Connectionist perspectives on lexical representation

David C. Plaut

1. Introduction

Words are often thought of as the building blocks of language, but the richness of their internal structure and the complexity of how they combine belies such a simple metaphor. Lexical knowledge encompasses multiple types of interrelated information—orthographic, phonological, semantic, and grammatical—with substantial degrees of ambiguity within each. It is perhaps not surprising then that, despite the extensive efforts put into studying lexical processing across multiple disciplines, our understanding of the cognitive and neural bases of word representation remains piecemeal.

The standard way of thinking about lexical representation is that each word is coded by some type of separate, discrete data structure, such as a "logogen" (Morton, 1969) or localist processing unit (Rumelhart & McClelland, 1981). Each such representation has no internal structure of its own but serves as a "handle" that links together the various types of information that comprise knowledge of the word. One interesting implication of this view is that, although words can be similar orthographically, phonologically, semantically, or grammatically, there's no sense in which, independent of these other dimensions, words can be similar *lexically*. That is, whereas the representation of each aspect of lexical knowledge defines a similarity space within which words can be more or less related to each other, lexical representations per se are fundamentally different in that each word is coded independently of every other word. In essence, lexical representations themselves have no relevant properties—they exist solely to solve a particular computational problem: how to bind together specific orthographic, phonological, semantic, and grammatical information so that each aspect can evoke the others and together contribute coherently to language processing more generally.

Although the traditional theory of lexical representation has considerable intuitive appeal, it runs into some difficulties when confronting the complexities of the internal structure and external relationships of words.

This chapter explores the possibility that a particular form of computational modelling—variously known as connectionist modeling, neural-network modeling, or the parallel distributed processing (PDP) approach, not only avoids these difficulties but, more fundamentally, provides a different solution to the problem that traditional lexical representations were created to solve in the first place. In particular, and as elaborated below, connectionist/PDP networks can learn the functional relationships among orthographic, phonological, semantic, and grammatical information even though no particular representation binds them all together in one place. In this way, connectionist/PDP modelling raises the radical possibility that, although there is certainly lexical knowledge and lexical processing, as traditionally construed *there is no such thing as lexical representation*.

2. Principles of connectionist representation

Connectionist models are composed of large groups of simple, neuron-like processing units that interact across positive- and negative-weighted connections. Typically, each unit has a real-valued activation level which is computed according to a non-linear (sigmoid) function of the weighted sum of activations of other, connected units. Different groups of units code different types of information, with some units coding input to the system and others coding the system's output or response to that input. Knowledge of how inputs are related to outputs is encoded in the pattern of weights on the connections among units; learning involves modifying the weights in response to performance feedback.

In thinking about how a group of units might represent entities in a domain, it is common to contrast two alternatives. The first is a *localist* representation, in which there is a one-to-one relationship between units and entities—that is, a single, dedicated unit corresponds to each entity. The second is a *distributed* representation, in which the relationship is many-to-many—that is, each entity is represented by a particular pattern of activity over the units, and each unit participates in representing multiple entities.[1]

The interactive activation (IA) model of letter and word perception (McClelland & Rumelhart, 1981) provides a useful context for clarifying this distinction. The model consists of three layers of interacting units: letter feature units at the bottom (various strokes at each of four positions), letter

units in the middle (one per letter at each position; e.g., B, L, U, and R), and word units at the top (one per word; e.g., BLUR). The IA model is usually thought of as localist because it contains single units that stand in one-to-one correspondence with words, but it is important to recognize that a representation is localist or distributed only relative to a specific set of entities. Thus, the word level of the IA model is localist relative to words, and the letter level is localist relative to (position-specific) letters. However, at the letter level, the presentation of a word results in the activation of multiple units (corresponding to its letters), and each of these units is activated by multiple words (i.e., words containing that letter in that position). Thus, the letter level in the IA model is localist relative to letters but distributed relative to words.

In practice, however, it can be difficult to distinguish localist from distributed representations on the basis of activity because localist units typically become active not only for the entity to which they correspond but also for entities that are similar to it. For example, in the IA model, the input for BLUR activates its word unit strongly but also partially activates the word unit for BLUE (see Bowers, 2009, p. 226). This off-item activation can be difficult to distinguish from the patterns that comprise distributed representations. Moreover, in most localist theories it is assumed that there are multiple redundant copies of each dedicated unit. Thus, in both localist and distributed representations, multiple units become active in processing a given entity, and each unit will become at least partially active for multiple entities.

A further consideration is that the number of active units in a representation—its *sparseness*—is a matter of degree. Localist representations constitute one extreme of sparseness, but distributed representations in which a very small percentage of units are active at any one time can be functionally quite similar, in that each pattern can have effects that are largely independent of the effects of other patterns. Even so, sparse distributed representations have a distinct advantage over strictly localist ones in that they provide far more efficient coding (O'Reilly & McClelland, 1994). Moreover, the degree of sparseness of learned internal representations within connectionist networks need not be stipulated *a priori* but arises as a consequence of the basic network mechanisms, the learning procedure, and the structure of the tasks to be learned. In general, systematic tasks—in which similar inputs map to similar outputs—yield denser activation to support general-

ization, whereas unsystematic tasks such as word recognition give rise to sparser activation to avoid interference (for discussion, see McClelland, McNaughton, & O'Reilly, 1995; Plaut, McClelland, Seidenberg, & Patterson, 1996).

An alternative characterisation of the locality of a representation is in terms of *knowledge* rather than activity (Bowers, 2009). That is, one can distinguish whether knowledge about an entity is encoded in the connections coming into or out of a particular unit or whether it is distributed across the connections of many units. For example, within the IA model, knowledge that the letter string BLUR is a word is coded only in the connections between the corresponding word unit and its letters; remove that single unit, and BLUR is no longer a word to the model.

Although this form of localist theory is clearly distinct from the types of knowledge typically learned by connectionist/PDP networks, it runs into difficulties when confronted with the general issue of the appropriate granularity of localist units—in particular, whether units should be allocated to individually encountered instances of entities or to some equivalence class of instances (Plaut & McClelland, 2010). The former case is problematic not only because it requires an unrealistic amount of storage but also because it doesn't explain how we recognize novel instances of familiar categories (e.g., a new car on the street, or this particular occurrence of the word BLUR). Assigning units to classes of instances is problematic because there will always be some further distinctions within the class that are important in some contexts but that are inaccessible because the instances are represented identically by the same localist unit. If both instance and class units are added, the knowledge about an entity is no longer localised to a single processing unit—that is, on this alternative formulation, the representation becomes distributed.

Although the issue of the granularity of localist representations is problematic in general, it could be argued that it is entirely straightforward in the specific case of words. That is, units should be allocated for each word, which corresponds to a class of instances (i.e., specific occurrences of that word). The reason this works is that words are s*ymbolic*—each instance of a word is exactly functionally equivalent to every other instance of the word, and so nothing is lost by representing them identically. Thus, even if localist representation is untenable in general, perhaps it is perfectly well-suited for lexical knowledge.

Unfortunately, localist representations face another challenge in this domain—capturing the internal structure of words.

3. The challenge of internal structure: Morphology

The real building blocks of language, if there were such things, would be morphemes. The traditional view of lexical representation is that words are composed of one or more morphemes, each of which contributes systematically to the meaning and grammatical role of the word as a whole (e.g., UNBREAKABLE = UN- + BREAK + -ABLE). If English morphology were perfectly systematic, lexical representation would have nothing to contribute beyond morphemic representation, and localist structures might be fully adequate for the latter. However, as is true of other linguistic domains, morphological systematicity is only partial. That is, the meaning of a word is not always transparently related to the meaning of its morphemes (e.g., a DRESSER is not someone who dresses but a piece of furniture containing clothes). Moreover, the meaning of a morpheme can depend on the word it occurs in (e.g., the affix -ER can be agentive [TEACHER], instrumental [MOWER], or comparative [FASTER], depending on the stem). In fact, some words decompose only partially (e.g., -ER is agentive in GROCER and instrumental in HAMMER, but what remains in each case [GROCE?, HAM?] is not a morpheme that contributes coherently to meaning). In short, the relationship of the meaning of a word to the meanings of its parts—to the extent it even has parts—is sometimes straightforward but can be exceedingly complex in general.

This complexity presents a formidable challenge to localist theories of lexical representation. First, the wealth of empirical data showing strong effects of morphological structure on the speed and accuracy of word recognition rules out a solution that involves units only for whole words. The fact that many words exhibit only partial semantic transparency also rules out having only morpheme units that contribute to meaning independently. The only viable approach would seem to be one in which both word and morpheme units are included, such that the word units compensate for any lack of transparency in the semantic contribution of individual morphemes (see, e.g., Taft, 2006). Even setting aside concerns about how the system would determine what morphemes are contained in a word, allo-

cate and connect the necessary units, and weight them relative to the word unit appropriately, the approach runs into problems because it forces morphological decomposition to be all-or-none. That is, a word either does or doesn't contain a morpheme, and if it does, the morpheme unit's contribution to meaning (as distinct from the word unit's contribution) is the same as in other words containing it. For instance, it seems clear that BOLDLY contains BOLD as a morpheme (in that it makes a transparent semantic contribution), whereas HARDLY doesn't contain HARD (and so HARDLY, despite the similarity in form, would not be decomposed). And, indeed, in a visually primed lexical decision experiment, BOLD primes BOLDLY but HARD doesn't prime HARDLY (relative to nonmorphological orthographic and semantic controls; Gonnerman, Seidenberg, & Anderson, 2007). But what about LATE in LATELY? On the localist theory, LATELY should behave either like BOLDLY if it is decomposed, or HARDLY if it's not, but empirically it exhibits an intermediate level of priming (Gonnerman et al., 2007). This finding is awkward for any theory that has no way to express intermediate degrees of morphological relatedness.

How might morphological structure be understood on a distributed connectionist approach? The first thing to point out is that morphemes, like word units, have no relevant internal structure but are posited to solve a particular problem: how to relate the surface forms of words to their meanings. We assume that (phonological) surface forms are coded by distributed patterns of activity over a group of units such that words with similar pronunciations are coded by similar patterns, and word meanings are coded over a separate group of units whose patterns capture semantic similarity. Mapping from one to the other is difficult precisely because, apart from morphological structure (and rare pockets of sound symbolism), phonological similarity is essentially unrelated to semantic similarity. This type of *arbitrary* mapping is particularly difficult for a connectionist network to learn, because units—due to their limited nonlinearity—are intrinsically biased to map similar inputs to similar outputs. In fact, when output similarity is very different from input similarity, the mapping cannot be implemented by direct connections between input and output units, and an additional layer of so-called *hidden* units are needed to mediate between the input and output. By modifying the input-to-hidden weights, the network can learn to re-represent the input patterns as a new set of patterns over the hidden units whose similarities are sufficiently close to those of the output patterns

that the hidden-to-output weights can generate the correct outputs. In this way, networks learn hidden representations that have a similarity structure that is in some sense halfway between the structure of the inputs and the structure of the outputs. This can always be done with a large enough hidden layer, but sometimes it is more efficient to use a series of smaller hidden layers instead.

Of course, spoken word comprehension is not a completely arbitrary mapping precisely because many words have morphological structure. On a connectionist account, however, the nature of this structure is not stipulated in advance (e.g., that words are composed of discrete parts) but is something that manifests in the statistical relationship between inputs and outputs and thus is discovered by the network in the course of learning.

Morphological structure introduces a degree of *componentiality* between inputs and outputs—that is, the degree to which parts of the input can be processed independently from the rest of the input. From a connectionist perspective, the notion of "morpheme" is an inherently graded concept because the extent to which a particular part of the phonological input behaves independently of the rest of the input is always a matter of degree (Bybee, 1985). Also note that the relevant parts of the input need not be contiguous, as in prefixes and suffixes in concatenative systems like English. Even noncontiguous subsets of the input, such as roots and word patterns in Hebrew, can function morphologically if they behave systematically with respect to meaning or syntax.

A network comes to exhibit degrees of componentiality in its behaviour because, on the basis of exposure to examples of inputs and outputs from a task, it must determine not only what aspects of each input are important for generating the correct output, but also what aspects are uninformative and should be ignored. This knowledge can then apply across large classes of items, only within small subclasses, or even be restricted to individual items. In this way, the network learns to map parts of the input to parts of the output in a way that is as independent as possible from how the remaining parts of the input are mapped. This provides a type of combinatorial generalisation by allowing novel recombinations of familiar parts to be processed effectively. In short, a network can develop mostly componential representations that handle the more systematic aspects of the task and that generalise to novel forms, while simultaneously developing

less componential representations for handling the more idiosyncratic aspects of the task, as well as the full range of gradations in between.

The graded componential structure of hidden representations is illustrated in a clear way by a simulation of morphological priming carried out by Plaut and Gonnerman (2000). A three-layer network was trained to map from the surface forms of words to their meanings for either of two artificial vocabularies (see Figure 1a). In each, sets of two-syllable words were assigned semantic features such that they varied in their semantic transparency. Each syllable was assigned a particular set of semantic features, such that a *transparent* word's meaning was simply the union of the features of it's component syllables. Such meanings are fully componential in that each semantic feature could be determined by one of the syllables without regard to the other. The meaning of an *intermediate* word was derived by determing the transparent meaning of its syllables and then changing a random third of its semantic features; for a *distant* word, two-thirds of the transparent features were changed. These meanings are progressively less componential than transparent meanings because the changed semantic features can be determined only from both syllables together. Finally, the meaning of an *opaque* word was derived by regenerating an entirely new arbitrary set of semantic features that were unrelated to the transparent meanings of its syllables.

Using these procedures for generating representations, two languages were created containing 1200 words each. In the morphologically *rich* language, the first 60 “stems” (first syllables), forming 720 words, were all transparent; in the *impoverished* language, they were all opaque. The remaining 480 words were identical across the two languages and were formed from 10 transparent stems, 10 intermediate stems, 10 distant stems, and 10 opaque stems. The simulation was designed to evaluate the degree of morphological priming among this shared set of words as a function of the nature of the remaining words in each of the two languages.

Figure 1b shows the amount of priming (difference in settling times following related vs. unrelated primes) as a function of level of morphological transparency and of language. The main relevant finding for present purposes is that, in both languages, morphological priming varies in a graded fashion as a function of semantic transparency, analogous to what was observed empirically by Gonnerman et al. (2006). The strong priming exhibited by transparent words suggests that the network's internal representa-

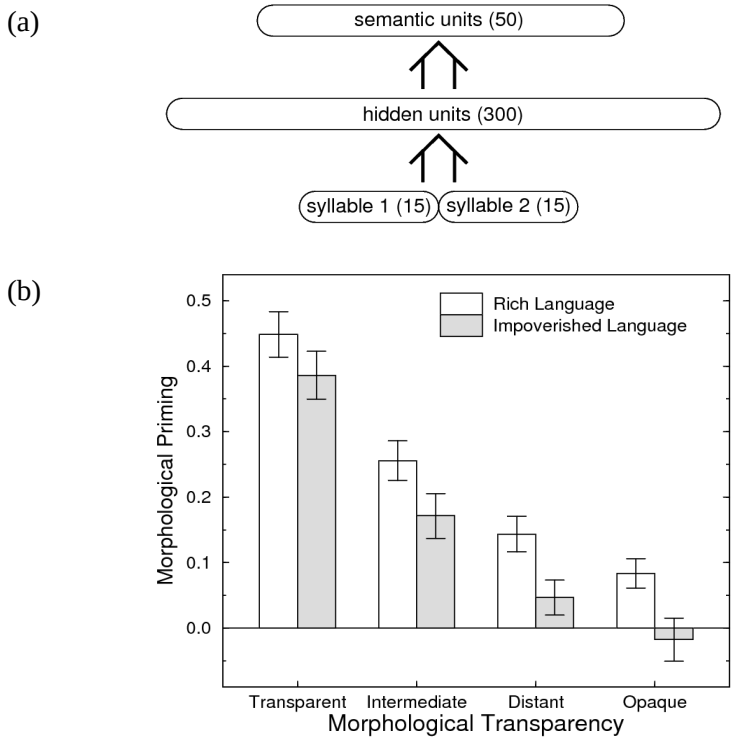

Figure 1. (a) The network architecture used by Plaut and Gonnerman (2000). Numbers of units in each group are shown in parentheses, and large arrows represent full connectivity between groups. (b) Priming results produced by the network as a function of the degree of morphological transparency and whether the network was trained on a morphologically rich or impoverished artificial language (Adapted from Plaut and Gonnerman, 2000).

tions have learned the systematic relationship between the shared stem's surface form and its (transparent) meaning, and in this sense it seems natural to describe the stem as a "morpheme" that is shared by the prime and target. But the intermediate and distant words benefit from sharing a stem to less of an extent, due to the fact that their internal representations overlap less. In these cases, what the stem contributes to the representation of the prime is not contained in or part of the representation of the target; rather, there is some degree of overlap but also some degree of divergence between the stem's contribution in the two words. At best, what could be said is that

the stem functions as a morpheme to some degree, and is contained by words to some degree; there is no discrete point at which words go from being fully componential to fully opaque. And based on the empirical findings, this characterization of graded morphological structure applies to human subjects as well as to the network.

In summary, unstructured or localist word representations can be augmented with similar morpheme representations to capture some aspects of the internal structure of words, but the processing of words with intermediate degrees of semantic transparency is awkward to explain. By contrast, because distributed connectionist networks start with the assumption that entities such as words are represented by patterns of activity with rich internal structure, such networks can more naturally capture the graded relationships between the surface forms of words and their meanings.

4. The challenge of external context: Ambiguity

Capturing the internal structure of words is not the only challenge facing theories of lexical representation. Another, often neglected problem concerns ambiguity in the relationships among different aspects of lexical knowledge. As it turns out, addressing this issue requires coming to terms with how words contribute to, and are influenced by, higher levels of language processing.

Every aspect of lexical knowledge suffers from ambiguity when words are considered in isolation: semantics (e.g., BANK [river] vs. BANK [money]), syntax (e.g., [the] FLY vs. [to] FLY); phonology (e.g., WIND [air] vs. WIND [watch]), and even orthography (e.g., COLOUR vs. COLOR). Most computational models of lexical processing, including connectionist ones, either actively avoid this problem by adopting simplified vocabularies and representations that lack ambiguity (e.g., Kello & Plaut, 1993), or perhaps include it only in phonology (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Harm & Seidenberg, 2004; Plaut et al., 1996). In simulations that include semantics, the presentation of a homophone like ATE/EIGHT in phonology, or a heterophonic homograph like WIND in orthography, typically gives rise to a blend of the semantic features of the relevant meanings, although such blending can be reduced by the introduction of disambiguat-

ing information, such as distinctive semantic for homophones or phonological information for homographs (see, e.g., Harm & Seidenberg, 2004).

A similar situation arises in simulations that include semantic ambiguity; that is, in which a given surface form (e.g., BANK) corresponds to more than one semantic representation (e.g., Joordens & Besner, 1994), although blends can be prevented for the most part by the use of an appropriate learning procedure (Movellan & McClelland, 1993; Rodd, Gaskell, & Marslen-Wilson, 2004). The selection of which of its multiple meanings a word produces on a given occasion is influenced by the relative frequency of the meanings but is otherwise the result of random processes within the network. This may suffice when accounting for data from words presented in isolation and in random order, but does not generalize to the way in which ambiguous words are understood in context.

Armstrong and Plaut (2008) developed a simple simulation of the use of context to disambiguate semantically ambiguous words, including both *homonymy* (i.e., words such as BANK [river/money] with multiple distinct meanings) and *polysemy* (i.e., words such as PAPER [document/material] with multiple distinct senses with a common meaning). Although these relations are often dichotomized in experimental designs, the degree of pattern overlap among distributed semantic representations provides a natural means of capturing the full continuum of relatedness among word meanings. The target phenomena for the simulation were findings by Hino, Pexman and Lupker (2006) that lexical decision typically produces only a polysemy advantage (i.e., faster responding to polysemous vs. unambiguous words) whereas semantic categorization produces only a homonym disadvantage (i.e., slower responding to homonymous vs. unambiguous words). Armstrong and Plaut's goal was to account for these findings, not in terms of task differences, but in terms of the time-course of cooperative and competitive dynamics within a recurrent connectionist network.

The architecture of the network included 25 orthographic units connected to 150 hidden units, which in turn where bidirectionally connected to 100 semantic units. In addition, 75 “context” units provided additional input to the hidden units that served as the basis for disambiguating words. The training patterns consisted of 128 unambiguous words, 64 homonymous words, and 64 polysemous words. Artificial patterns were generated to approximate the relationship among written words and their meanings. Specifically, orthographic, context, and semantic representations were gen-

erated by probabilistically activating a randomly selected 15% of the units in a group (ensuring that all patterns differ by at least three units). Unambiguous words consisted of a single pairing of a randomly selected orthographic pattern, context pattern, and semantic pattern. Homonymous words were represented as two separate input patterns which shared the same orthographic pattern but were associated with a different randomly selected context and semantic pattern. Polysemous words were similar except that their semantic patterns were both originally derived by distoring the same "prototype" pattern to ensure that they shared 60% of their features with each other. To instantiate the bottom-up salience of orthographic stimuli, context input was presented only after 10 unit updates with orthographic input alone.

After training with a continuous version of recurrent back-propagation, the network was successful and activating the correct semantic features of each word given the appropriate context representation. Figure 2 shows the number of semantic units in the model that were activated strongly (i.e., above 0.7) over the course of processing polysemous, homonymous, and unambiguous words. Early in semantic processing (time A), polysemous words show an advantage over both homonymous and unambiguous words (which do not differ much). This advantage arises because the shared features among the overlapping meanings mutually support each other. In contrast, late in processing (time C), homonymous words show a disadvantage relative to both polysemous and unambiguous words (which do not differ). This disadvantage is due to competition among the non-overlapping features of the alternative unrelated meanings of homonymous words. Thus, the network exhibits the pattern of results observed by Hino et al. (2006), not because of task differences (as there are none in the model), but because of changes in the dynamics among sets of semantic features in the model. The model accounts for the empirical data if we assume that lexical decisions can be made relatively early in the course of semantic processing, whereas semantic categorization requires a more precise semantic representation that takes longer to activate.

On this account, it should be possible to shift from a polysemy advantage to a homonymy disadvantage within a single task solely by increasing difficulty (and thus degree of semantic processing). Armstrong and Plaut (2008) tested and confirmed this prediction by varying the wordlikeness (summed bigram frequency) of nonword foils in a lexical decision task.

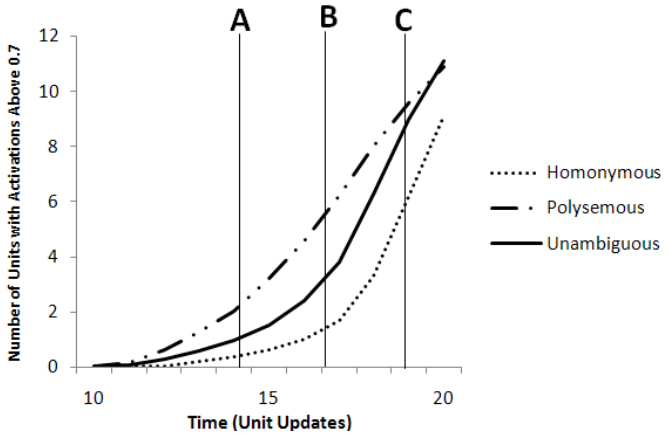

Figure 2. The average number of semantic units in the Armstrong and Plaut (2008) model that were active above 0.7 for polysemous, unambiguous, and homonymous words. Note that these trajectories do not reflect pre-semantic visual and orthographic processing; the zero time-point reflects the onset of semantic processing only, and no semantic units were active above 0.7 before unit update 10. (Adapted from Armstrong and Plaut, 2008)

Moreover, by using moderately wordlike nonwords, they confirmed the model's that, with an intermediate amount of semantic processing (time B), both effects should be observed (see Figure 2).

The Armstrong and Plaut (2008) model illustrates—in admittedly highly oversimplified form—how context can serve to disambiguate words of varying degress of ambiguity in a way that is consistent with at least some aspects of human comprehension processes (see also Gaskell & Marslen-Wilson, 1997). But in many ways the model begs the question of where the appropriate context representations come from in the first place. One possible answer is that the network activation left behind by the previous word might serve as the relevant context. However, while some models have used this approach to model lexical semantic priming effectively (e.g., Plaut & Booth, 2000), the meaning of a single word is insufficient in general to capture the richness and complexity of how previous (and even subsequent) linguistic input can serve to alter the meaning of a word. A full treatment of context effects on word comprehension requires embedding lexical processing within a broader framework for sentence understanding.

As an example of how sentence-level syntax and semantics must be used to determine word meanings, consider the following:

1. The pitcher threw the ball.

Here, every content word has multiple meanings in isolation but an unambiguous meaning in context. The same is true of vague or generic words, such as CONTAINER, which can refer to very different types of objects in different contexts, as in

2. The container held the apples.
3. The container held the cola.

Finally, at the extreme end of context dependence are implied constituents which are not even mentioned in the sentence but nonetheless are an important aspect of its meaning. For example, from

4. The boy spread the jelly on the bread.

most people infer that the instrument was a knife.

To address how sentence context can inform word comprehension (among other issues), St. John & McClelland (1990; McClelland, St. John, & Taraban, 1989) developed a connectionist model of sentence comprehension which instantiates sentence comprehension as a constraint satisfaction process in which multiple sources of information from both syntax and semantics are simultaneously brought to bear in constructing the most plausible interpretation of a given utterance. The architecture of the model, in the form of a simple recurrent network, is shown in Figure 3. The task of the network was to take as input a single-clause sentence as a sequence of constituents (e.g., THE-BUSDRIVER ATE THE-STEAK WITH-A-KNIFE) and to derive an internal representation of the event described by the sentence, termed the Sentence Gestalt. Critically, this representation was not predefined but was learned from feedback on its ability to generate appropriate thematic role assignments for the event given either a role (e.g., Agent, Patient, Instrument) or a constituent that fills a role (e.g., busdriver, steak, knife) as a probe.

Events were organized around actions and had a probabilistic structure. Specifically, each of 14 actions had a specified set of thematic roles, each of which was filled probabilistically by one of the possible constituents. In this process, the selection of fillers for certain roles biased the selection for other roles. For example, for eating events, the busdriver most often ate steak

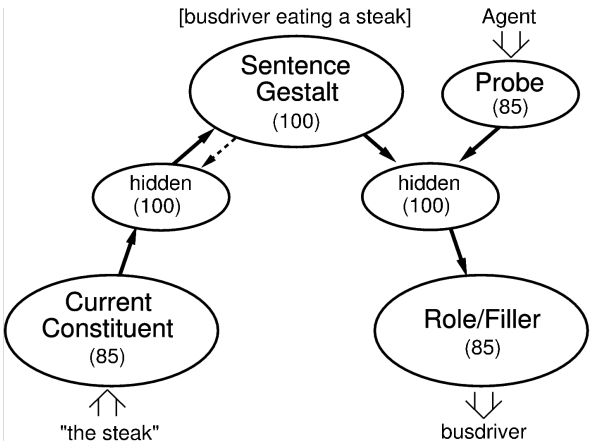

Figure 3. The architecture of the St. John and McClelland (1990) model of sentence comprehension. The number of units in each layer is shown in parentheses. The large arrows identify which layers receive input (incoming arrow) or produce output (outgoing arrow). The dashed arrow indicates a projection from "context" units (omitted for clarity) whose states are copied from the Sentence Gestalt layer for the previous time step. The indicated content of representations is midway through the sentence THE BUSDRIVER ATE THE STEAK WITH A KNIFE. (Adapted from St. John & McClelland, 1990).

whereas the teacher most often ate soup, although occasionally the reverse occurred. These probabilistic biases in the construction of events were intended to approximate the variable but non-random structure of realworld events: some things are more likely than others to play certain roles in certain activities.

The choice of words in the construction of a sentence describing the event was also probabilistic. The event of a busdriver eating a steak with a knife might be rendered as THE-ADULT ATE THE-FOOD WITH-A-UTENSIL, THE-STEAK WAS-CONSUMED-BY THE-PERSON, SOMEONE ATE SOMETHING, and so on. This variability captures the fact that, in real life, the same event may be described in many different ways and yet understood similarly. Overall, given the probabilistic event structures and the lexical and syntactic options for describing events as sentences, there were a total of 120 different events (of which some were much more likely than others) and 22,645 different sentence-event pairs.

During training, sentence-event pairs were generated successively and the constituents of each sentence were presented one at a time over the Current Constituent units (see Figure 3). For each constituent, the network updated its Sentence Gestalt representation and then attempted to use this representation as input to generate the full set of role/filler pairs for the event. Specifically, with the Sentence Gestalt fixed and given either a role or a filler over the Probe units, the network had to generate the other element of the pair over the Role/Filler units. For example, after the presentation of THE-STEAK in the sentence THE-STEAK WAS-EATEN-BY THE-BUSDRIVER, the network was trained to output, among other things, the agent (busdriver), the patient (steak), the action (eating), and the instrument (fork). It was, of course, impossible for the network to do this with complete accuracy, as these role assignments depend on constituents that have yet to occur or are only implied. Even so, the network could do better than chance; it could attempt to predict missing information based on its experience with the probabilistic dependencies in the event structures. More specifically, it could (and did) generate distributions of activity over roles and fillers that approximated their frequency of occurrence over all possible events described by sentences that start with the-steak. Note that these distributions could, in many cases, be strongly biased towards the correct responses. For example, steaks typically fill the patient role in events about eating and (in the environment of the network) steaks are most commonly eaten by busdrivers using a fork. In this way, the training procedure encouraged the network to extract as much information as possible as early as possible, in keeping with the principle of immediate update (Marslen-Wilson & Tyler, 1980). Of course, the network also had to learn to revise the Sentence Gestalt appropriately in cases where its predictions were violated, as in THE-STEAK WAS-EATEN-BY THE-TEACHER.

The network was trained on a total of 630,000 sentence-event pairs, in which some pairs occurred frequently and others—particularly those with atypical role assignments—were very rare. By the end of training, when tested on 55 randomly generated sentence-event pairs with unambiguous interpretations, the network was 99.4% correct.

St. John and McClelland (1990) carried out a number of specific analyses intended to establish that the network could handle more subtle aspects of sentence comprehension. In general, the network succeeded at using both semantic and syntactic context to 1) disambiguate word meanings

(e.g., for THE-PITCHER HIT THE-BAT WITH-THE-BAT, assigning flying bat as patient and baseball bat as instrument); 2) instantiate vague words (e.g., for THE-TEACHER KISSED SOMEONE, activating a male of unknown age as patient), and 3) elaborate implied roles (e.g., for THE-TEACHER ATE THE-SOUP, activating spoon as the instrument; for THE-SCHOOLGIRL ATE), activating a range of foods as possible patients).

Disambiguation requires the competition and cooperation of constraints from both the word and its context. While the word itself cues two different interpretations, the context fits only one. In THE-PITCHER HIT THE-BAT WITH-THE-BAT, PITCHER cues both container and ball-player. The context cues both ball-player and busdriver because the model has seen sentences involving both people hitting bats. All the constraints supporting ball-player combine, and together they win the competition for the interpretation of the sentence. In this way, even when several words of a sentence are ambiguous, the event which they support in common dominates the disparate events that they each support individually. The processing of both instances of BAT work similarly: the word and the context mutually support the correct interpretation. Consequently, the final interpretation of each word fits together into a globally consistent understanding of an entire coherent event.

There is no question that the Sentence Gestalt model has important limitations in its theoretical scope and empirical adequacy. The model was trained on sentences restricted to single clauses without embeddings and pre-parsed into syntactic constituents, and the use of event structures composed of probabilistic assignment to fixed thematic roles was also highly simplified (although see Rohde, 2002, for an extension of the model that addresses these limitations). Nonetheless, it is useful to consider the nature of word meanings, and lexical representations more generally, in light of the operation of the model.

The first thing to note is that there is no real sense in which each word/constituent[2] in the input is assigned a particular semantic representation—in the form of a pattern of activity over a group of units—even when disambiguated by context. Rather, the current word combines with the current context—coded in terms of the existing activation pattern within the network—to determine a new internal representation (over the hidden units) that then serves to revise the model's sentence interpretation (over the Sentence Gestalt layer). While it is true that the contribution of the current

word is carried out via a relatively stable set of weights—those coming out of the unit (or units) coding it as input—the actual impact of this knowledge on active representations within the model is strongly dependent on context. This dependence can vary from introducing subtle shading (for polysemous words) to selection of an entirely distinct interpretation (for homonymous words), and everything in between. In this way, in the context of the model, it would be a mistake to think of words as "having" one or more meanings; rather, words serve as "cues" to sentence meaning—for some words, the resulting sentence meanings have considerable similarity whereas for others, they can be quite unrelated.

In the context of a typical psycholinguistic experiment, where words are presented in isolation and in a random order, the representation of "sentence context" is generally unrelated and unbiased relative to the contexts that a word typically occurs in, and so the resulting representation evoked over the Sentence Gestalt layer reflects general implications of a word across all of its context—in some ways analogous to what happens in the model for the initial word/constituent of a sentence. Such a pattern may be systematically related to other types of knowledge (e.g., pronunciation) but it wouldn't constitute a specific part of some larger lexical representation. In the model, and perhaps in the human language system as well, words are not assigned specific representations but solely serve as vehicles for influencing higher-level linguistic representations. It is in this sense that, as claimed at the outset of this chapter, distributed connectionist modelling gives rise to a view of language in which lexical knowledge and processing play a fundamental role in language understanding, without any explicit role for *lexical representation* per se.

5. Conclusions

Despite broad agreement on the critical roles that words play in language, there is very little clarity on the nature of word representations and how they interact with other levels of representation to support linguistic performance. Early theories of lexical representation used words as unstructured "handles" or pointers that simply linked together and provided access to phonological, orthographic, semantic, and grammatical knowledge. However, such a simple account is undermined by careful consideration of both

the effects of the internal structure of words and of the subtleties in how words are influence by the contexts in which they occur.

Distributed connectionist modeling provides a way of learning the functional relationships among different types of information without having to posit an explicit, discrete data structure for each word (or morpheme). Rather, the similarity structure of activation patterns within and between each domain can capture various aspects of morphological relatedness, and an emerging sentence-level interpretation can modulate the contributions that words make to meaning. Indeed, if the goal of language processing is cast as the comprehension and production of larger-scale utterances, individual words can be seen as contributing to these processes in context-sensitive ways without themselves being represented explicitly. Although the resulting theory of language processing runs against strong intuitions about the primacy of lexical representation in language, it might nonetheless provide the best account of actual language performance.

Notes

1. The many-to-one case, where many units code one and only one entity, is essentially a redundant version of a localist code. The one-to-many case, where entities correspond to single units but a given unit represents multiple entities, is too ambiguous to be useful.
2. Although St. John and McClelland's (1990) Sentence Gestalt model took constituents rather than words as input (e.g., THE-BUSDRIVER), Rohde's (2002) extension of the model took sequences of individual words as input.

References

Armstrong, Blair C., and David C. Plaut

2008 Settling dynamics in distributed networks explain task differences in semantic ambiguity effects: Computational and behavioral evidence. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bowers, Jeffrey S.

2009 On the biological plausibilty of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review, 116*: 220-251.

Bybee, Joan
1985 *Morphology: A study of the relation between meaning and form.* Philadelphia: Benjamins.

Coltheart, Max, Kathleen Rastle, Conrad Perry, Robyn Langdon, Johannes Ziegler
2001 DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*: 204-256.

Gaskell, M. Gareth, and William D. Marslen-Wilson
1997 Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12*: 613-656.

Harm, Michael W. and Mark S. Seidenberg
2004 Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review, 111:* 662-720.

Hino, Yashushi, Penny M. Pexman, and Stephen J. Lupker
2006 Ambiguity and relatedness effects in semantic tasks: Are they due to semantic coding? *Journal of Memory and Language, 55*: 247-273.

Joordens, Steve, and Derek Besner
1994 When banking on meaning is not (yet) money in the bank: Explorations in connectionist modeling. *Journal of Experimental Psychology: Learning Memory and Cognition,* 20: 1051-1062.

Kello, Christopher T., and David C. Plaut
2003 Strategic control over rate of processing in word reading: A computational investigation. *Journal of Memory and Language, 48: 2*07-232.

Marslen-Wilson, William D., and Lorraine K. Tyler
1980 The temporal structure of spoken language understanding. *Cognition, 8*: 1-71.

McClelland, James, L., Brian L. McNaughton, and Randall 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, James L., and David E. Rumelhart
1982 An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*: 375-407.

McClelland, James L., Mark St. John, and Roman Taraban
1989 Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes, 4*: 287-335.

Morton, John

1969 The interaction of information in word recognition. *Psychological Review, 76*: 165-170.

Movellan, Javier R. and James L. McClelland

1993 Learning continuous probability distributions with symmetric diffusion networks. *Cognitive Science, 17*: 463-496.

O'Reilly, Randall C., and James L. McClelland

1994 Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *Hippocampus, 6:* 661-682.

O'Reilly, Randall C., and James L. McClelland

1994 Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *Hippocampus, 6:* 661-682.

Plaut, David C., and James R. Booth

2000 Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review, 107*: 786-823.

Plaut, David C., and James L. McClelland

2010 Locating object knowledge in the brain: A critique of Bowers's (2009) attempt to revive the grandmother cell hypothesis. *Psychological Review, 117*: 284-290.

Plaut, David C., James L. McClelland, Mark S. Seidenberg, and Karalyn Patterson

1996 Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103:* 56-115.

Rodd, Jennifer M., M. Gareth Gaskell, and William D. Marslen-Wilson

2004 Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science, 28*: 89-104.

Rohde, Douglas L. T.

2002 *A connectionist model of sentence comprehension and production.* Ph.D. dissertation, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA. Available as Technical Report CMU-CS-02-105.

St. John, Mark F., and James L. McClelland

1990 Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46*: 217-257.

Taft, Marcus

2006 A localist-cum-distributed (LCD) framework for lexical processing. In *From inkmarks to ideas: Current issues in lexical processing*, Sally Andrews (ed.), 76-94. Hove, UK: Psychology Press.