2 Capturing Gradience, Continuous Change, and Quasi-Regularity in Sound, Word, Phrase, and Meaning

JAMES L. MCCLELLAND

1. Visions of Language

One vision of the nature of language holds that a language consists of a set of symbolic unit types, and a set of units of each type, together with a set of grammatical principles that constrain how these units can be used to compose other units, and a system of rules that project structured arrangements of such units onto other structured arrangements of units (for example, from syntactic to semantic structure). An alternative vision of the nature of language holds that it is often useful to characterize language *as if* the above statements were true, but only as a way of approximately notating or summarizing aspects of language. In reality, according to this alternative vision, approximate conformity to structured systems of symbolic units and rules arises historically, developmentally, and in the moment, from the processes that operate as users communicate with each other using sound or gesture as their medium of communication. These acts of communication leave residues that can be thought of as storing knowledge in the form of the continuous-valued parameters of a complex dynamical system (i.e. a system characterized by continuous, stochastic, and non-linear differential equations). Greatly influenced by the work of Joan Bybee (1985, 2001) and others who have pointed out some of its advantages, I am a disciple of this alternative vision (Bybee and McClelland, 2005; McClelland and Bybee, 2007).

As argued in the Bybee and McClelland papers just cited, neural network models that rely on distributed representations (sometimes called connectionist or parallel-distributed processing models) provide one useful way of capturing features of this vision. Such models are, in general, just the sort of continuous, stochastic, non-linear systems that are needed to capture the key phenomena, and the connection weights and other variables in such networks are the continuous-valued parameters in which

The Handbook of Language Emergence, First Edition. Edited by Brian MacWhinney and William O'Grady.

the relevant knowledge is stored. The present chapter reviews this vision and the use of distributed neural networks to capture it, covering motivations for the approach based on phenomena of language, some extant models using the approach, and prospects for the further development of this approach to understanding the emergence of language.

2. Motivations for an Emergentist Vision

2.1 Continuous variation and continuity of change in the units of language

Some of the basic elements of motivation for this alternative vision have been laid out in the papers cited above; here I review some of the key elements. First, a fundamental motivation for avoiding a fixed taxonomy of units is the existence of continuous variation in the characteristics of the purported building-blocks of language. Indeed, even the presence vs. absence of a purported linguistic unit can be a matter of degree. To mention but a few examples: (1) Attempts to identify a universal phonemic inventory founder in the face of graded differences in the realizations of phonemes both within and across languages. Even within a local dialectal community and in identical local phonetic context, phonemes vary continuously in a way that depends on frequency. For example, the /t/'s in *softly* and *swiftly* differ in the duration of silence and the amplitude of the burst: the former is shorter and the latter is smaller in *softly,* the more frequent of the two words. (2) Similar factors affect syllabic status. The word *livery* clearly has three syllables, and *every* generally only has two, but *memory* is intermediate, and greater reduction is associated with greater frequency. (3) Morphology – even the presence of an inflectional marker – can be a matter of degree. The regular English past-tense marking is more reduced in some words than others, and again frequency is a factor that affects this. Many frequent words that are past tense-like lack differentiation between their present- and past-tense forms (*hit* and *cut* being two examples). (4) In derivational morphology we see clear signs of variation in the extent to which a word should be treated as a single unit or as a composition of two or more subunits. Bybee (1985) illustrated the problem by considering a range of words beginning in *pre*. In some cases, such as *prefabricate*, it seems adequate to treat the item as consisting of two morphemes, while in others, this is less adequate: In cases like *predict* and (to a greater extent) *prefer*, *pre* loses its phonological identity (with the vowel becoming weaker and weaker), the remainder of the word has little or no independent status, but treating the item as a single atomic unit loses the characteristic of coming or being placed before that is still present. (5) One other domain in which a taxonomy of units seems particularly problematic is that of word meanings (McClelland, 1992). Fodor and Pylyshyn (1988) claimed that the verb *love* contributes the same thing to the meaning of *John loves Mary* and *Mary loves John*. However, the meaning of *love* changes in *John loves ice cream*, *The pope loves sinners*, and *Jimmy Swaggart loves prostitutes*. It could be argued that *love* has many different meanings, each of which can be listed separately in the lexicon, but where do we draw the line? I would argue that even in the case of *John loves Mary* and *Mary loves John*, the meaning of *love* is slightly different, and that, in general, the meanings of words are not selected from a fixed taxonomy of alternatives, but take on different shades in different contexts that cannot be captured by a fixed taxonomy.

In summary, the constituents of linguistic expressions appear to exhibit continuous variation that makes any fixed taxonomy of types problematic. Very importantly, there is a tendency for fragments of utterances to become more and more compressed and less and less analyzable as languages evolve over time (Bybee, 2006). One seeks a modeling framework that avoids any pre-commitment to any particular taxonomy of types, allows the presence of constituent elements within larger items to be a matter of degree, and also allows for a completely gradual and continuous change in the extent of presence and the detailed characteristics of these constituents to the point of their disappearance or merger with other constituents.

2.2 *Quasi-regularity and sub-regularity*

A further set of issues arises when one attempts to characterize lawful relationships with a system of rules. The problem starts with the fact that linguistic systems (as well as other structured bodies of knowledge) exhibit both regular items and exceptions. One can attempt to address this while maintaining a relatively pure and abstract system of rules by treating the exceptions as items that must simply be listed explicitly as such, but simple forms of this idea miss two pervasive characteristics of exceptions: the fact that they often share in the regular patterns at least to a degree and the fact that they tend to come in clusters. Seidenberg and McClelland (1989) introduced the term *quasi-regularity* to refer to these characteristics. For present purposes, I will use the term *quasi-regularity* to refer to the tendency for forms to exhibit partial consistency with the so-called regular patterns typical of other forms and/or with so-called regular mappings typical of other items; a *quasi-regular* item will be one that exhibits such partial consistency. I will use the term *sub-regularity* to refer to the tendency for irregular forms to exist in clusters with similar characteristics: a *sub-regular* item will be an item that participates in one of these structures. I begin with two simple example domains that illustrate the concept, one from the English past tense and one from the English spelling–sound system. As we shall see, in both cases, quasi-regularity and sub-regularity often co-occur with each other.

2.2.1 The English past tense The English past tense is characterized by a fairly pervasive regularity: we form the past tense of a verb by adding /d/, /t/, or /ɪd/ depending only on simple phonological features of the final segment of the stem. However, the past tense of the word *say* does not rhyme with *played* as it would if it were fully regular: Instead, the past tense of *say* is *said*. This is a quasi-regular item in that the past tense preserves most of the phonological properties of the stem, and, like other words ending in a vowel, adds the voiced stop, /d/. The item would be fully regular were it not for a reduction of the vowel. An example of a simultaneously quasi-regular and sub-regular pattern is the pattern exhibited by *keep* and many other verbs ending in *-eep* (including *creep*, *weep*, and *sleep*, but not *beep*). Here, the unvoiced stop /t/ is added after the final unvoiced consonant of the stem as it would be in fully regular items, but the items are exceptions to this pattern in that the vowel is reduced. Similar points apply to a set of verbs that rhyme with *feel*, though here what would regularly be a /d/ becomes a /t/, as in *feel-felt*, *deal-dealt*, *kneel-knelt*, etc. (McClelland and Patterson, 2002b). The English past tense also includes sub-regular patterns that do not add a /d/ or /t/ to a past-tense form, as in clusters of items like *sing-sang*, *ring-rang*, etc.

2.2.2 Mapping from spelling to sound Although the mapping from spelling to sound in English is known to be rife with exceptions, nearly every exceptional form is quasi-regular, and quasi-regularity generally co-exists with sub-regularity. The case of the word PINT is typical: its pronunciation is not regular /pɪnt/ where /ɪ/ represents the vowel in HINT, MINT, LINT, but /paːint/, where /aːi/ represents the vowel in WINE, PIKE, TIDE, etc. Two things are critical here. The first is that the phonemes corresponding to the letters P, N, and T are completely consistent with the most typical case, so that PINT could be said to be at least three-quarters regular. The second is that the exceptional pronunciation of the letter I is not completely inconsistent with its use in other cases. Not only is this the typical pronunciation of I in the context of a following consonant and a final E as in the examples above, but it also arises in cases like BIND, MIND, FIND, and KIND, which share orthographic and phonological features with PINT. Again, there is nothing atypical about these characteristics; quasi-regularity and sub-regularity are pervasive characteristics of the spelling–sound system of English. In summary, in these two domains we find that nearly all exceptional items are largely consistent with the regular pattern found in other items and/or that their idiosyncratic properties are shared with other items. Such sharing with other items is especially likely for items that are themselves of low frequency.

The presence of quasi-regularity as well as sub-regularity in the English past tense challenges the approach of characterizing language knowledge as a system of rules since it requires decisions to be made about (1) when a rule should be invoked and (2) whether a rule applies to an item or not. Attempts have been made to address these issues, and I do not wish to suggest that systems with these characteristics could not be made to work in particular cases. The phenomena do, however, strongly blur the line between the productive and the non-productive elements of language, and have motivated many to search for explanatory frameworks in which a single homogeneous mechanistic framework deals simultaneously with regular and exceptional items. Before turning to a consideration of such models, we briefly consider three other domains in which similar issues arise.

2.2.3 Natural kinds While this domain might be excluded from language by some, for those who see language as exemplifying domain-general principles, not to mention reflecting the structure of the natural world, this is an important domain to consider alongside of more properly linguistic domains. It might even be argued that cognitive mechanisms that evolved in pre-linguistic hominids evolved to be useful for capturing the quasi-regular structure of the natural world. This domain clearly exhibits quasi-regularity, in the sense that many items are partially but not totally consistent with the typical features of their taxonomic category. *Elephants* and *turkeys* are good examples. *Elephants* have many of the typical properties of mammals, so it is clearly useful to see them as members of this class, but they also have several idiosyncratic properties. Their large floppy ears and trunks are unique, while they share having tusks with a few other animals. *Turkeys* share many properties of birds, but are members of a sub-regular cluster of flightless birds (though flightlessness itself a matter of degree – wild turkeys can get off the ground for short distances), and they tend to share with such birds their superior edibility compared to many birds that fly. Clearly, then, the domain of natural kinds exhibits both quasi-regularity and sub-regularities.

2.2.3.1 Derivational morphology Returning to a topic within language, derivational morphology is also rife with quasi-regular and sub-regular patterns. Derived morphological forms include cases that appear to arise from a very productive process (e.g. the addition of *ness* to turn an adjective into a noun, as in *bold-boldness*) as well as cases that arise from less productive processes (e.g. *profound-profundity*; Aronoff, 1976). The less productive cases could be thought of as sub-regular patterns, but with the twist that the meanings of the participating derived forms tend to exhibit a degree of idiosyncrasy while also partially reflecting the semantic characteristics of the other items sharing the same affix. Bybee's (1985) examples *predict* and *prefer* both illustrate this: In both cases, there is a sense of priority (either in time or attractiveness), though the exact sense is not fully predictable by a simple rule or by a strict composition of the meanings of the parts. I see these cases as being yet another example of quasi-regularity, which is to say: we cannot account for the item's properties fully by treating it as part of a regular pattern or superordinate class, but we would be ignoring some degree of participation in a pattern shared by other items if we treated the item as though it were a completely unanalyzable word form separate from other forms with which it partially shares structure.

2.2.3.2 Meanings of multi-word patterns As a final example, I consider the quasi-regularity associated with the meanings of multi-word structures. These phenomena are generally discussed under the heading of *constructions* (Goldberg, 1995; Croft, 2001). Again we see a range of cases, from those that seem predictable enough from a rule-based compositional perspective to those that seem highly idiosyncratic. *She hit the ball* falls at one end, recognizably instantiating the canonical NP-(V-NP) pattern referenced in *Syntactic Structures* and triggering the mapping SVO→Actor-Action-Object proposed by Bever (1970), but what about *She hit the scene* or *She hit the wall*? The first of these is an instance of a relatively open construction (*X hit the Y*, where X is a person and Y is a social event or setting), whereas the second is far more restrictive at least with respect to the object constituent. In both cases, however, there is a degree of idiosyncrasy and context-specificity of the contribution of the verb (*hit*) to the overall meaning of the expression. Furthermore, there are additional cases such as *She cooked his goose* and *She kicked the bucket* where the meaning of the whole appears to be progressively more "opaque" and idiosyncratic. It may be useful to see the range of cases as divided into types with different labels (fully productive, constructions, collocations, and idioms, perhaps) – but at the same time it is important to see that they all admit to some degree of variation in such things as tense, aspect, and number, in accordance with standard patterns. In all cases, there is a degree of consistency with the regular patterns in language, with progressively increasing degrees of specificity and idiosyncrasy: treating different types of cases differently ignores the continuity among them. A goal for a theory of language would then be to offer a single homogeneous approach to address the full range of cases.

3. Modeling Graded Constituency, Continuous Change, and Quasi-Regularity

Having noted the graded nature and gradual changes in linguistic units and the quasi-regularity that characterizes all kinds of linguistic expressions, we consider

ways of approaching the development of models that might address these kinds of phenomenona.

3.1 Rules plus similarity-based generalization among exceptions

One approach is the rules-and-exceptions approach advocated by Pinker (1991, 1999) and subsequently by Jackendoff (2007). According to this approach, there are two types of items: those that are fully consistent with the regular patterns of language and those that are not. Similar ideas have been proposed by Coltheart, Curtis, Atkins, and Haller (1993) in the domain of reading. A problem for the simplest form of such views is that they offer no basis for understanding either the sub-regularity or the quasi-regularity that one finds in exceptions. Pinker (1991), recognizing the presence of sub-regular clusters in exceptions, proposed that the exception system exploits a similarity-based activation mechanism, similar to that offered by the connectionist model of past-tense formation that Rumelhart and I proposed (Rumelhart and McClelland, 1986). Items that are similar to other items could then enjoy support from such items, explaining the tendency for low-frequency exceptions only to be found in the present-day language if they are parts of a cluster of similar items, and even explaining the observation that occasionally, forms are attracted into such clusters (Pinker and Prince, 1988, cited *kneel-knelt* as a possible example of this kind, joining a cluster including *deal-dealt* and other items). However, Pinker and colleagues argued that such processes were characteristic only of the lexicon, and not the rule systems of language, which are fully categorical and "algebra-like" in nature.

While the rules-plus-similarity-based-generalization-among-exceptions view can address sub-regularities, it does not explain why so many irregular items have so much in common with the regular forms, and it made claims about dissociations between regular and exceptional forms that did not stand up to further scrutiny (McClelland and Patterson, 2002a; Seidenberg and Plaut, in press). In my view the fundamental problem facing this approach is to explain why so many exceptions are quasi-regular, if regulars and exceptions are produced by distinct processing mechanisms. The above review of the pervasiveness of quasi-regularity suggests that quasi-regularity is not an accident but is instead a fundamental characteristic of language and other natural forms of structured knowledge.

Exemplar models. Another framework that can capture many of the phenomena is an exemplar model framework (Nosofsky, 1984; Pierrehumbert, 2001). The idea here is that items that are similar in, say, phonological form to a given input will all be partially activated when the form is experienced. Semantic features of these items will then contribute to the representation of meaning. In this way phonological forms that are similar to other past tenses will seem to convey pastness, even if they lack a past-tense morpheme. In exemplar models, highly similar forms generally carry greater weight than those that are less similar, thereby providing a mechanism for the partial override of general patterns by a cluster of similar examples that have similar features. Such models can address change over time and with experience if they include a further process whereby items that are predictable and/or occur frequently will be subject to a compressive shortening which can then rob the item of its similarity to other forms, allowing it to become

more independent of these in meaning (Bybee, 2001; Pierrehumbert, 2001). Reciprocally, similarity in meaning can help preserve similarity of form, and this too can perhaps be captured in exemplar models. The idea would be that semantic similarity of a given item *x* to a collection of other known items would cause aspects of the known items' form to become active, helping to protect the form of item *x* from changing as much as it would in the absence of such similarity of meaning. Such situations arise, for example, in inflectional morphology, where the meaning of the inflection (e.g., tense or number) is largely independent of the meaning of the item inflected. The consistency of meaning in these cases, as Bybee (1985) argued, helps explain the consistency of form, and exemplar models can help explain this.

Even though I have often relied on exemplar models myself (McClelland, 1981; McClelland and Rumelhart, 1981; Kumaran and McClelland, 2012), I see these models as another form of sometimes useful approximate characterizations of what are underlyingly distributed neural networks. Going to the distributed network level allows us to address two problems facing such models. The first is the problem of specifying whether exemplars should be represented at the type or the token level. If we have one exemplar for each alternative type, we must then confront the problem of deciding when an item is just another example of an existing type, and when a new type representation should be created (Plaut and McClelland, 2010). That is, in the face of the considerable variability among tokens of the same item, how can we know which ones to combine in a single type representation and which ones are actually tokens of different types? The alternative of assuming complete storage of full detail of each encountered token of each type may be a way to avoid this issue, but it creates a new problem, namely that every experience must be stored, severely taxing memory capacity. The second problem is that of specifying a similarity metric for exemplar models. This arises in assigning tokens to types in models that represent exemplars at the type level, and in deciding on the contribution of each stored exemplar during processing of a current input in both types of exemplar models. In my view, it is unlikely that a fixed, universal similarity metric exists; rather, similarity is a matter of language- and culture-specific convention and so the similarity metric must arise in part from experience-dependent processes. While there are exemplar models of categorization that provide a rudimentary form of adjustment of the similarity metric by allowing differential weighting of pre-specified dimensions (Nosofsky, 1984; Kruschke, 1992), neither model allows the construction of the actual dimensions of similarity themselves, something that is possible with learned distributed representations, as we shall discuss below.

4. Distributed Neural Network Models

4.1 Earliest efforts

The type/token issues facing exemplar models were among those that led me and Rumelhart to our early explorations of models that used distributed representations. We explored this idea in a distributed model of memory (McClelland and Rumelhart, 1985), and in our model of past-tense inflection (Rumelhart and McClelland, 1986). Such models do preserve a shred of the key idea in exemplar models – each experience leaves a residue in the system – but unlike exemplar models, the residue left behind is not

construed to be a distinct memory trace requiring separate storage. Instead, the residue is the set of adjustments that the experience makes to the connection weights among the processing units in the system. The adjustments made by different experiences are all superimposed in the ensemble of connection weights, so that experiences can cumulate without requiring the allocation of additional storage for each new experience, and each experience can have an effect on processing without requiring it to be stored separately. Items in memory (objects and their names in the memory model) and examples of present- and past-tense forms (in the past-tense inflection model) are not stored as such: all that is stored is the superimposed, cumulated result of the set of example-by-example changes that have been made to the connections.

Importantly for the issues under consideration here, both of these early models showed how one and the same ensemble of connection weights could simultaneously exhibit sensitivity to typical or regular patterns while also capturing idiosyncratic properties of individual items. In both cases exceptional items were generally quasi-regular, in that they shared some properties with other examples. For example, in the distributed memory model, McClelland and Rumelhart (1985) considered an exceptional dog that had some idiosyncratic properties as well as some properties it shared with other dogs, and to a degree with cats also seen by the model. In the past-tense model, Rumelhart and McClelland (1986) examined the model's performance with fully regular past-tense items (*shape-shaped*), arbitrary one-off exceptions (*go-went*), quasi-regular items (*say-said*), including those occurring in clusters (*keep-kept)*, and items occurring in other types of sub-regular clusters (*sing-sang*, etc.). In both cases the models used a simple, homogeneous, learning procedure and a single integrated network architecture to simultaneously deal with all of these different kinds of items. In particular, the same connections that were used to inflect regular *shape* to form its past tense *shaped* contributed in inflecting quasi- and sub-regular *keep* and *sleep* to their past-tense forms *kept and slept*; and the connections that allowed the network to capture the reduction in the vowels were shared, so that the similar items contributed to the knowledge each used in the formation of its past tense. Not only did the model capture all of these types of known forms; it also exhibited a tendency to capture the productivity of both the regular and the irregular past tense, producing regular inflections for most of the novel items it encountered as well as extending quasi- and sub-regular patterns to previously unseen examples (*weep-wept*, *cling-clung*).

4.2 Learning in distributed neural network models

Distributed neural network models generally make use of what is often called an "error-correcting learning algorithm." A good way to view these algorithms is to see them as imposing a constraint on the values of connection weights based on the characteristics of the full ensemble of patterns used to train them. The models already reviewed used the two basic paradigms that are used in many neural network learning models: *pattern-association* and *auto-association*. In pattern association, used in the past-tense model, one pattern is associated with another: in this case the pattern for the present tense of a word is associated with the pattern for the word's past tense. In auto-association, used in the distributed memory model, a pattern is essentially associated with itself. The two ideas can blur into each other, when we consider that two

patterns can often be considered to be parts of a single larger pattern or as sub-patterns to be self- and inter-associated.

The models just reviewed differed from almost all of their successors in using a single layer of modifiable connection weights, thereby limiting their learning capabilities. Just after this work was completed, it became possible to train multiple layers of connection weights, using the back-propagation learning algorithm (Rumelhart, Hinton, and Williams, 1986), which extends the error-correcting learning idea to networks with hidden units – units whose activation values are not specified directly by the inputs or target patterns presented to the network. Such models have the potential to learn both how to represent their inputs as patterns of activation across their hidden units and how to use these representations, and so have the potential to address how learning and experience can affect the representations used for given inputs, and to address how representations change dynamically over developmental and historical time.

The remainder of this chapter considers such distributed neural network models further. I argue that, in spite of the trenchant criticisms of early versions of such models, they have much to offer – certainly, as one among several approaches – in helping us capture the gradient nature of linguistic structures and processes, the gradual nature of change, and the presence of quasi-regular and sub-regular structure among items that other approaches often exclude from the core mechanisms of language as exceptions. These models are useful, I believe, because they have the potential to allow us to address the problem of understanding how languages map between meaning and sound without pre-specification of a taxonomy of units and unit types and without relying on an artificial division between regular and exceptional items that prevents the quasi-regularity in the exceptions from being captured in the regular system. I will proceed by (1) briefly noting several of the bodies of modeling work that have attempted to address each of the domains discussed above, (2) examining some of the challenges that have confronted these models, and (3) describing exciting new directions in the exploration of such models that indicate that some of the limitations have been or may soon be overcome.

4.3 Distributed neural network models applied to language, reading and semantic representation

4.3.1 The English past tense The distributed neural network model of the English past tense introduced by Rumelhart and McClelland had the positive features noted above, but led to a barrage of criticisms addressing limitations of the model and calling its core tenets into question (Pinker and Prince, 1988). One criticism fell on the choice of input representation, said by Lachter and Bever (1988) to presuppose the solution to the problem, but said by Pinker and Prince to be woefully insufficient to capture aspects of linguistic regularities. Another fell on the fact that the model was only partially successful in applying the regular pattern of the English to novel forms; and a third fell on the model's unrealistic characterization of the training experiences that allowed it to capture U-shaped over-regularization of exceptions (Marcus, Pinker, Ullman et al., 1992). All three of these criticisms were addressed by subsequent simulations by others. MacWhinney and Leinbach (1991) showed how, with a different choice of input representation, a distributed neural network model could easily master the pervasive regular

pattern. Plunkett and Marchman (1991) chose to focus on the U-shaped pattern of over-regularization, showing that this pattern, as it is exhibited in the corpora of Adam, Eve, and Sarah (Brown, 1973), can arise with much more realistic assumptions about the training experiences of young children. Other work, by Daugherty, MacDonald, Petersen, and Seidenberg (1992) and Hoeffner and McClelland (1993), extended the model to address the important role of semantic as well as phonological influences on past-tense inflections. The work of Plunkett and Marchman (1993) was very important in stressing how a distributed neural network model would naturally capture the tendency for exceptions to occur in clusters.

In all of these models, the knowledge that underlies the correct production of the regular past tense is at work in the network whenever an item is presented. To take the MacWhinney and Leinbach model as an example, this knowledge would largely have been confined to connections from the input units representing post-vocalic segments of the final syllable of the uninflected form of a word and connections to output units for a possible post-stem inflection (/d/, /t/ or /ɪd/). These connections would have been in play whenever a regular item such as *play* or *beep* was presented or a exceptional item such as *say*, *keep*, or *feel* was presented, and so they would participate in the production of the regular aspects of the past tenses of such forms. The network would learn to capture idiosyncratic aspects of particular exceptions by using connections arising from throughout the input to adjust the output in item-specific ways. The ability to do this was, quite naturally, a joint result of the extent of the modification required (thereby favoring modest stem-to-past alterations), the frequency of the item itself, and the combined frequency of other similar items involving similar modifications.

4.3.2 Other aspects of morphology and sound–meaning relationships Although the past tense of English has been subject to the most intense scrutiny, aspects of derivational morphology have also been considered using neural network modeling approaches. At issue here is the graded semantic compositionality of many kinds of inflectional forms, including those signaled by phonological and sometimes orthographic changes, and those not signaled by such changes (e.g., *rewrite* vs. *return*, Gonnerman, Seidenberg, and Andersen, 2007). Distributed neural network models have been used to capture graded priming effects observed with such items (Plaut and Gonnerman, 2000). It is also worth noting the usefulness of distributed neural network models to capture the graded constraints that shape the phonological patterns associated with grammatical gender (MacWhinney, Leinbach, Taraban, and MacDonald, 1989), including subtle influences of the partial association of grammatical gender with biological gender (Dilkina, McClelland, and Boroditsky, 2007). The German *–s* plural, treated by Marcus, Brinkmann, Clahsen, Wiese, and Pinker (1995) as an example *par excellence* of a case of an algebra-like rule of language, has not yet been modeled using a distributed neural network approach, but it is worth noting that it exhibits sensitivity to complex phonological and semantic influences as generally expected under the present perspective (see McClelland and Patterson, 2002a, for a review of the relevant findings). This and many other aspects of inflectional systems found in the world's languages are ripe for future modeling within a distributed neural network framework.

4.3.3 Spelling-to-sound models The initial effort in this domain was undertaken by Sejnowski and Rosenberg (1987), using a simple distributed neural network model in

which each letter in a text was moved sequentially across the inputs to a multi-layer network. For each letter, the network was trained to produce the corresponding phoneme or (in the case of silent letters or letters after the first in multi-letter graphemes such as SH) no phoneme in its output. The network successfully learned to translate text into the appropriate sequence of outputs as specified in its training corpus and exhibited suggestive developmental transitions but was not systematically applied to reading data.

Seidenberg and McClelland (1989) used input and output representations similar to those used in the Rumelhart and McClelland (1986) past tense model to begin to address developmental and adult patterns in reading words aloud. This model captured a considerable body of word reading data but like the Rumelhart and McClelland past tense model it did not adequately capture the human ability to read non-words, leading critics to argue for the importance of maintaining a separation of systems for processing exceptions on the one hand and novel items consistent with rules of spelling–sound correspondence on the other. However, subsequent models by Plaut, McClelland, Seidenberg, and Patterson (1996) used an improved input representation, and successfully demonstrated that a simple, three-layer distributed neural network model with an appropriate choice of input and output representations could adequately address the same body of word reading data addressed by the earlier model, and could also achieve human-like levels of success in reading non-words. Plaut et al.'s analysis of the model centered on the way in which its reading of exceptional items such as PINT and BOOK simultaneously exploited the same connections underlying the reading of other items with which each item overlapped. All items beginning with, say, B would naturally exploit the same connection weights (from the input unit for orthographic onset B to the hidden layer, and from the hidden layer to the output unit for phonological onset /b/). Input units for vowel graphemes such as I or OO tended to activate all possible correspondences of these when presented in isolation. When surrounded by other letters, such as final K in the case of the word BOOK, these activations would be shifted to favor the short-vowel reading of OO typical of short-vowel contexts; but onset SP as in SPOOK would override this and shift the activation back toward the long-vowel correspondence found in this item. In general, all context letters were necessary for the model to read an exception word correctly; with less context it generally tended to activate the most probable correspondence for the given fragment. In reading non-words such as GROOK the model distributed its responses among what Patterson, Ralph, Jefferies et al. (2006) have called the *alternative legitimate renderings* of OO, in this case the vowel in BOOK and the vowel in SPOOK and TOOL, just as human participants do. When subjected to damage, frequent and regular items tended to be preserved much more than less frequent and less regular items, as observed in patients with brain damage producing reading disorders (for details, see Plaut et al., 1996).

4.3.4 Models of natural kind semantics The characteristics of the distributed neural network models described above were very much in mind as Rogers and I began to consider the interesting patterns of behavior exhibited by neuropsychological patients undergoing progressive degeneration of the anterior temporal lobes, producing the condition known as semantic dementia. Such patients exhibited a striking pattern of errors as their disease progressed, revealing strong sensitivity to typicality and frequency (see McClelland, Rogers, Patterson, Dilkina, and Lambon Ralph 2009 for a review). Perhaps the most

striking finding is the tendency of such patients to exhibit over-regularization errors in past tense inflection, in spelling–sound correspondence, and in the generation of properties of objects (Patterson et al., 2006). When reading PINT, the patient might produce the regular form /pɪnt/; when inflecting *sing*, the patient might say *singed*; and when drawing a picture of a duck, the patient might add two extra legs, consistent with other animals often seen walking about on the ground. Correspondingly Rogers and I were struck by the existence of parallel phenomena in semantic development, whereby young children attribute to objects properties that they do not have in accordance with typical properties of superordinate categories (a phenomenon some had termed "illusory correlations"), or in which they overgeneralize names of frequently occurring objects. Using the distributed neural network of semantics introduced by Rumelhart (1990; Rumelhart and Todd, 1993), Rogers and I simulated the semantic findings described above as well as many other aspects of semantic and conceptual development and the disintegration of semantic knowledge in semantic dementia (Rogers and McClelland, 2004; Rogers, Lambon Ralph, Garrard et al., 2004), and there are now models that simultaneously capture aspects of both the spelling-to-sound and the semantic errors seen in such patients (Dilkina, McClelland, and Plaut, 2008). These models, like the ones described above, all use relatively generic neural network architectures involving input and output units for each of several different types of information about an item (for example, the semantic, visual characteristics of an item and the orthographic and phonological characteristics of the word of the item). As before, the knowledge responsible for generating the typical aspects of an item (be they orthographic, phonological, visual, or semantic aspects) is shared across many items and is more robustly represented because of this sharing, accounting for its tendency to override less pervasive and idiosyncratic information both in development and in degeneration.

4.3.5 Distributed neural network models of sentence processing Shortly after the initial wave of distributed neural network modeling work on past tense and spelling to sound, interest arose in applying similar ideas in the domain of sentence processing. Miikkulainen and Dyer (1991) and Pollack (1990) were among those exploring this issue from a computer science perspective. While the modeling work here tended to address issues other than gradedness and quasi-regularity *per se*, the models nevertheless shared the characteristics of the above models in that they sought to avoid commitment to linguistic units of particular types or the explicit formulation of linguistic rules. Elman's (1990, 1991) use of simple recurrent networks exemplifies the approach. These papers showed that a very simple distributed neural network could learn to make appropriate predictions consistent with various types of explicit linguistic representations and rules: that is, the networks acquired sensitivity to key features of the sequential dependency structure of English. From a training corpus consisting only of a steady stream of words generated according to a generative grammar, the network learned to predict each upcoming word by using preceding words. With training, it came to be able to predict successor words of the appropriate syntactic category, and, within these, to restrict its predictions to items that obeyed selectional restrictions embodied in the generative grammar. No negative evidence was needed: the network learned simply from the stream of words that formed grammar-consistent sentences. In the 1991 paper, the grammar included embeddings that required the neural network to learn grammar-appropriate sensitivity to long-distance dependencies. This occurred without

the network having any prior knowledge of syntactic categories or of the characteristics of the grammar that generated the training examples.

In a parallel effort undertaken at about the same time, St. John and McClelland (1990; McClelland, St. John, and Taraban, 1989) undertook to address the problem of mapping from strings of words to meanings. This work was an example of a radical eliminative or emergentist approach in that it completely avoided making any commitments to explicit representation of syntactic structure as an intermediary between a string of words on the one hand and the meaning of the sentence on the other. Instead, the model learned simply from pairings of strings of words representing stripped-down sentences (e.g. *The boy kissed the girl*, *The bus driver ate the steak with gusto*) and a simplified representation of the set of role-filler pairs in a simple frame-like representation of the sentence. Even though it lacked any explicit notion of syntactic structure the model could successfully learn to recover the appropriate meaning representation for both active and passive sentences. The model also correctly inferred implied arguments (e.g. the instrument in *The boy cut the steak*), correctly conformed to selectional restrictions on arguments embodied in the sentence–event pairs it was trained on (e.g., since all kissing in the model was between humans of opposite sexes, the model could anticipate that the object of the incomplete sentence *The boy kissed* … must be a human female). The knowledge of lexical meaning, syntactic convention, and selectional constraints among constituents was embedded homogeneously in the connection weights in the network and acquired as a result of exposure to examples of sentence–event pairs.

There has been a large body of other relevant work using distributed neural networks to address aspects of sentence processing and comprehension (Reali and Christiansen, 2005; Chang, Dell, and Bock, 2006; Bryant and Miikkulainen, 2001; Rohde, 2002). Some of this work has improved on the models described above by exploring the consequences of learning using naturalistic corpora and/or has addressed shortcomings of the earlier work, such as the restriction of event representations to a flat role-filler representation in St. John and McClelland (1990). This effort appears to have slowed in recent years, however, due in part to computational limitations. As we shall see below, some of these limitations have recently been overcome by research in machine learning.

4.3.6 Representations learned by the models As important as the successes of these models in capturing linguistic and semantic knowledge and human language-processing behavior was the analysis of the internal representations the models used in achieving the solutions they found, and the progressive changes in these representations over the course of learning. We focus first on the findings from Elman's 1990 model, trained strictly on word sequences forming sentences generated by a simple generative grammar. The key point is that the representations found capture key syntactic categories and subcategorizations identified by linguists without having these categories pre-specified for them. That is, the models assigned to words distributed internal representations such that (1) all nouns were more similar to each other than to verbs and *vice versa*, (2) within nouns, animates were distinguished from inanimates, and (3) within verbs, intransitive, transitive, and ditransitive subtypes were all distinguished. Importantly also, these representations were modulated by context, in ways that were systematic with respect to the selectional restrictions applying to a given word in a given context. For example, the patterns representing the nouns *boy* and *girl* would change similarly when these words occurred in subject vs. object position. Thus, the model captured general aspects

of grammar as well as structured context-specific variation. The issue of context-specific variation was further explored in Elman 1991, where it was found that the representation associated with the head noun of a main-clause noun (such as *man* in *the man who the boys chase walks dogs*) would be approximately the same both at the end of the simple noun phrase *the man* and at the end of the entire complex noun phrase *the man who the boys chase*, indicating that the model had learned structured expectations consistent with the structural constraints embodied in the training corpus.

The representations that emerged from learning in the model of word semantics by Rogers and McClelland (2004) had characteristics similar to those arising in Elman's model, but Rogers and McClelland explored both the developmental course of such representational changes and issues related to sub-regularities and quasi-regularity. Here the key findings were that: (1) the representations used in the model undergo progressive differentiation in the course of development, first capturing the gross, superordinate category distinctions (e.g. between animate and inanimate objects) and then later capturing finer and finer distinctions; the representations in the network exhibited periods of relatively little change punctuated by relatively rapid transitions in which subcategories became differentiated, capturing finer categorical distinctions; (2) developmentally and as a function of frequency and degree of typicality, representations of items captured shared and idiosyncratic aspects of items to varying degrees, with more overall experience, higher frequency of an item, and greater idiosyncrasy of an item leading to relatively greater degrees of differentiation; yet (3) even differentiated representations still captured the gross categorical structure of the domain, in that even highly differentiated animals remained more similar to each other than any of the animals were to any of the plants.

4.3.7 Complementary learning systems All of the distributed neural network models considered thus far rely on the back-propagation learning algorithm or other closely related error-correcting learning algorithms. A characteristic of such models is that they tend to learn relatively slowly: change occurs gradually, in what I have often termed "developmental time." In general such gradual learning appears psychologically well justified, capturing the gradual change in children's acquisition of inflectional patterns (as Brown, 1973 first noted – and contra claims by Marcus, Pinker, Ullman et al. 1992 – all of the different inflections in English are acquired gradually over a period of about a year: see McClelland and Patterson, 2002a for details), the gradual change in children's ability to read (as modeled by Seidenberg and McClelland, 1989), and gradual changes in semantic cognitive abilities, including reorganization of semantic representations (as modeled by Rogers and McClelland, 2004), over the age range from six to 12 and beyond. Yet children and adults can learn new things quickly. Early attempts to explore such rapid learning using distributed neural networks led to the discovery that they were susceptible to *catastrophic interference*: Any attempt to rapidly learn new information even partially inconsistent with knowledge already stored in the system led to disruption of the knowledge already stored in the connections (McCloskey and Cohen, 1989). This finding contributed to a loss of enthusiasm for distributed neural networks as models of learning and memory among some researchers.

However, a consideration of the human amnesic syndrome, as exhibited by patients with bilateral damage to the specialized brain areas in the medial temporal lobes, suggested that the brain might have evolved two complementary learning systems

that provide a solution to the catastrophic interference problem – something it would have to do if the basis of knowledge of language, semantics, and many other things is indeed to be found in the kinds of distributed neural network models reviewed above. In the amnesic syndrome, patients exhibit a profound inability to learn new arbitrary information rapidly, while still retaining the full complement of linguistic and semantic knowledge they had acquired prior to sustaining damage. Critically, all aspects of prior linguistic and semantic knowledge – including knowledge of exceptional aspects of words and things – are spared in the amnesic syndrome. Such patients do have difficulty with the rapid acquisition of new lexical items and with the formation of new episodic memories, and exhibit selective loss of memory for episodes occurring within a window of time ranging from months to years prior to the loss of medial temporal lobe function.

To address these findings, my colleagues and I proposed the *complementary learning systems theory* (McClelland, McNaughton and O'Reilly, 1995). In this theory it is proposed that the bulk of the forebrain including all areas of the neocortex outside of the medial temporal lobes is part of a structured learning system in which highly overlapping patterns of activation and therefore highly overlapping ensembles of connection weights are used for the representation and processing of related items. All of the models reviewed above exemplify these characteristics. In relevant simulations, we focused on the semantic network model introduced by Rumelhart (1990) and explored further by Rogers and McClelland (2004). Attempts to teach such networks new arbitrary information in a focused manner (without interleaving with ongoing exposure to examples illustrating the full distribution of characteristics across items) led to catastrophic interference (see McClelland, 2013, for a recent detailed examination of these issues). We argued (drawing on an earlier related proposal by Marr, 1971) that the pattern of findings in amnesia suggested that the rapid learning of the contents of new experiences – including experiences of objects and linguistic expressions – was primarily subserved by learning mechanisms in the medial temporal lobes. This scheme allows the rapid learning of new material without interference with existing knowledge; ultimately, though, our theory (like Marr's, and consistent with the suggestions of Milner, 1966 and Squire, 1992) still proposes that semantic and linguistic knowledge becomes integrated into the distributed neural networks in the neocortex. This integration occurs through gradual learning, interleaved with learning of other items, as all of the above models propose. Such learning may occur either through ongoing experience with relevant information during waking life or through replay of relevant patterns initially stored in the medial temporal lobes during off-line activities, including sleep.

An important aspect of the theory is that the medial temporal lobes are heavily interconnected with the neocortex: they receive their inputs from it, and send their outputs back to it. While the representations used for individual items in the medial temporal lobes are thought to be distributed patterns of activation, the theory holds that these patterns are relatively sparse and rely on specialized hippocampal circuitry to minimize overlap, so that they can be usefully approximated as though they were exemplar-like in character (Kumaran and McClelland, 2012). Critically, however, the inputs to the hippocampus arise from the neocortex, and so depend on the learned distributed representations that arise gradually in the neocortical learning system (McClelland and Goddard, 1996). This allows even the exemplar representations to depend critically on gradual structured learning, overcoming one of the key limitations of classical exemplar models noted above.

In summary, the complementary learning systems theory provides a more complete overall theory of learning and memory than that provided by the various distributed neural network models described above. The simple exemplar-like learning system in the hippocampus complements the more structured learning system in the neocortex, which remains the principal substrate for semantic and linguistic knowledge, and which plays the critical role in addressing both regular and quasi-regular aspects of language and other forms of skilled knowledge-dependent cognition.

5. Modeling the Emergence of Quasi-Regular Forms through Graded Constraints on Phonological Representations

I have argued above that distributed neural network models provide an opportunity to model the process of gradual language change over historical time as well as gradual representational change over developmental time. Several neural network models have been applied to language change, including an early model of the coalescence of the English regular past tense out of the strong and weak verb system characteristic of Middle English (Hare and Elman, 1995). Here we briefly consider a model offered by Lupyan and McClelland (2003) that re-examined this issue with specific reference to several of the themes of the present chapter.

We began from an observation by Burzio (2002) that regular English past tenses have phonological forms that violate the phonotactic constraints observed by monomorphemic English word forms, but irregular past tenses do not violate these constraints. As examples, consider regular *taped* and irregular *kept*. There are no monomorphemic English word forms whose rhymes contain both a long vowel and two stop consonants, but there are such word forms that contain both a short vowel and two stop consonants, such as *inept*, *apt*, and *act*. Reducing the vowel in *keep* preserves these phonotactic constraints. Regular words like *taped* maintain the stem of the verb, and thus are both regular and transparent but at the expense of excessive phonological complexity, while irregular *kept* reduces this complexity, paying a price in terms of reduced (but, importantly, not completely eliminated) regularity and transparency. For the most part, as Burzio noticed, the irregular past tense forms of English are no more complex phonologically than their stems, and sometimes they can even be seen as slightly less complex. In other cases, they trade the addition of a segment (e.g. the final /t/ added to *keep*) for a vowel reduction, at least partially ameliorating the added complexity due to the added segment. A key further observation is that many quasi-regular past tense forms with this or similar reductions are past tenses of very high frequency, including *did*, *said*, *had*, and *made*: the first two of these involve the regular inflection with a vowel reduction while the second two involve the regular inflection on a reduction of the stem. It is clear that at least some of these forms evolved to their reduced forms since Middle English, where, for example, the verb that is now *make* was a regular member of the weak verb system, with past tense *makode*. Thus *irregularization* occurred for this form over historical type.

Lupyan and McClelland sought to model these changes using a distributed neural network model that simultaneously embodied the constraints of (1) correctly specifying

the semantic content of each verb (including its tense semantics) when presented with either the present or the past tense phonology of the item, while (2) adjusting the phonological form of each item to minimize its length while still leaving it sufficient for successful communication of its meaning. The model was not intended to represent a single learner, but rather to capture the pressures operating on the language system over historical time, and to capture how these pressures could produce the gradual reduction and in some cases the complete elimination of particular phonemes as the system of verb forms evolved under these joint pressures.

The model employed a set of input units over which the phonology of monosyllabic, monomorphemic word forms could be represented by activating units corresponding to individual phonemes in sets corresponding to the onset, vowel nucleus, and consonant coda of the word form, while one additional unit was provided to allow for a possible additional /d/ or /t/ phoneme corresponding to an inflection. To capture differences in vowel complexity, each simple vowel was represented by a single active unit in the vowel pool while each long vowel or diphthong was represented by two active units. The degree to which a phoneme was present in a word form was captured by the extent of its activation, represented as a real number between 0 and 1. Each word form in the corpus had paired with it a semantic pattern; each word was presented in both a present and a past tense form. We compared learning of two corpora, one in which the semantics was fully compositional (same semantic pattern for the present and past tense form, with past tense represented by the same small set of additional active units added to the base semantic patterns of the item), and another in which there were small, idiosyncratic (randomly generated) differences in the semantics associated with the present and past tense forms, so that the semantic representations might be described as quasi-compositional.

The phonological representations used in the model were subjected to two graded pressures: first, to correctly produce the semantic pattern corresponding to the present and past tense phonology of each stem, and, second, to keep the phonology as simple as possible. To capture the first pressure, back-propagation was used not only to adjust the connection weights in the network, but also to adjust the activations of the phoneme units to ensure that the phonological input was capable of producing the correct target activation (this technique was first used by Miikkulainen and Dyer, 1987, to learn representations for words in distributed neural network models). To capture the second pressure, there was a cost associated with the degree of activation of each phoneme unit. This cost was imposed directly on the activation of each phoneme in each word, so that there would be a tendency to reduce its activation, allowing phonemes to gradually disappear from the representations of words if they were unnecessary for successful communication (see Lupyan and McClelland, 2003 for details).

The simulations were successful in showing how the two graded pressures described above could allow for the emergence of quasi-regular items. In one simulation, a fully regular initial training corpus was used, and this led to a reduction of the stem in most of the highest-frequency items. In a second simulation, verbs from several Old English strong verb clusters (such as those exemplified by *sing* and *think*) were included in the corpus together with their irregular past tenses (*sang*, *thought*). These clusters tended to be preserved while initially regular high-frequency items like *make* became quasi-regular by reduction or elimination of activation of stem phonemes. The model is but a first step toward addressing language change and has several limitations, one of which we consider below. Nevertheless, it successfully illustrated the gradual, continuous reduction

of phonological content of items. The joint effects of a pressure to communicate effectively while maintaining the simplest possible representation of each item resulted in the emergence of quasi-regular forms with phonological reductions similar to many of the quasi-regular past tense forms in English.

6. Evaluation of the Distributed Neural Network Models and Comparison to Other Contemporary Approaches

Most of the distributed neural network models reviewed above were introduced in the 1980s or early 1990s. Although work with such models is ongoing, many investigators now pursue other approaches, including structured probabilistic models and dynamical systems models. One might reflect on this and ask whether these alternative frameworks (or others yet to be introduced) should replace distributed neural network models, because of limitations inherent in the approach that the other approaches might be able to overcome, or whether, instead, apparent limitations of distributed neural network models that might have led some to explore alternatives might eventually be addressed. Here, I will first consider what I see as the specific advantages of distributed neural network models relative to structured probabilistic models and dynamical systems models. Then I will consider some of the factors that may have limited the appeal of existing distributed neural network models. Finally, I will point to recent signs of resurgence of interest in such models, and to reasons for believing that they will continue to play an important role in the future development of attempts to understand processing, representation, development, and historical change in natural cognitive domains such as language and natural kind semantics.

6.1 Comparison to structured statistical models

Many contemporary approaches to understanding cognition and language rely on structured statistical models (Griffiths, Chater, Kemp, Perfors, and Tenenbaum, 2010). Such models approach language and cognition as abstract computational problems framed as a search for a structured ensemble of hypotheses selected from a complex hypothesis space. Selection among hypotheses is constrained jointly by considerations of simplicity and of correctly accounting for the training data, which might be, for example, a corpus of sentences. These models have much in common with distributed neural network models in that both can involve finding a good solution to a set of simultaneous constraints, which may be graded or continuous in nature.

The key differences between such approaches and distributed neural network approaches appear to be differences in the pre-specification of a formal representation language for capturing alternative hypotheses. While structured statistical models pre-specify a space of possible hypotheses, sometimes in a formal language such as context-free rewrite rules, Boolean expressions, or first-order predicate logic, distributed neural network models attempt to make minimal assumptions about such representations, and leave the representation of such structure implicit in the knowledge stored in the connections among the units in the system. For example, Perfors, Tenenbaum, and Regier (2011) considered how a structured statistical model could use training data to

select among three alternative hypotheses about the nature of the grammar underlying the sentences heard by a child learning language. They found that a context-free rewrite grammar provided a better account than a simple transition network grammar or a third alternative. The approach appears to show, as distributed neural network researchers have known since Elman's work, that aspects of English grammar can be learned from a training corpus, but on close inspection, Perfors et al. only shows that once one has the right form of hypothesis to compare to other alternatives, selection among them can be made using statistics. Elman's work appears to go further in showing that no pre-commitment to any formal representation language (other than the generic language of multi-layer neural networks) is necessary to acquire the structure of natural language. Similar points can be made about the approach to representing the structured knowledge people have of natural kind statistics taken by Kemp and Tenenbaum (2009). Recent work with analytically tractable versions of models like those used by Rogers and McClelland shows that learned distributed representations that capture human knowledge of natural kind semantics can closely approximate the hierarchical and other structures considered by Kemp and Tenenbaum, without needing to build such representations in advance (Saxe, McClelland, and Ganguli, 2013).

More recent work within the structured statistical framework has been useful in capturing aspects of language structure, such as the distribution of kinship terms (Regier, Kemp, and Kay, chapter 11, this volume). This model, however, adheres to the characteristics of classical models in that it adopts a pre-specified taxonomy of concepts and a system of rules for constructing complex expressions from other expressions. These representations provide a useful high-level summary of some of the factors that affect the selection of kinship systems, but take a great deal as given. I would conjecture that further research from a more fully emergentist perspective will acquire representational systems of comparable expressivity without prior stipulation of such concepts and rules.

6.2 *Comparison to dynamical systems models*

A comparison of neural network models with dynamical systems models is made difficult in part by the diversity of approaches that fall under the heading of "dynamical systems" (see the chapters in Spencer, Thomas, and McClelland, 2009 for many examples). Dynamical systems researchers tend to seek simple characterizations of complex systems in terms of qualitative signatures, including such concepts as attractors, bi- or multi-stability, inaccessible regions, bifurcations, and so on. If, however, a dynamical system is thought of as a continuous time-varying system governed by non-linear, stochastic differential equations, then neural networks are examples of dynamical systems, and the concepts of dynamical systems analysis can be applied to them (McClelland and Vallabha, 2009). I believe that stochastic, continuous time activation dynamics applies to all aspects of human cognitive processing (McClelland, 1993), and the presence of trial-to-trial variability in human response times in every task supports this belief. Some distributed networks use a single deterministic activation step to compute outputs from given inputs, but to me this is a simplification adopted for tractability (McClelland, 2009) rather than a claim about the nature of processing.

Treated as examples of dynamical systems, neural networks exhibit many of the features that protagonists of dynamical systems approaches point to, though some

neural networks models exhibit such features more clearly than others. Perhaps, for example, the transitions between states of knowledge exhibited by some neural network models are not as abrupt or noisy as those seen in certain human developmental transitions. However, close scrutiny of developmental data often reveals that transitions are more gradual than previously thought, leaving open questions about whether existing distributed neural network models are sufficient or not (Schapiro and McClelland, 2009).

One last, but very important, point worth making about the difference between distributed neural network models and dynamical systems models is the fact that the latter often fail to provide a mechanistic or process-based characterization of developmental or learning-based change (McClelland and Vallabha, 2009). For example, Schutte, Spencer, and Schöner (2003) offer a dynamical systems characterization of differences between children of different ages, in terms of differences in the widths of the basins of attraction these investigators use in characterizing distortions in the reaching behavior of young children. While the width of a basin of attraction may provide an adequate descriptive characterization of the patterns of responses made by children of different ages, it fails to provide an explanation of how it is that the widths of these basins of attraction change. One view of this matter is to construe the characterization Schutte and colleagues offer as a higher-level descriptive account of the characteristics that might arise in a distributed neural network model that gradually improves the precision of its representations through learning.

In spite of the differences between dynamical systems and neural network modeling approaches, I would certainly encourage further efforts to integrate the two approaches, as proposed in Spencer, Thomas, and McClelland (2009).

6.3 *Limitations of distributed neural networks*

While distributed neural network models have many virtues, many of which I have attempted to enumerate above, they suffer also from limitations that have contributed to the appeal of alternative approaches. Here I will briefly mention two such limitations, as well as two other areas of controversy surrounding many of the distributed neural network models reviewed above.

6.3.1 Stipulation and discreteness of input representations Virtually all of the distributed neural networks considered in this chapter employed surface representations (i.e., patterns used as inputs or outputs of the distributed neural network) that were specified by the modeler, and quite often characteristics of these surface representations are themselves problematic, particularly in that they tend to be discrete and categorical in nature. Sometimes, "localist" input representations for items such as letters or words are used; these representations presuppose, and treat as discrete, units such as phonemes, and words, even though they are in fact far from discrete in real spoken language. Such units are often used in models that show that more abstract levels of structure can emerge, but the use of such units still presupposes and builds in too much in my view. One case in point is the phonological representation used in the model of Lupyan and McClelland, where separate units were provided for onset, nucleus, and coda phonemes. Though the presence of each phoneme could, in this model, be treated as a matter of degree, the phonemes themselves were still discrete and such reductive processes as palatalization

or neutralization could not be effectively modeled. What is needed is a way to model the processing of spoken input directly from the speech stream, so that representations at all levels of structure can be captured directly as emergent phenomena.

6.3.2 Use of restricted corpora A second limitation of many of the models considered above is the very limited nature of the training corpora they employ. While the models of single-word reading and inflection have tended to use corpora based on characteristics of real language, models of sentence processing and semantic knowledge representation have tended to use far more restricted, and often entirely synthetic, corpora, thereby leaving themselves open to the criticism that they might not scale up to address the full complexity of real natural language. With regard to the first point, many models have restricted themselves to simple forms, such as monosyllabic word forms or one-clause sentences, thereby raising questions about the framework's ability to extend to these more complex structures.

The two remaining issues we now consider are potentially more controversial. While many consider these inherent weaknesses, it is not entirely clear that they really are intrinsic shortcomings of the models.

6.3.3 Lack of transparency and analytic tractability The first such issue we will consider is the lack of transparency of the representations and processes embedded in distributed neural network models. When such models succeed, their success may still require further explication. Why did they succeed? What features of the model were essential and which only incidental? What features are responsible for insufficiencies of the models? Difficulties of this sort have led some to wonder in just what sense we ought to see such models as offering any explanation for observed patterns of behavior. In contrast, the stipulation of a simple rule or set of rules may appear to offer a sense of greater clarity at least about what is being claimed by the protagonists of a particular model. My own position on this issue is somewhat circumspect. I appreciate that it is often useful to be able to offer an explicit quantitative theory capturing the processes at work in a model; but we should not necessarily expect such a theory to be easy to develop, nor should we expect a truly simple formal system to provide a fully adequate characterization. The beauty and simplicity of the grammars Chomsky enticed us with in *Syntactic Structures* (1957) and *Aspects* (1965) turned out to be illusory, as have similar claims for the more contemporary Minimalist program (Chomsky, 1995; see Newmeyer, 2003). While some may still seek the deep insight that would allow a very simple and still complete characterization, an emergentist perspective holds that such a characterization must always be partial and approximate.

6.3.4 Insufficient respect for structure The second controversial issue lies in the concern that the representations used in distributed neural network models are insufficient to allow them to capture the full systematicity and productivity of language or other forms of human cognition (Fodor and Pylyshyn, 1988; Griffiths, Chater, Kemp, Perfors, and Tenenbaum, 2010). To be sure, some models use restricted inputs that cannot do full justice to the complexities of the thoughts that minds can entertain. For example, the Rumelhart network used in the simulations of natural kind semantics by Rogers and McClelland (2004) can only process simple propositions consisting of an item, a relation, and a single attribute or other item, such as *canary can fly* or *robin is a bird*. Clearly these

propositions do not reflect the full expressive power of natural languages. The question that remains open for debate is whether the use of explicit recursively defined and hierarchically structured representations of the kind provided by a syntactic parse tree is a necessary component of a successful model of language, as has recently been argued by Berwick, Pietroski, Yankama and Chomsky (2011). While it seems clear that sentences have constituent structure, this structure may not always be clear and in any case may be emergent; the explicit representation of that structure as such may turn out not to be necessary.

6.4 Future prospects for distributed neural network models

The question we now face is whether the above limitations and controversies facing distributed neural network models are inherent and insurmountable or whether the future will lead to superior models that address these issues. While I have no crystal ball, I see reasons for optimism in the recent work using neural networks in large-scale machine learning applications and in developing deeper mathematical analyses of such networks. Below I consider some of these developments.

6.4.1 Avoiding stipulation and discreteness in surface representations The ultimate inputs to the human cognitive system are the time-varying patterns of light of various wavelengths that reach the retina, the time-varying pattern of acoustic pressure that reaches the ear, and time-varying inputs in other sensory modalities. In accordance with this, in the domains of both vision and speech, contemporary distributed neural networks used in the field of machine learning are working directly from minimally preprocessed inputs. Such neural networks now allow mobile phone service providers to interpret spoken requests involving arbitrarily complex naturally spoken sentences (Mohamed, Dahl and Hinton, 2009), and allow machine categorization and detection of the objects present in images and videos at ever-improving degrees of specificity (10,000 distinct categories, including a large number of subcategories, are included in current category taxonomies; Le, Ranzato, Monga et al., 2012). These neural networks often involve many layers, each trained using back-propagation or a related algorithm to form an internal representation sufficient to reconstruct its input on its output, and also constrained to minimize complexity of the internal representation. The successes of such models in capturing aspects of the representations neurophysiologists find when recording from neurons suggest that the constraints operating on learning in such systems are sufficient to extract human-like representations without supervision, and thereby allow one to imagine future cognitive models that would place much less reliance than earlier models did on stipulation of features of input representations. *Use of restricted corpora*. Contemporary neural network research in machine learning has also overcome the restricted scope of the corpora used in the models described earlier in this chapter. Huge corpora are used to train the networks for machine speech perception and object recognition cited above, and Socher, Bauer, Manning, and Ng (2013) have trained what they call a "Matrix-Vector Neural Network Model" of sentence processing to classify the sentiment expressed in single-sentence descriptions of movies, using a corpus of 10,000 such sentences for which humans have provided sentiment ratings. While some of the expressions of sentiment are fairly easy to categorize, others are conveyed in highly complex sentences, including the following examples: "Doesn't come close to justifying the hype

that surrounded its debut at the Sundance film festival two years ago," and "Not always too whimsical for its own good, this strange hybrid of crime thriller, quirky character study, third-rate romance and female empowerment fantasy never really finds the tonal or thematic glue it needs." Currently this model represents the state of the art, beating other models of sentiment classification. The contemporary machine learning models make use of a number of enhancements to the most basic multi-layer neural network architectures, but none of these enhancements fundamentally changes the basic commitment to the use of simple neuron-like processing units without predefined meaning which is the hallmark of distributed neural network research since the introduction of back-propagation. While these enhancements are likely to contribute to the success of these current models, another reason for their success may be the large-scale corpora (and large-scale computer clusters) that are available for use in these models' training.

6.4.2 Use of compressed compositional representations The model of Socher and colleagues does make use of a tree-like representation of sentence structure. That is, the model derives representations of word sequences by combining pairs of constituents from the bottom of the tree upward, and replacing each pair with an equal-length pattern vector representing the combined expression as a whole, as proposed initially by Pollack (1990). Interestingly, the choice of which constituents to combine may be guided either by an explicit parse tree provided by a structured probabilistic syntactic parser, or by considering at each step in the upward pass which pairs of constituents fit together best (see Socher, Perelygin, Wu et al., 2013 for details). Thus, while some reliance on grouping words into larger constituents may contribute to the model's success, future research is needed to determine whether even this level of concession to an explicitly structured sentence representation is necessary. One alternative is the possibility that a future model trained with the same amount of data could work as well, simply progressively updating a representation of sentiment (or other evaluation of an aspect of the meaning conveyed by the sentence) as it works its way forward through a spoken sentence, as in the Sentence Gestalt model of St. John and McClelland (1990).

6.4.3 Developments in formal theory of learning in multi-layer neural networks The recent success of neural networks for machine learning comes, for the most part, from using "deep networks," composed on many layers between inputs and outputs, and/or from the use of learned distributed representations of words and larger constituents that are not explicit with respect to their meaning or content. This being so, these models may be even less analyzable than the models reviewed earlier in this chapter. There has, however, been some progress in developing an analytic understanding of the learning trajectories of a useful simplified version of multi-layer neural networks, one in which the non-linear processing units standard in such networks are replaced for analytic tractability with simpler, linear processing units. Multi-layer networks of linear units are restricted in the computations they can perform, but nevertheless reveal interestingly complex learning dynamics similar to what is seen in networks with non-linear processing units (one reason for this is that networks are typically initialized in such a way that they perform in an approximately linear regime, at least during the initial stages of learning). For example, the progressive differentiation of representations of natural kinds seen in the deep non-linear networks used by Rogers and McClelland is

also exhibited by simplified networks that employ linear units, and closed-form mathematical expressions that characterize the trajectory of learning in such networks as a function of the statistical structure present in the training corpus have been developed (Saxe, McClelland, and Ganguli, 2013). It will be interesting to see whether such analyses can be extended further, to allow greater analytic understanding of the outcome and trajectory of learning in a fuller range of contemporary network architectures.

7. Summary and Conclusion

In this chapter I have reviewed the evidence for graded constituent structure, gradual change, and quasi-regularity in several sub-domains of language and cognition. This evidence motivated the use of distributed neural network models to explore how well they could capture aspects of language without requiring an explicit taxonomy of units and rules for combining and manipulating them. The models reviewed in the main body of the chapter captured many of the motivating aspects of language, although these models do have some limitations. While the attention of some has recently shifted toward structured probabilistic and dynamical systems models, I have argued that the future prospects for modeling language and cognition using distributed neural networks are very bright. The ability to avoid stipulating and discretizing the surface representations used as inputs to such models and the availability of large training corpora and large-scale computational resources for training such models may overcome many of the earlier models' limitations. It remains to be seen how far such models can go in allowing language to be captured as arising historically, developmentally, and in the moment from the processes that operate as users communicate with each other using sound or gesture as their medium of communication.

REFERENCES

Aronoff, M. 1976. *Word Formation in Generative Grammar*. Linguistic Inquiry Monographs, 1. Cambridge, MA: MIT Press.

Berwick, R, P. Pietroski, B. Yankama, and N. Chomsky. 2011. Poverty of the stimulus revisited. *Cognitive Science* 35: 1207–1242.

Bever, T. G. 1970. The cognitive basis for linguistic structures. *Cognition and the Development of Language* 279(362): 1–61.

Brown, R. 1973. *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.

Bryant, B. D. and R. Miikkulainen. 2001. From word stream to Gestalt: A direct semantic parse for complex sentences. Technical Report AI98-274, AI Lab, University of Texas at Austin, June.

Bybee, J. 1985. *Morphology: A Study of the Relations Between Meaning and Form*. Philadelphia, PA: John Benjamins.

Bybee, J. 2001. *Phonology and Language Use*. Cambridge: Cambridge University Press.

Bybee, J. 2006. From usage to grammar: The mind's response to repetition. *Language* 82: 529–551.

Bybee, J. and J. L. McClelland. 2005. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *Linguistic Review* 22(2–4): 381–410.

Burzio, L. 2002. Missing players: Phonology and the past-tense debate. *Lingua* 112: 157–199.

Chang, F., G. S. Dell, and K. Bock. 2006. Becoming syntactic. *Psychological Review* 113(2): 234–272.

Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton.
Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
Chomsky, N. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.
Coltheart, M., B. Curtis, P. Atkins, and M. Haller. 1993. Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review* 100(4), 589–608.
Croft, W. 2001. *Radical Construction Grammar*. Oxford: Oxford University Press
Daugherty, K. G., M. C. MacDonald, A. S. Petersen, and M. S. Seidenberg. 1993. Why no mere mortal has ever flown out to center field but people often say they do. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, pp. 383–388. Hillsdale, NJ: Erlbaum.
Dilkina, K., J. L. McClelland, and L. Boroditsky. 2007. How language affects thought in a connectionist model. In D. S. McNamara and J. G. Trafton (eds.), *Proceedings of the 29th Annual Conference of Cognitive Science Society*, pp. 215–220. Austin, TX: Cognitive Science Society.
Dilkina, K., J. L. McClelland, and D. C. Plaut. 2008. A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology* 25(2): 136–164.
Elman, J. L. 1990. Finding structure in time. *Cognitive Science* 14(2): 179–211.
Elman, J. L. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7(2–3): 195–225.
Fodor, J. A. and Z. W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition* 28(1): 3–71.
Goldberg, A. E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: Chicago University Press.
Gonnerman, L. M., M. S. Seidenberg, and E. S. Andersen. 2007. Graded semantic and phonological similarity effects in priming: Evidence for a distributed connectionist approach to morphology. *Journal of Experimental Psychology: General* 136(2): 323–345.
Griffiths, T. L., N. Chater, C. Kemp, A. Perfors, and J. B. Tenenbaum. 2010. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences* 14(8): 357–364.
Hare, M. and J. L. Elman. 1995. Learning and morphological change. *Cognition* 56(1): 61–98.
Harm, M. W. 2002. Building large scale distributed semantic feature sets with WordNet. *CNBC Tech Report PDP.CNS.02.1*.
Hoeffner, J. H. and J. L. McClelland. 1993. Can a perceptual processing deficit explain the impairment of inflectional morphology in development dysphasia? A computational investigation. In E. V. Clark (ed.), *The Proceedings of the Twenty-Fifth Annual Child Language Research Forum*, pp. 38–49. Stanford, CA: Center for the Study of Language and Information.
Jackendoff, Ray. 2007. Linguistics in cognitive science: The state of the art. *Linguistic Review* 24: 347–401.
Kemp, C. and J. B. Tenenbaum. 2009. Structured statistical models of inductive reasoning. *Psychological Review* 116(1): 20–58.
Kruschke, J. K. 1992. ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review* 99(1): 22–44.
Kumaran, D. and J. L. McClelland. 2012. Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review* 119: 573–616.
Lachter, J. and T. G. Bever. 1988. The relation between linguistic structure and associative theories of language learning: A constructive critique of some connectionist learning models. *Cognition* 28(1): 195–247.
Le, Q. V., M. A. Ranzato, R. Monga, M. Devin, K. Chen, G. S. Corrado, J. Dean, and A. Y. Ng. 2012. Building high-level features using large scale unsupervised learning. In John Langford and Joelle Pineau (eds), *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, June 27–July 3, 2012, pp. 81–88. Madison, WI: Omnipress. http://www.icml.cc/2012/files/handbook.pdf. Also available as arXiv preprint arXiv:1112.6209.
Lupyan, G. and J. L. McClelland, 2003. Did, made, had, said: Capturing quasi-regularity in exceptions. In R. Alterman and D. Hirsh (eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pp. 740–745. Mahwah, NJ: Lawrence Erlbaum Associates.
MacWhinney, B. and J. Leinbach. 1991. Implementations are not conceptualizations: Revising the verb learning model. *Cognition* 40(1): 121–157.

MacWhinney, B., J. Leinbach, R. Taraban, and J. McDonald/ 1989. Language learning: Cues or rules? *Journal of Memory and Language* 28(3): 255–277.

Marcus, G. F., U. Brinkmann, H. Clahsen, R. Wiese, and S. Pinker. 1995. German inflection: The exception that proves the rule. *Cognitive Psychology* 29(3): 189–256.

Marcus, G. F., S. Pinker, M. Ullman, M. Hollander, T. J. Rosen, F. Xu, and H. Clahsen. 1992. Overregularization in language acquisition. *Monographs of the Society for Research in Child Development* 57(4): 1–178.

Marr, D. 1971. Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 262(841): 23–81.

McClelland, J. L. 1981. Retrieving general and specific information from stored knowledge of specifics. In *Proceedings of the Third Annual Conference of the Cognitive Science Society*, pp. 170–172.

McClelland, J. L. 1992. Can connectionist models discover the structure of natural language? In R. Morelli, W. M. Brown, D. Anselmi, K. Haberlandt, and D. Lloyd (eds.), *Minds, Brains and Computers: Perspectives in Cognitive Science and Artificial Intelligence*, pp. 168–189. Norwood, NJ: Ablex.

McClelland, J. L. 1993. Toward a theory of information processing in graded, random, interactive networks. In D. E. Meyer and S. Kornblum (eds.), *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence and Cognitive Neuroscience*, pp. 655–688. Cambridge, MA: MIT Press.

McClelland, J. L. 2009. The place of modeling in cognitive science. *Topics in Cognitive Science* 1(1): 11–38.

McClelland, J. L. 2013. Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *Journal of Experimental Psychology: General* 142(4): 1190–1210. doi: 10.1037/a0033812.

McClelland, J. L. and Bybee, J. 2007. Gradience of gradience: A reply to Jackendoff. *Linguistic Review* 24: 437–455.

McClelland, J. L. and Goddard, N. 1996. Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus* 6: 654–665.

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

McClelland, J. L. and K. Patterson. 2002a. Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Science* 6(11): 465–472.

McClelland, J. L. and K. Patterson. 2002b. "Words *or* Rules" cannot exploit the regularity in exceptions. *Trends in Cognitive Science* 6(11): 464–465.

McClelland, J. L., T. T. Rogers, K. Patterson, K. N. Dilkina, and M. R. Lambon Ralph. 2009. Semantic cognition: Its nature, its development, and its neural basis. In M. Gazzaniga (ed.), *The Cognitive Neurosciences*, 4th ed., pp. 1047–1966. Cambridge, MA: MIT Press.

McClelland, J. L. and D. E. Rumelhart. 1981. An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review* 88: 375–407.

McClelland, J. L. and D. E. Rumelhart. 1985. Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General* 114: 159–197.

McClelland, J. L., M. St. John, and R. Taraban. 1989. Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes* 4: SI 287–335.

McClelland, J. L. and G. Vallabha. 2009. Connectionist models of development: Mechanistic dynamical models with emergent dynamical properties. In J. P. Spencer, M. S. C. Thomas, and J. L. McClelland (eds.), *Toward a Unified Theory of Development: Connectionism and Dynamic Systems Theory Re-considered*, pp. 3–24. New York: Oxford University Press.

McCloskey, M. and N. J. Cohen, 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation* 24: 109–164.

Miikkulainen, R. and M. G. Dyer. 1987. Building distributed representations without microfeatures. UCLA-AI-87-17. Artificial Intelligence Laboratory, Computer Science Department, University of California, Los Angeles.

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

Milner, B. 1966. Amnesia following operation on the temporal lobe. In C. W. M. Whitty and Oliver L. Zangwill (eds.), *Amnesia*, pp. 109–133. London: Butterworth.

Mohamed, A., G. Dahl, and G. Hinton. 2009. Deep Belief Networks for phone recognition. *Science* 4(5): 1–9. doi:10.4249/scholarpedia.5947.

Newmeyer, F. 2003. Review of *On nature and language*, by N. Chomsky. *Language* 79: 583–599.

Nosofsky, R. M. 1984. Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10(1): 104–114.

Patterson, K., M. A. Lambon Ralph, E. Jefferies, A. Woollams, R. Jones, J. R. Hodges, and T. T. Rogers. 2006. "Presemantic" cognition in semantic dementia: Six deficits in search of an explanation. *Journal of Cognitive Neuroscience* 18(2): 169–183.

Perfors, A., J. B. Tenenbaum, and T. Regier. 2011. The learnability of abstract syntactic principles. *Cognition* 118(3): 306–338.

Pierrehumbert, J. B. 2001. Exemplar dynamics: Word frequency, lenition and contrast. *Frequency and the Emergence of Linguistic Structure* 45: 137–157.

Pinker, S. 1991. Rules of language. *Science* 253: 530–535.

Pinker, S. 1999. *Words and Rules*. New York: Basic Books.

Pinker, S. and A. Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28(1): 73–193.

Plaut, D. C. and L. M. Gonnerman. 2000. Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes* 15: 445–485.

Plaut, D. C. and J. L. McClelland. 2010. Locating object knowledge in the brain: A critique of Bowers' (2009) attempt to revive the grandmother cell hypothesis. *Psychological Review* 117: 284–288.

Plaut, D.C., J. L. McClelland, M. S. Seidenberg, and K. Patterson. 1996. Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review* 103: 56–115

Plunkett, K. and V. Marchman. 1991. U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition* 38(1): 43–102.

Plunkett, K. and V. Marchman. 1993. From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition* 48(1): 21–69.

Pollack, J. B. 1990. Recursive distributed representations. *Artificial Intelligence* 46(1): 77–105.

Reali, F. and M. H. Christiansen. 2005. Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science* 29(6): 1007–1028.

Rogers, T. T., M. A. Lambon Ralph, P. Garrard, S. Bozeat, J. L. McClelland, J. R. Hodges, and K. Patterson. 2004. Structure and deterioration of semantic memory: A neuropsychological and computational investigation. *Psychological Review* 111: 205–235.

Rogers, T. T. and J. L. McClelland. 2004. *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge, MA: MIT Press.

Rohde, D. L. 2002. A connectionist model of sentence comprehension and production. Doctoral dissertation, Carnegie Mellon University.

Rumelhart, D. E. 1990. Brain style computation: Learning and generalization. In S. F. Zornetzer, J. L. Davis, and C. Lau (eds.), *An Introduction to Neural and Electronic Networks*, pp. 405–420. San Diego: Academic Press.

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

Rumelhart, D. E. and J. L. McClelland. 1986. On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, and the PDP research group (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume II: Psychological and Biological Models*, pp. 216–271. Cambridge, MA: MIT Press.

Rumelhart, D. E. and P. M. Todd. 1993. Learning and connectionist representations. In D. E. Meyer and S. Kornblum (eds.), *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*, pp. 3–30. Cambridge, MA: MIT Press.

Saxe, A. M., J. L. McClelland, and S. Ganguli. 2013. Learning hierarchical category structure in deep neural networks. In M. Knauff, M. Paulen, N. Sebanz, and I. Wachsmuth (eds.), *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*, pp. 1271–1276. Austin, TX: Cognitive Science Society.

Schapiro, A. C. and J. L. McClelland. 2009. A connectionist model of a continuous developmental transition in the balance scale task. *Cognition* 110(1): 395–411.

Schutte, A. R., J. P. Spencer, and G. Schöner. 2003. Testing the dynamic field theory: Working memory for locations becomes more spatially precise over development. *Child Development* 74(5): 1393–1417.

Seidenberg, M. S. and J. L. McClelland. 1989. A distributed, developmental model of word recognition and naming. *Psychological Review* 96: 523–568.

Seidenberg, M. S. and D. Plaut. In press. Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive Science*.

Sejnowski, T. J. and C. R. Rosenberg, 1987. Parallel networks that learn to pronounce English text. *Complex Systems* 1(1): 145–168.

Socher, R., J. Bauer, C. D. Manning, and A. Y. Ng. 2013. Parsing with compositional vector grammars In *ACL 2013: 51st Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference. Volume 1: Long Papers*, pp. 455–465. Stroudsburg, PA: Association for Computational Linguistics, http://aclweb.org/anthology/P/P13/P13-1045.pdf.

Socher, R., A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. 2013. Recursive deep models for semantic compositionality over a Sentiment Treebank. In *2013 Conference on Empirical Methods in Natural Language Processing: Proceedings of the Conference*, pp. 1631–1642. Stroudsburg, PA: Association for Computational Linguistics. http://www.aclweb.org/anthology/D13-1170.

Spencer, J. P., M. S. C. Thomas, and J. L. McClelland. 2009. *Toward a Unified Theory of Development: Connectionism and Dynamic Systems Theory Re-Considered*. New York: Oxford University Press.

Squire, L. R. 1992. Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review* 99(2): 195.

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