repertoires of concepts (Stewart and Eliasmith, 2012). Assessed on its computational properties, it is an attractive model of neural binding. It is very hard to seek evidence for it, since representations are large distributed patterns of activity. One strong prediction is that the distributed representation of the binding of concepts \( A \) and \( B \) bears no resemblance to the representation of either \( A \) or \( B \) by themselves. Many studies of PFC-based working memory find that working memory representations of bound concepts tend to include identifiable representations of the individual concepts being bound (see e.g. Rainer et al., 1998). But these studies also tend to find that there are cells which encode specific combinations of bound concepts, which are what is predicted by the model. Perhaps better tests could be devised using the methodology of Mitchell et al. (2008), which maps distributed word representations to whole fMRI brain scans.

A final attractive model of neural binding is INSOMNET (see e.g. Mayberry and Mikkalainen, 2008). This scheme combines elements of distributed and blackboard binding models. It draws on a sophisticated symbolic scheme for representing sentence meanings, called **Minimal Recursion Semantics** (MRS; Copestake et al., 1999). In MRS, a sentence meaning is expressed as an unstructured set of frames. Each frame has a label or **handle** \( (h_n) \), and introduces one semantic element of the sentence, encoding the relation of this element to other semantic elements by binding relevant roles to the handles of other frames. (Thus, for instance, a frame with handle \( h_0 \) introducing the action concept HIT might bind the **AGENT** and **PATIENT** roles to the handles \( h_1 \) and \( h_2 \), which identify frames holding information about the agent and patient respectively.) INSOMNET has a fixed-size set of assemblies encoding prototypical frame structures: for instance, frames introducing transitive and intransitive actions, and frames introducing referential elements of different kinds might all be represented by different frame assemblies. Assemblies represent frame structures in a distributed manner: they are subsequently translated into localist representations by a ‘decoder’ network. (The decoder network is constrained to use the same weights to decode each assembly, so similar assemblies are forced to use similar distributed schemes for representing frame structures.) Importantly, while each assembly represents a frame with a particular structure, it can represent many different token assemblies with
this structure. (For instance, the same assembly might be able to represent a HIT or an
GRASP action concept, because these concepts have a similar structure.) The front-end of
INSOMNET is a fairly standard SRN, which takes a sequence of words and produces a pat-
ttern of activation over the set of frame assemblies, which are then translated into a set of
localist frame representations by the decoder network (aided by a mechanism for selecting
particular frames which I will not discuss here). The whole network is trained in a super-
vised manner: for each sentence, a sequence of words is presented as input and the network
is trained to produce the associated set of localist frame representations as output. (An
interesting feature of INSOMNET is that it is trained on a very large corpus of sentence-
meaning pairs, created automatically by a wide-coverage symbolic sentence interpretation
system and postprocessed by hand.) As well as supervised training, the network also uses
a self-organising map to structure its set of assemblies into groups which represent similar
frame structures. The trained network is very good at identifying the structure of the set
of MRS frames representing unseen sentences, though somewhat less good at identifying
the token concepts which fill these structures. Its strength is in its ability to learn a set of
frame representations which collectively suffice to model the structure of a large range of
real sentences in a semantically sophisticated way.

5 Syntactic representations and processes

In this section, I consider how syntactic processing is implemented in neural circuitry.
I begin in Section 5.1 by defining the processing which will be under discussion, and
in Section 5.2 I give some broad ideas about the neural areas where this processing is
implemented. In Section 5.3 I will discuss the dominant idea in connectionist models of
syntax: that syntactic processing is carried out by circuitry in which a simple recurrent
network plays an important role. In Section 5.4 I will discuss some limitations of standard
SRNs as models of syntax, and describe some ideas about how to extend or modify SRNs
to overcome these limitations.

5.1 Conceptions of syntax

The syntax of a natural language can be defined in two ways. On the one hand, we
can think of the syntax of a given language as a specification of what counts as a ‘well-
formed’ sentence in that language. On the other hand, we can think of the syntax of a
language as the principles which determine how the meanings of the words in a well-formed
sentence are composed together to yield the meaning of the whole sentence. These two
conceptions of syntax are linked, in that they both require well-formed sentences to be
represented as hierarchical structures, rather than as flat lists of words. The number of
well-formed sentences in any language is infinite, for all practical purposes. This means
that we must model well-formedness by defining a system of rules or principles which sentences
must conform to, rather than by enumerating individual sentences. In traditional models
of grammar, this system defines a sentence as being composed of phrases, or groups of
words. Phrases are defined recursively as composed of sub-phrases of their own, a definition terminating in elementary phrases consisting of single words. Phrases are also implicated in the conception of syntax which supports semantic interpretation of sentences: the phrase structure of a sentence is the guide which indicates how the meanings of its words combine together. In traditional models of compositional semantics, each phrase has a semantic interpretation, which is formed by combining the meanings of the phrases from which it is formed. The key question in theoretical syntax is therefore how to state the rules or principles which specify how sentences decompose into phrases.

One of the interesting debates concerns to what extent these rules abstract away from the lexical items which actually make up sentences. In one school of thought, some core set of syntactic rules is provided innately: infants arrive in the world ‘knowing’ certain (quite abstract) facts about the syntax of all human languages, and only have to learn about the syntactic idiosyncracies of the particular language they are exposed to. This idea is associated with the Chomsykan syntactic tradition (see e.g. Chomsky, 1981; 1989; 1995). In this tradition, the core syntactic rules make no reference to words whatsoever. In another school of thought, infants learn all the syntactic principles of their native language by finding patterns in the sentences they are exposed to: in other words, they learn a more or less complete model of syntax from their exposure language. This idea is associated with constructivist linguists (see e.g. Goldberg, 1995; Tomasello, 2003). In this tradition, rules are created by a mechanism which takes concrete sentences as input, and there is nothing to prevent them making reference to specific words. Constructivist theories of grammar are good at modelling ‘surface’ patterns in language, such as those found in idiomatic or semi-idiomatic constructions (e.g. John let the cat out of the bag; Letting the cat out of the bag might be a good idea), as well as graded notions of grammaticality (e.g. the fact that some sentence structures are ‘unusual, but not impossible’). They are also good at modelling the trajectory of syntactic development in infants; there is evidence that infants’ earliest multi-word constructions are structured around specific words, and that it takes some time for properly abstract syntactic constructions to develop (see again Tomasello, 2003).

5.2 Neural areas involved in syntactic processing

There is some evidence that the neural circuitry which represents well-formedness of sentences is distinct from that which maps word forms onto word meanings. Patients with damage to Broca’s area and surrounding left inferior frontal cortex have difficulty generating word inflections, as noted in Section 3.4, but also have a wider difficulty generating syntactically well-formed sentences, and also in interpreting sentences, at least if they have complex syntactic structure (see e.g. Saffran, 2000). At the same time, if damage is limited to these areas, patients’ knowledge of word meanings is often quite well preserved. The opposite pattern of damage is associated with damage to left posterior superior temporal cortex, as already noted in Section 3.3. Some patients with damage in this area produce sentences which are fluent and syntactically well-formed, but which fail to convey any meaning at all; they are also unable to understand single words, or to meaningfully produce
single words. Damage in these cases is assumed to be to the circuitry which maps between word forms and word meanings, while the circuitry which encodes well-formed sentences remains intact. This classic double dissociation sees syntactic processing implemented in left inferior frontal cortex.

However, more recent work also implicates the left anterior superior temporal cortex in syntactic processing, particularly syntactic interpretation; in fact, syntactic interpretation deficits are more strongly associated with damage to this area than with damage to left frontal areas (Dronkers, 2004). This area is activated more by syntactically well-formed utterances than by lists of words (see e.g. Friederici et al., 2000), and more activated by sentences reporting events than by nonlinguistic auditory stimuli depicting events (Humphries et al., 2001). In summary, there is syntactic circuitry in the brain in left anterior temporal cortex as well as in left inferior frontal cortex.

5.3 Basic SRN-based models of syntax

In neural models of language, the key question concerns how syntactic knowledge is represented in neural circuitry. Again, we can think of syntax in two ways: either as a specification of well-formed sentences, or as the principles which allow word meanings to be composed into full episode representations. I will give examples of models emphasising each of these conceptions of syntax.

While traditional linguistics represents phrases symbolically, most connectionist models of syntax represent them subsymbolically, or implicitly. Nearly all connectionist models have an empiricist flavour: they are set up to learn syntactic rules from training data, rather than hardwired to encode particular rules. The dominant idea in recent connectionist models is that an implicit representation of phrase structure is learned by something akin to a simple recurrent network (SRN). The network receives (or generates) sentences as flat sequences of words; through exposure to many sentences during training, it learns something about phrase structure in the weights which map its input and context units to its hidden layer.

To explain this idea, I will illustrate by describing Elman’s (1990) original model of syntactic processing. This model emphasises the conception of syntax as a specification of well-formed sentences. It takes the form of an SRN which receives the words in a sentence one by one, and in each case is trained to predict the next word. Of course there are many possible next words; what the network really learns to predict is the probability distribution for the next word, given the words in the sentence so far. This probability distribution embodies a graded notion of grammaticality: rather than a binary distinction between words which can legally occur next and words which cannot, the network represents which words are most likely to occur next, combining syntactic information with semantic information about what situations commonly occur (or are commonly reported). It also incorporates a model of commonly occurring surface patterns in language such as idioms.

How does a trained word-predicting SRN come to incorporate a model of phrase structure? To explain how this happens, it is useful to think of the SRN as a function which steps...
through a sequence of context representations when given a sequence of words as input. This sequence can be interpreted graphically as a trajectory defined in the $n$-dimensional space of the network’s possible context unit activations. Information about phrase structure is present in the trajectories which the network learns. To illustrate, imagine a very simple language, in which a sentence is made up of two phrases $P_1$ and $P_2$ (in that order), and where there are no dependencies between these two phrases, so any instance of $P_1$ can be followed by any instance of $P_2$. An SRN trained on sentences from this language will learn to establish the same context representation after being presented with any sequence of words instantiating $P_1$, so that its subsequent predictions (about $P_2$) are properly independent of the particular instance of $P_1$ that it has just seen. The trained SRN represents the phrase boundary between $P_1$ and $P_2$ by having all its trajectories in context-unit space pass through the same point when this boundary is reached. (In more realistic situations where there is some degree of syntactic dependency between two phrases, so that particular varieties of $P_1$ require particular varieties of $P_2$, the SRN will learn different invariant points for the various different varieties, but the idea of a phrase boundary as a point where trajectories in context-unit space converge still carries over to these situations.) Importantly, the representations of phrase structure learned by an SRN are only approximations of the notion of phrase structure defined in symbolic models of grammar. I will return to this point in Section 5.4.

Other SRN-based models of syntax emphasise the role of syntax in representing the compositional semantics of sentences, i.e. in defining a mapping between individual word meanings and the meanings of whole sentences. These models either receive or produce a sequence of words, but they are trained by associating word sequences with episode representations. In a model of sentence interpretation, the SRN receives a sequence of words as input, and learns to produce an episode representation as output. We have already seen examples of this kind of network: the network models of McClelland et al. (1989) and of Mayberry and Miikkulainen (2008) both have this structure. In a model of sentence generation, the SRN takes an episode representation as input and produces a sequence of words as output: an example of this kind of model is Chang (2002). The purpose of the context layer in these networks is no longer to explicitly represent well-formed sentences, but to encode a mapping between word sequences and sentence meanings. However, these networks still learn something about the phrase structure of the training sentences they are exposed to, in the way already described.

One attractive feature of an SRN as a model of syntactic processing is that the circuitry it requires is amply attested in the frontal and temporal neural areas where syntactic processing is thought to happen. An SRN is a simple recurrent circuit, and there is evidence of recurrent loops throughout cortex (see e.g. Alexander et al., 1986): these loops pass through subcortical areas, and return back to the same area of cortex from which they originated. There is also evidence that inferior frontal areas are involved not only in syntactic computations, but also in sequence processing more generally. Not all sequence processing depends on inferior frontal cortex—for instance, Broca’s aphasics can learn simple sequences of stimuli (Goschke et al., 2001). However, damage to inferior frontal cortex impairs the learning of sequences with hierarchical structure, and of sequences of
abstract symbols (Conway and Christiansen, 2001; Dominey et al., 2003; Hoen et al., 2006). As I will discuss in the next section, a basic SRN has to be augmented with additional circuitry in order to allow it to learn more complex sequences of these kinds. In the remainder of this discussion, I will focus on computational arguments about the additional circuitry needed; ideas about the neural plausibility of this circuitry are still quite sketchy.

5.4 Limitations of standard SRNs, and some alternative models

As noted above, the way a standard SRN represents phrase structure is just an approximation of the way phrase structure is represented in a symbolic model of syntax. In this section I will discuss two of the computational shortcomings of SRNs.

5.4.1 Generalisation to unseen patterns

For one thing, SRNs generalise quite poorly to patterns they have never encountered during training: the patterns they learn are strongly tied to the token words found in these patterns. The most problematic cases involve generalisations which allow a word encountered in one syntactic position to be used or understood in a new syntactic position. Imagine a language learner encounters a new word describing an animal, in sentences where this word always happens to occur in subject position. The learner should be able to understand (or generate) sentences where this word occurs in object position, even though she has never been encountered in that position. Syntactic formalisms which can generalise over syntactic positions in this way are said to possess ‘strong systematicity’ (Hadley, 1994). A standard SRN exhibits little systematicity of this kind.

In some cases, this lack of strong systematicity is a good thing. The constructivist conception of syntax mentioned in Section 5.1 positively requires that syntactic constructions can be defined with reference to specific words and word combinations. But there are clearly also cases where properly abstract rules are required. So there are two questions: firstly, how can we create networks which exhibit strong systematicity, and secondly, how can we tell which constructions should retain reference to specific words and which should abstract away from them?

Most work has focussed on the first of these questions. There are now several network models which exhibit a good measure of strong systematicity. One of these is the network of Chang (2002; see also Chang et al., 2006). This network has an SRN at its core; the main innovation is that this SRN learns to sequence abstract semantic roles (AGENT, PATIENT, etc) rather than actual words. This network uses the ‘binding-by-connection’ scheme introduced in Section 4.4: a message is encoded by temporarily associating word meanings with particular roles. The words in an incoming sentence activate a sequence of role representations, and the SRN learns about the structure of these sequences. Chang et al’s model has quite wide syntactic coverage: it can learn the right order for semantic roles for a wide variety of episode types. And because the rules it learns are framed in terms of semantic roles, it shows good strong systematicity. On the other hand, by generalising away from individual words, the network loses an ability to encode idiomatic patterns.
which refer to surface structures in language, which is one of the things a normal SRN does quite well.

Another model which achieves a form of strong systematicity is that of Takac et al. (2011; see also Knott, 2012). This model uses binding-by-serial-position (see again Section 4.4) to associate participants in an episode with semantic roles. The novel element in this network is that semantic roles can feature more than once in the canonical sequence representing an episode: for instance, the sequence of roles representing a transitive sentence could be summarised as agent, patient, action, agent, patient. The doubling of agent and patient roles reflects the fact that these participants are apprehended in different sensorimotor modalities when a transitive episode is observed. They must first be attended to as objects, to establish the context in which a motor action can be executed or monitored; during the course of this execution or monitoring, they are reactivated, but this time in a motor modality. In Takac et al.’s model, producing a sentence which expresses an episode involves internally replaying the sensorimotor sequence characterising the episode in a special mode where sensorimotor signals can have linguistic side-effects. When replaying the episode, there are multiple opportunities to pronounce the agent and the patient; what the syntactic network has to learn is which opportunity to take. Like Chang’s network, Takac et al.’s network learns abstract word-ordering conventions, and therefore exhibits strong systematicity. However, in Takac et al.’s model, a specific sequence of roles is provided in advance, and the network only needs to learn binary choices about when to pronounce the agent and the patient, while in Chang’s network, a sequence of roles needs to be learned from scratch. Takac et al.’s network is in fact explicitly framed as a nativist model in the Chomskyan tradition: knowledge of the sensorimotor sequences associated with episodes is construed as innate linguistic knowledge, and acquiring abstract word-ordering conventions just involves setting the values of a small number of parameters (see Knott, 2012 for more about this interpretation of Chomskyan syntax). At the same time, Takac et al.’s network also provides a mechanism for learning surface structures in language. The network includes a simple SRN, which receives a semantic message as input and is trained to predict the next word in the sentence as output. Since messages have sequential structure, the SRN receives the message one element at a time rather than all at once. As long as the SRN can confidently predict the next word from the current semantic input, it can work autonomously, generating an ‘idiomatic’ sequence of several words; when it can no longer do so, the next element in the message sequence is provided. This scheme allows the network to learn a range of idiomatic and semi-idiomatic structures, as well as abstract word-ordering conventions.

A final interesting approach to systematicity comes from Frank et al. (2009). The key suggestion in this model is that ‘episodes’ are not a well-defined unit of representation in their own right, but should rather be thought of as components in larger representations of ‘situations’, which combine information about many episodes and their patterns of co-occurrence. In Frank et al.’s model, semantic representations are stored in a structure similar to a self-organising map, which is trained on symbolic representations of whole situations, which somewhat resemble the ‘discourse context’ representations discussed in Section 4.3. Symbolic situations are defined in relation to a microworld, in which a finite
set of \( n \) episodes can occur. Each symbolic situation is represented in an \( n \)-dimensional vector whose components specify the truth value (1 or 0) of every one of these episodes. These vectors are presented as training data to the self-organising network (which has a larger number of nodes). The network learns patterns of co-occurrence between the input episodes. Within the trained network, each episode is represented by a distinct vector, and Boolean combinations of episodes can be neatly derived from these basic vectors. Frank et al.’s sentence interpretation network is the simplest possible kind of SRN: it takes word sequences as input and is trained to generate appropriate situation representations as output. Nonetheless, it shows some degree of systematicity, being able to make good guesses about the semantics of unseen sentences, even when these feature unseen combinations of concepts. Interestingly, this network makes no attempt to represent the internal compositional structure of episodes. The previously described networks, by contrast, achieve systematicity precisely by representing this structure. One interesting possibility is that speakers make use of both kinds of episode representation, with each producing its own type of linguistic systematicity.

5.4.2 Representing nested phrase structure

Another problem with using an SRN in a model of syntax is that SRNs are not good at representing long-distance syntactic dependencies, of the kind which are seen in deeply nested relative clauses (e.g. the agreement between \textit{man} and \textit{likes} in \textit{The man [who my friend at Cadbury’s plays snooker with] likes grappa}). These kind of dependencies are what motivate linguists to think of syntactic representations as hierarchical. SRNs can learn something about phrases, as discussed in Section 5.3, but they are not explicitly designed to represent hierarchical phrase structures.

Encoding long-distance dependencies between words in an SRN is not impossible, but it requires reference to subtle features of context representations, reflecting words which occurred many iterations ago. Since information about the past fades exponentially in the context layer, these features are hard for learning algorithms to identify and exploit. Humans and SRNs in fact have difficulty with the same types of long-distance dependency: the hardest dependencies in each case are nested ‘centre-embedded’ clauses (\textit{The man [who the woman [who danced] chased] sang}). So in some ways this deficit speaks in favour of SRNs as a cognitive model of syntax (Christiansen, 1992). But simple SRNs also have difficulty with long-distance dependencies which pose no trouble to human speakers. Most modern SRN-based models include additional circuitry, for instance adding additional recurrent loops (Chang, 2002) or using explicit representations of the recent words in a sentence (Mayberry and Miikkulainen, 2008). But some networks go further, building in explicit representations of hierarchical, nested structures.

A key component in many of these latter networks is a \textbf{recursive auto-associative memory (RAAM)} network (Pollack, 1990). This is a network which takes a number of distinct vectors of size \( n \) as input, and learns to reproduce these inputs on a similarly structured output layer, through a hidden layer whose size is the same as the input vectors (see Figure 6). After training, the hidden layer serves to compress the combined input
vectors into a single vector which can be presented again to the network and combined with an additional vector. This property makes the RAAM a recursive data structure.\footnote{Note that the vector-symbolic representations described in Section 4.4, which achieve a similar effect without learning. But most work on nested syntactic structures to date has been conducted with variants on RAAMs. Most of the networks I describe below could probably be redesigned to use vector-symbolic representations.} Of course, compressing input vectors is a lossy operation, so there is a limit to the amount of recursive structure a RAAM can hold.

One interesting application of a RAAM is in a neural network implementation of a \textbf{stack}. A stack is a data structure which holds a number of items: an item is added with a ‘push’ operation, and retrieved with a ‘pop’ operation. The ‘pop’ operation returns the item most recently pushed onto the stack. To implement a stack, a RAAM network needs two inputs and outputs, as shown in Figure 6. The current state of the stack is represented by the RAAM’s hidden layer: to represent an empty stack, this layer is initialised with an arbitrary conventional vector. Items to store in the stack are represented as input vectors, with the same dimensionality as the hidden layer. When a new item is pushed onto the stack, the RAAM is given this item and its own current state as inputs (and training outputs), and learns a new hidden layer which represents the new state of the stack; this process can then be repeated to push additional items onto the stack. When an item is popped from the stack, we can simply evoke the stack’s current state in the hidden layer and read the popped item and the new state of the stack from the two vectors of the output layer. Of course, the limited capacity of a RAAM means that it can only store a certain number of items before it starts to make mistakes when popping items. But its performance degrades gradually: if it does not return the correct item, it is likely to return a similar one.

Early models of syntactic recursion used a RAAM to encode explicit representations of phrase structure. For instance, Reilly (1992) trained an RAAM to encode a number of different hierarchical parse trees in its hidden layer. Trees were strictly binary-branching, so the RAAM just needed two vectors on its input and output; it recursively learned compressed representations of each binary node in each tree, starting from the simplest nodes and progressing to the more complex ones. In a separate learning phase, a standard SRN was trained to map sequences of words onto the parse tree representations learned by the RAAM. Subsequent models (e.g. Berg, 1992; Ho and Chan, 1997) combined the...
RAAM and the SRN into a single network, but preserved the idea that parse trees are explicitly computed by a sentence processor.

An interesting alternative strategy is proposed in Miikkulainen’s (1996) SPEC network. In this model, an SRN receives the words of a sentence one by one, but works in clause-sized units, mapping each clause in a complex sentence onto an assembly representing role-filler bindings for that particular clause. For instance, when interpreting *The girl [who liked the dog] saw the boy*, it creates a representation of ‘The girl liked the dog’ (AGENT=GIRL, ACT=LIKE, PATIENT=DOG) when processing the the nested clause and a representation of ‘The girl saw the boy’ (AGENT=GIRL, ACT=SAW, PATIENT=BOY) when processing the matrix clause. Thus the output of the system for a complex sentence is a sequence of clause-sized semantic representations, rather than a single monolithic representation. The SRN itself never has to deal with the nested structure of clauses: this job is performed by two additional networks, which together act to store and load the SRN’s context representations at clause boundaries. One, called the segmenter network, learns to recognise contexts where nested clauses begin and end. It takes the SRN’s context and ‘current word’ representations as input, and learns to generate a control signal specifying what should happen at the next iteration of the SRN, as well as a signal which can be used as the SRN’s context representation at this next iteration. At the start of a nested clause, the control signal indicates that the SRN’s current context representation should be pushed onto a stack, and replaced with a new context representation appropriate for processing the nested clause. At the end of a nested clause, the control signal indicates that the SRN’s context representation should be replaced with the representation obtained by popping the stack. Where there is no clause boundary, it indicates that the context representation at the next iteration should be what the SRN would normally use (i.e. its current hidden layer). The stack is implemented by a RAAM network, as described above, configured to store vectors with the same dimensionality as the SRN’s hidden layer. When a new context representation is pushed onto the stack, the RAAM learns to autoassociate this context representation with its own current hidden layer, creating a new hidden layer representing the new state of the stack. When the stack is popped to restore a previous context, this context (and the new state of the stack) are read from the two vectors of the RAAM’s output layer. Dividing the task of sentence interpretation up between these three networks allows the model to learn good generalisations: the SRN learns generalisations which extend over all clauses, whether they are main clauses or nested ones, and the segmenter network learns generalisations about where clauses begin and end, regardless of their content or how deeply nested they are. The SRN and segmenter networks can learn what they need to know from a relatively small set of training sentences, at which point the SPEC network as a whole can successfully interpret a large space of complex sentences far greater than the size of its training set. At the same time, the RAAM’s limited capacity results in psychologically realistic difficulties interpreting multiply centre-embedded clauses.
6 Summary

Neural network models provide many insights into the neural circuitry implementing the human capacity for language. We have interesting models of phonological representations, word forms, and word meanings, and of the circuits which learn connections between word forms and word meanings. We also have interesting models of the representation of sentence-sized episodes, both in long-term memory and in working memory, and of the representation of the discourse contexts in which sentences are interpreted. Finally, we have interesting models of the syntactic machinery which implements the mapping between sequences of words and episode representations. There is a trend away from ‘toy models’ towards models with wide syntactic coverage, echoing a similar trend in applied computational linguistics, where there is now a similar emphasis on machine learning techniques and naturalistic data sets. There is still real debate, however, in almost all the areas which have been discussed, especially in the areas of syntax and sentence/discourse semantics. I have described several radically different models of episode representations, and several radically different models of syntactic processing; all of them have their limitations, so the correct model may not bear much resemblance to any of the ones I have reviewed here. Caveat lector!

References


