Bridging language with the rest of cognition: computational, algorithmic and neurobiological issues and methods

Shimon Edelman
Department of Psychology
232 Uris Hall, Cornell University
Ithaca, NY 14853-7601, USA
http://kybele.psych.cornell.edu/∼edelman

August 31, 2004

Abstract

The computational program for theoretical neuroscience initiated by Marr and Poggio (1977) calls for a study of biological information processing on several distinct levels of abstraction. At each of these levels — computational (defining the problems and considering possible solutions), algorithmic (specifying the sequence of operations leading to a solution) and implementational — significant progress has been made in the understanding of cognition. In the past three decades, computational principles have been discovered that are common to a wide range of functions in perception (vision, hearing, olfaction) and action (motor control). More recently, these principles have been applied to the analysis of cognitive tasks that require dealing with structured information, such as visual scene understanding and analogical reasoning. Insofar as language relies on cognition-general principles and mechanisms, it should be possible to capitalize on the recent advances in the computational study of cognition by extending its methods to linguistics.

The possibility of integrating linguistics into a unified science of cognition — a desideratum put forward in many of the relevant disciplines — depends on the degree to which common computational principles (Marr and Poggio, 1977) and brain mechanisms are shared by language and by the other cognitive functions. To explore this possibility, we need to bring together ideas from several fields, which as yet have seen little intellectual cross-fertilization. The first of these is cognitive linguistics (Langacker, 1987; Bernárdez, 1999) — a natural home discipline for the integration project, which consistently produces valuable insights into the psychology of language, yet is little concerned with algorithmic or implementational issues. The second is computational linguistics (Jurafsky and Martin, 2000), including statistical natural language processing

(Manning and Schütze, 1999) — a field that examines the mathematical nature of language-related tasks and generates important applications, yet pays little attention to behavioral or neurobiological issues. Lastly, there is the Marr-Poggio computational framework (Marr and Poggio, 1977), which is used across cognition and which spans all the relevant levels of analysis, but has not yet been extended to the study of language.

This chapter discusses some of the general computational principles that emerge as useful for understanding cognition, focusing on those that are likely to be especially relevant in dealing with structured knowledge. It then brings these principles to bear on a theory of language that is rooted both in cognitive and in computational linguistics, and that views language as an incrementally learnable system of redundant, distributed representations akin to those found by neurobiologists in olfaction, audition, vision, and motor control.

1 Common principles of cognitive representation and processing

The view that cognition hinges on the representation of knowledge by the brain is widely accepted in linguistics, psychology, neuroscience, and the philosophy of mind (Chomsky, 1957; Miller, 1962; Shepard, 1975; Marr, 1982; Gallistel, 1990; Cummins, 1996). Most importantly, representations play a central role in those theories of mind/brain that construe cognition as computation defined over representational states (Baum, 2004).[1] A representational state in a cognitive system is characterized by its covariation with certain aspects of the relevant state of affairs in the world, and, crucially, by having counterfactually supported observable effects.[2]

1.1 How to garner empirical support for posited representations

Whereas thirty years ago linguists were expected to prove that the representations they posit are psychologically real (Fodor et al., 1974) by predicting and then demonstrating such effects, contemporary formal linguistics has, lamentably, given up on this requirement (Edelman and Christiansen, 2003). Consider, for example, the following passage from an online introduction to a course in neurolinguistics: "We know already what isn't the right question: What is the psychological reality of linguistic entities and operations?"[3] As

[1] Such states need not, and probably cannot, be wholly internal to the brain; cf. "The primary function of perception is to keep our internal framework in good registration with that vast external memory, the external environment itself" (Reitman et al., 1978, p.72). Thus, in many respects, the world is its own best representation (O'Regan, 1992).

[2] A counterfactual is a logical conditional statement whose antecedent is taken to be contrary to fact by those who utter it; cf. "If linguistics were what the author claims, syntactic trees would be visible in CAT scans." (Postal, 2004, ch.11).

[3] Quote found at a web site for *Neurolinguistics*, course #24.944, taught by Alec Marantz (Head, Department of Linguistics and Philosophy, MIT) in 2000; see
http://web.mit.edu/linguistics/www/marantz/marantz.home.24944f00intro.html.

a result, stipulated linguistic *entia*, from Deep Structure and transformations in the 1960s to Logical Form, traces, Move, and Merge in the 1990s, have multiplied over the decades, arguably *praeter necessitatem*.[4] To date, none of the rare attempts to obtain psycholinguistic (behavioral) evidence for such entities have yielded unequivocal results. For example, in the recent study by (Nakano et al., 2002), only 24 subjects out of the original 80 performed consistently with the predictions of a trace/movement theory, while 39 subjects exhibited the opposite behavior (the data from the rest of the subjects were discarded because their error rate was too high).

The shakiness of the empirical foundations of generative linguistics appears to be especially disappointing when seen in the broader context of successful representational theories that have emerged in other cognitive domains. Indeed, the need to demonstrate the psychological (behavioral), and, eventually, the neurobiological, reality of the theoretical constructs exists in all of cognition, including human vision, where, as in language, direct observation of the underlying mechanisms is difficult. An excellent example of how vision scientists have risen to this challenge is found in the history of the concept of multiple parallel spatial frequency channels (Figure 1), a representational hypothesis that had been introduced in the late 1960s, then completely vindicated by purely behavioral means over the following decade; see, e.g., (Wilson and Bergen, 1979).

More generally, the logic of looking for empirical signatures of the posited representations can be put to work in a number of ways that are all well known to cognitive scientists. For example, the reality of a distinction between two representations or processes can be indicated by a *double dissociation*, that is, a situation in which each of the two can be obtained in isolation from the other, either as a result of "complementary" lesions in different patients (Damasio and Tranel, 1993), or through experimental manipulation of stimuli presented to normal subjects (Pulvermüller et al., 1996). Likewise, one can use *priming* (Tulving and Schacter, 1990; Ochsner et al., 1994): if a representation can be primed — that is, if its manifestation in response to a stimulus can be modified by prior exposure to a related stimulus — then it is real, and can be accounted for by the mechanism that embodies the memory trace for this class of stimuli (Wiggs and Martin, 1998). Finally, the worries of some cognitive scientists that representations are merely epiphenomenal to cognition can be assuaged ultimately by demonstrating the *causal effectiveness* of representational mechanisms through direct intervention, such as microampere-level current injection at the appropriate brain site which brings about the predicted perceptual/behavioral change (Salzman et al., 1990).

[4] Occam's Razor, often stated as *entia non sunt multiplicanda praeter necessitatem* [entities should not be multiplied beyond necessity], is a fundamental principle of the scientific method, which has recently assumed a central role in statistical inference and learning theory (Rissanen, 1987; Blumer et al., 1987).

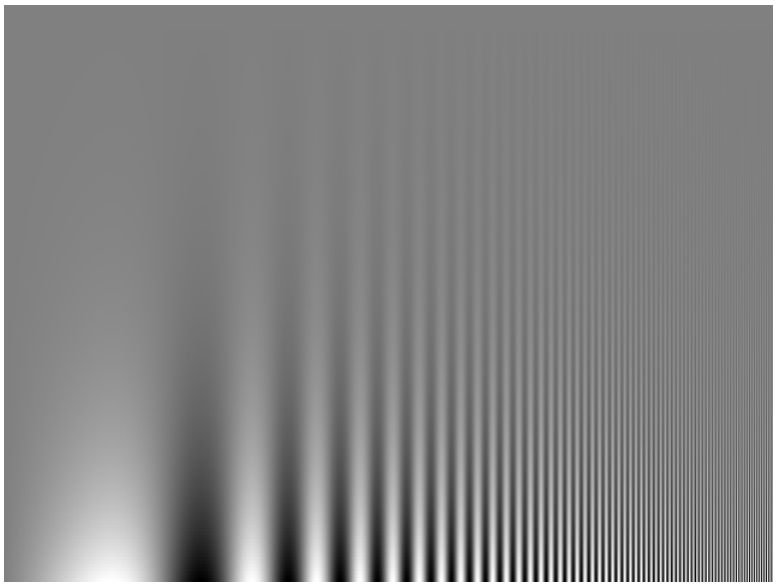

Figure 1: An illustration of the basic concepts required for a behavioral demonstration of the psychological reality of spatial frequency channels, which is a kind of visual representation found in the brain. In the image shown here, spatial frequency (rate of change of intensity across space) varies along the abscissa and contrast (difference between dark and light) along the ordinate (Campbell and Robson, 1968). As you can see, more contrast is required to perceive the grating (the alternation between dark and light) for low and high frequencies, compared to intermediate frequencies. Researchers have postulated early on that the perception of spatial frequency stimuli is supported by multiple channels, each tuned to a particular band (much like in a sound system's graphic equalizer). Evidence corroborating this idea comes from three kinds of psychophysical experiments. **(1) Adaptation:** exposure to a high-contrast grating of a specific frequency reduces the sensitivity (that is, raises the detection threshold) for gratings of other spatial frequencies to an extent depending on the frequency difference. In the context of the variable frequency and contrast test grating shown here, adaptation would manifest itself as a notch in the boundary along which the grating fades into an apparently uniform field, situated at the frequency of the adapting stimulus: at the adapted frequency, more contrast is needed for the grating to be perceived. **(2) Sub-threshold summation:** the extent to which two sub-threshold (that is, imperceptible) gratings of differing spatial frequencies shown together combine to elicit a supra-threshold percept depends on the difference between their frequencies. **(3) Masking:** the threshold for a faint test grating is elevated by a high-contrast mask grating superimposed on it to the extent that their frequencies match. In all three cases, the effects fade when the difference in spatial frequency is larger than about one octave (a factor of two). This would not be the case, were the contrast information not processed by a set of independent mechanisms, each tuned to an octave-wide band of spatial frequencies.

1.2 Some common characteristics of cognitive representations

What kinds of representations does one find in the brain? In domains as diverse as olfaction, vision, reasoning and memory, the representations are typically *distributed* in that an ensemble of neurons (rather than a single neuron) is involved in each task, *overlapping* in that the response profiles of the members of an ensemble are redundant (rather than mutually exclusive), and *graded* in that each neuron's response depends smoothly (rather than in an all-or-none fashion) on the represented quantity. Thus, theories involving distributed, overlapping, and graded representations (Pouget et al., 2000) have enjoyed the most consistent and wide-ranging explanatory success across cognition.

While thinking about distributed representations, it is important to realize that embracing the terminology of parallel distributed processing, often referred to as "connectionism", does not by itself constitute a particularly illuminating explanation. That the brain is, on the level of mechanism, a connectionist device is a trivial observation; understanding it will take coordinated action on computational (problem) and algorithmic (process) levels as well (Marr and Poggio, 1977). This is precisely what is happening now in the cognitive sciences: researchers are converging on a few classes of problems to which various aspects of cognition can be reduced, and are forging an understanding of a few classes of computational processes, operating over distributed representations, that are common to a wide range of cognitive tasks. Some examples of such general-purpose computational building blocks of biological information processing are outlined below (see Appendix A for a brief overview of the relevant mathematical concepts).

Function approximation. In numerical analysis, function approximation is the problem of recovering the form of an unknown function from a set of given argument-value pairs. It has been noted that this generic problem description fits well the standard scenario of supervised learning (Poggio, 1990). In vision, for instance, an observer may be exposed to various views of a given shape, then required to determine whether a test view belongs to the same object. This can be done by approximating an indicator function that encodes the appearance of the object in question in the space of all views of all possible objects, then evaluating it at the test view (Poggio and Edelman, 1990). In motor control, the function to be learned may map the intended action to a vector of muscle activations, and so on for other cognitive tasks (Poggio, 1990).

Density estimation. An idea that is computationally related to function approximation is the estimation of the probability density of some quantity of interest over the relevant variables. Such estimation may proceed in an unsupervised fashion, or combined with class information; it leads to powerful methods for statistical inference and decision making, via the mathematical apparatus of Bayes theory and related approaches. Statistical inference is widely acknowledged to be an indispensable tool for cognition, in areas ranging from vision

(Kersten and Schrater, 2000) to conceptual learning (Tenenbaum, 1999) and language acquisition (Clark, 2001), where the relevant domain may be the set of constructions acquired from a corpus, and the output of the inference procedure — a breakdown of the probabilities of each construction depending on the context.

Dimensionality reduction. Learning theorists know that for function approximation (in particular, density estimation) to work well it must be conducted over a low-dimensional domain: because the required number of data points grows exponentially with dimensionality, function approximation in high-dimensional spaces is intractable (Bellman, 1961). The problem of dimensionality reduction, also known in cognition as feature detection, is increasingly often seen as fundamental in language (Landauer and Dumais, 1997) and in vision (Intrator and Edelman, 1997).

In some cases, these abstract problems and the principles and algorithms used to address them have been mapped onto the function of the brain and its circuitry, resulting in explanatory models that span all three levels of Marr's program. For example, in olfaction the anatomy and the physiology of the pathway leading from the sensory epithelium to the glomeruli in the olfactory bulb (Lancet, 1991; Shepherd, 1992) can be seen as filtering data through a bank of radial basis functions (Poggio, 1990). This operation implements what is known to be a universal approximation algorithm (Hartman et al., 1990) that can be used in learning from examples (Poggio, 1990). The same algorithmic approach can support visual object recognition, as demonstrated by the Chorus of Prototypes model (Edelman, 1999), in which the stimulus is represented by its similarities to (processed) memory traces of past stimuli. Recent single-cell studies in the monkey found neurons that are broadly and redundantly tuned to particular object categories (Freedman et al., 2001) and that embody an ensemble representation of inter-object similarities that is veridical with respect to the distal stimuli (Op de Beeck et al., 2001). Both these findings had been predicted by the Chorus of Prototypes model (Edelman, 1998; Edelman, 1999).

2 Dealing with structure: a special challenge?

To be relevant to language, the computational principles behind these findings must be extended to situations that require highly structured representations. Recent work in various areas of cognition has been pursuing such an extension. For example, in complex analogy tasks a similarity-based model performs very well when the distributed representations it uses are made to reflect the structure of the input (Plate, 1995; Eliasmith and Thagard, 2001). Likewise, in vision the Chorus of Fragments model (derived from the Chorus of Prototypes), which aims at dealing with structured objects and scenes (Edelman and Intrator, 2003), is based on the twin principles of distributed representation by similarities (mentioned above) and of the use of visual space to

anchor the various shape fragments (Edelman, 2002), introduced next.

2.1 The role of space in representing structure

The idea that space should serve as a natural scaffolding for supporting structured representations, whose roots go back to the ancient mnemonic Method of Loci (Neisser, 1976, p.137), is stated forcefully in Wittgenstein's *Tractatus* (Wittgenstein, 1961, proposition 3.1431):

> The essential nature of the propositional sign becomes very clean when we imagine it made up of spatial objects (such as tables, chairs, books) instead of written signs. The mutual spatial position of these things then expresses the sense of the proposition.

In vision, sorting shape cues by their location in the visual field goes a long way toward solving the binding problem in the representation of object and scene structure (Edelman, 1999; Clark, 2000; Edelman, 2002; Edelman and Intrator, 2003). In particular, various components of a scene or an object need not be bound to each other in any special manner, as long as each of them is bound to its proper location in the visual space, merely by virtue of its appearance there.

In neurobiology, the spatial scaffolding approach to the representation of visual structure is consistent with the omnipresence in the monkey inferotemporal and prefrontal cortex of *what+where* neurons, which are both shape-tuned (signaling *what* is the stimulus), and location-selective (signaling *where* it appears) (Rao et al., 1997; Op de Beeck and Vogels, 2000). On a larger scale, the neural substrate of the perceptually defined external space may be the cortical surface itself, as indicated by the ubiquity of map-like representations (Gallistel, 1990) in vision (Ward et al., 2002), olfaction (Joerges et al., 1997), and audition (Shamma, 2001).

2.2 Spatial representations for language

Functional (problem-level) analogies between language and vision suggest various parallels between the manner in which structure is dealt with in these cognitive domains (Minsky, 1985). For instance, the treatment of a sentence with an embedded relative clause may be compared to the processing of a scene with occlusion (Figure 2, left; additional parallels are illustrated and discussed in Figure 2, right). In recent years space has been conjectured to play a central role in neurolinguistics. Consider, for example, the notion of iconicity in syntax (Simone, 1995): "...not only motor but also cognitive operations such as language, which do not appear to have any intrinsic spatial organization, are maintained in registration with spatial systems, and [...] this attention-requiring linkage confers a processing advantage" (Coslett, 1999). The iconicity hypothesis is intimately connected to Construction Grammar (Fillmore, 1985; Goldberg, 1995): linguistic freezes or

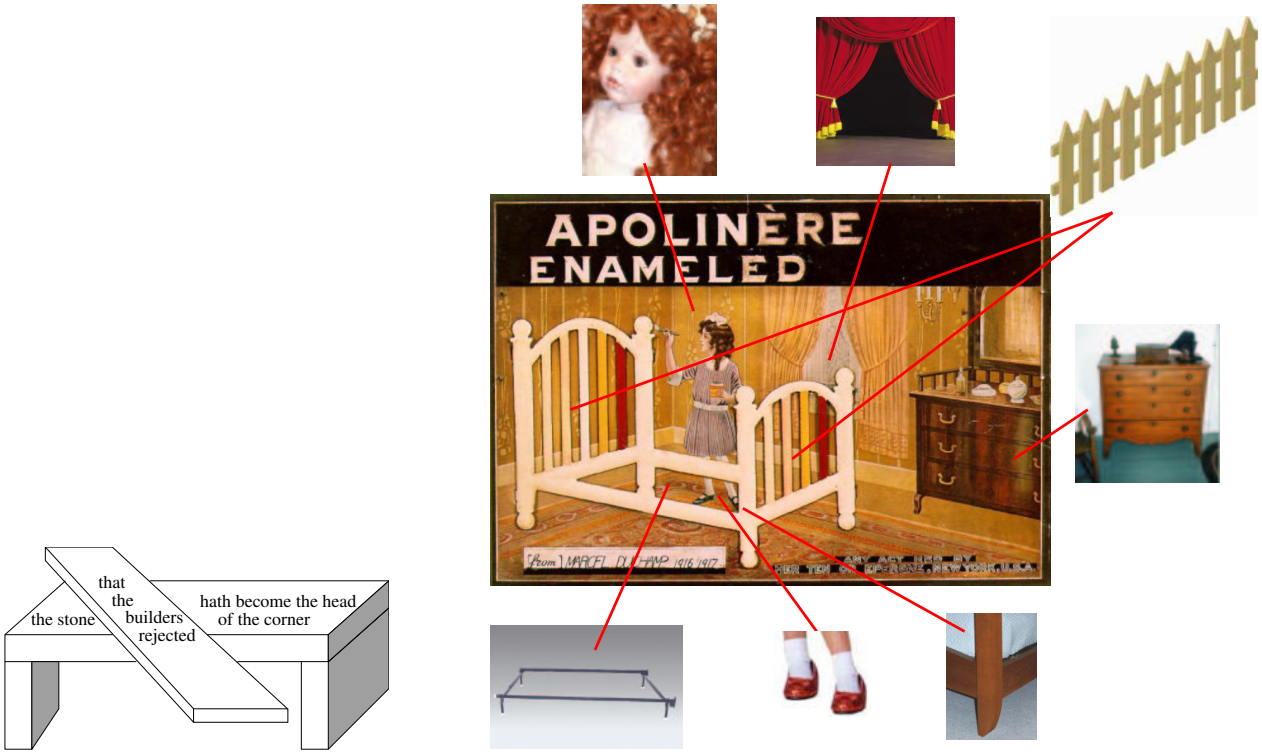

Figure 2: *Left:* there is a task-level analogy between interpreting a composite scene with occlusion and the processing of a sentence that contains an embedded relative clause (adapted from (Minsky, 1985, p.269)). The existence of such analogies between vision and language on the abstract level of the computational tasks (Marr and Poggio, 1977) faced by the brain, along with the uniformity of the underlying low-level cortical mechanisms (Phillips and Singer, 1997), suggests that cross-cognition commonalities should be sought also on the algorithmic level. *Right:* the postcard shown here (*Apolinère Enameled* by M. Duchamp) can be used to make the same point about parallels between language and vision (cf. the occlusion of the girl's legs by the bed-frame), and more. For instance, Duchamp's painting could be represented (and understood) in terms of its local similarities to various familiar images (Chorus of Fragments (Edelman, 2002; Edelman and Intrator, 2003)), which need not match the scene perfectly (Edelman, 1999); likewise, an utterance could be represented (and understood) in terms of the cloud of constructions (Chorus of Phrases (Solan et al., 2003b)) it evokes, as illustrated in Figure 3. Furthermore, just as many viewers fail to notice that the bed in this scene would be useless (look closely at the frame), subjects exposed to ungrammatical sentences of moderate complexity may rate them as no less felicitous than similarly structured grammatical ones. For example, the sentence "The apartment that the maid who the service had sent over was well decorated" tends to be rated as no worse (Gibson and Thomas, 1999) – and, in some settings, better (Christiansen and MacDonald, 2003) – than "The apartment that the maid who the service had sent over was cleaning every week was well decorated"; cf. (Gibson and Pearlmutter, 1998; Chipere, 1997; Chipere, 2001).

prefabs (Landsberg, 1995) that are spatially (or, equivalently, temporally; cf. Figure 3) iconic become constructions when parameterized (Erman and Warren, 2000). Psycholinguistic and neuropsychological evidence in support of linguistic iconicity has been recently reviewed in (Chatterjee, 2001).

On the level of neurobiology, in those areas of the human brain that support language, the counterpart to the visual *what+where* neurons mentioned above may be *what+when* neurons, tuned to particular structures appearing in a particular sequence (as illustrated in Figure 3). The possible role of temporal response properties of neuron assemblies in implementing sequence-sensitive processing has been discussed by (Pulvermüller, 2002); parallels between the brain representations of space and time in vision and in audition have been pointed out, among others, by (Shamma, 2001).

3 Treatment of structure in computational cognitive linguistics

Examples of space-based representations, which abound both in vision and in language, show that the goal of structure-sensitive processing by a distributed architecture is less forbidding than commonly thought, and that it is already within reach of cognition-general principles and mechanisms. This section outlines a cognitive approach to language, which is based on these foundations, and which casts the relevant computational, algorithmic and implementational issues in cognition-general terms.

3.1 Computational approach: the Chorus of Phrases and Construction Grammar

When applied to language, the idea of a distributed representation of structure based on similarities to multiple structured exemplars, called the Chorus of Fragments in the setting of visual scene processing (Edelman and Intrator, 2003), translates into a *Chorus of Phrases*: a redundant ensemble of potentially overlapping, mutually reinforcing phrase fragments that, as Langacker puts it, "motivate" the sentence they cover:

> "... rather than seeing a composite structure as an edifice constructed out of smaller components, we can treat it as a coherent structure in its own right: component structures are not the building blocks out of which it is assembled, but function instead to *motivate* various aspects of it." (Langacker, 1987, p.453), italics in the original.

A schematic illustration of the Chorus of Phrases (CoPh) in action is shown in Figure 3, where a stimulus (which could be an entirely novel sentence) evokes a cloud of associations, pointing to snippets of previously encountered phrases, each of which approximately matches parts of the input, and which together cover all of it.

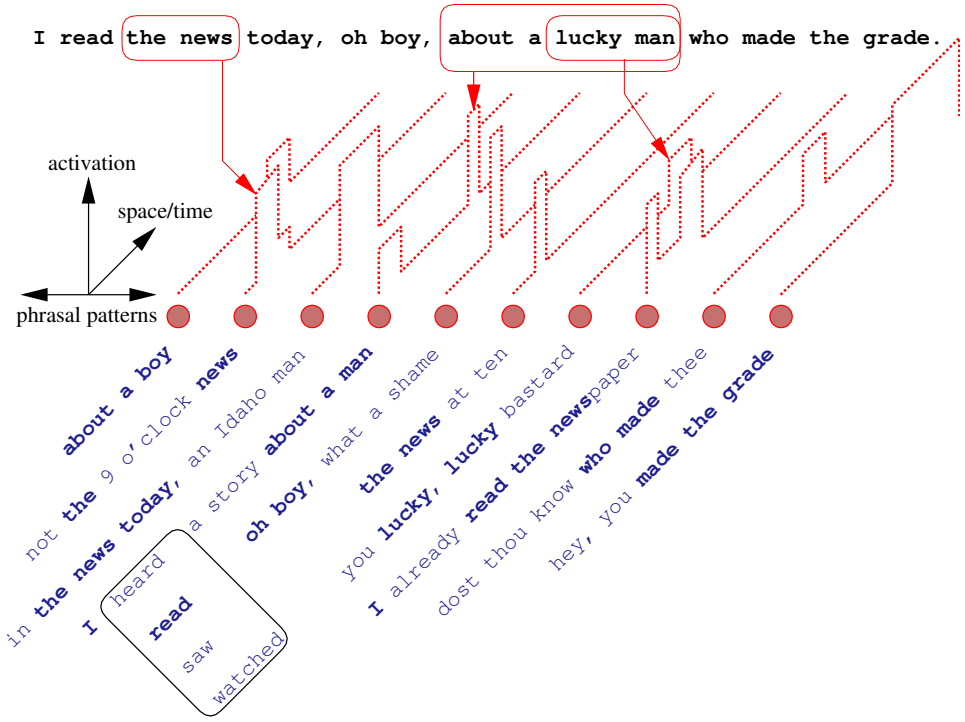

Figure 3: A schematic illustration of the Chorus of Phrases in sentence processing (for actual data from the ADIOS project, see (Solan et al., 2003a; Solan et al., 2003b; Edelman et al., 2004)). An input sentence is shown along with a subset of phrases it evokes from memory, each of which matches some word, sequence of words, or, generally, a parameterized pattern (in cartouche: **I** `heard, read, saw, watched` `a story` **about a man**) in the input. The unfolding of each pattern's activation (which reflects its time-domain "receptive field") may be important (Pulvermüller, 2002), but even without it the ensemble of active patterns is a highly informative representation, just as its counterparts in vision are (Edelman, 1999; Edelman and Intrator, 2003). The members of the ensemble disambiguate each other by supplying multiple interacting constraints on the interpretation. Consequently, it should be possible to process various queries about the input, both syntactic (voice, aspect, etc.) and semantic (thematic, connotational, conceptual). Moreover, it may be possible to use for that purpose generic cortical mechanisms (Phillips and Singer, 1997; Maass et al., 2003) that would map the distributed phrase activation patterns onto the corresponding required outputs, as in the scenario of function approximation found across cognition (Poggio, 1990; Intrator and Edelman, 1997). From the computational standpoint, it is interesting to observe that one can reconstruct the input sentence itself, should that be required for some reason, from a number of subsequence (phrase or pattern) queries that is on the order of $n \log \alpha + \alpha \log n$, where n is the length of the sentence and α is the size of the lexicon (Skiena and Sundaram, 1995). This computational complexity, which is quite benign in view of the α-fold parallelism inherent in a distributed lexicon, can be further reduced by allowing matching that is approximate (Jiang and Li, 1996) in the sense of (Valiant, 1984).

On the abstract, computational (Marr and Poggio, 1977) level, the view of language as based on context-specific structural generalizations, exemplified by the CoPh approach, differs radically from that of generative theories such as the Minimalist Program (Lasnik, 2002), which attempt to describe language in terms of universally valid syntactic rules projected by a categorically annotated lexicon. Recall that the basic theoretical challenge at the computational level is to specify what it is that needs to be done in the given task — in the present case, in language comprehension and production. According to the CoPh framework, comprehension involves constructing a distributed representation of the stimulus in terms of its structure-dependent similarities to multiple stored exemplars, which convey information both about form (the exemplars are, generally, patterns with slots; see Figure 3) and about meaning. Production consists of letting a set of exemplars chosen for their semantics interact and constrain each other until a fully specified linear sequence of terminals is ready for output.

This distributed approach, which does not distinguish between syntactic and semantic representations and processes, is broadly compatible with the tenets of the Cognitive school in linguistics (Langacker, 1987), and, more specifically, with Construction Grammar (Fillmore, 1985; Goldberg, 1998; Goldberg, 2003; Croft, 2001). CoPh is, however, more than a mere metaphor for constructions, for several reasons. First, CoPh is deeply rooted in computational principles (multidimensional similarity spaces, distributed representations), algorithmic methods (statistical inference) and neural mechanisms (receptive fields and maps) that proved instrumental in analyzing other aspects of cognition. Second, by steering the goals of syntactic (and semantic) analysis toward those of cognition in general, CoPh brings to the fore a collection of mathematical tools hitherto not considered by linguists (see Appendix B: mathematical tools). Third, an implemented model of language acquisition and processing situated within the CoPh framework provides empirical support and constraints for the construction grammar theories, as described briefly below.

3.2 Algorithmic and implementational issues: the ADIOS model

The algorithm behind this working model of language acquisition and processing (ADIOS, or Automatic DIstillation Of Structure) learns, in an unsupervised fashion, a streamlined representation of linguistic structures from untagged, large-scale natural-language corpora (Solan et al., 2003b; Solan et al., 2004; Edelman et al., 2004). The algorithm represents sentences as paths on a graph whose vertices are, initially, words. Significant patterns are defined as sets of paths in which a common prefix and suffix form a context surrounding a slot where locally distributionally equivalent (Harris, 1954) elements may appear. In each iteration, such patterns, determined by context-sensitive statistical inference, form new vertices, and the graph is rewired, leading to the emergence of progressively more complex, hierarchically structured representations. The algorithm stops

when no new patterns are found in a given iteration. Linguistic constructions thus correspond to trees composed of patterns and their associated equivalence classes. An entire utterance is typically represented by several such constructions (a Chorus of Phrases; cf. Figure 3), which may be activated to different degrees, depending on their fit to the input. Previously unseen inputs are processed by pursuing structural and lexical similarities to familiar patterns.

The probabilistic principle that drives the context-sensitive, hierarchical pattern abstraction process in the ADIOS model is related both to the notion of "suspicious coincidences" long thought to be the key to unsupervised learning in neural systems (Barlow, 1959; Barlow, 1989) and to the Minimum Description Length (MDL) criterion for representational efficiency (Bienenstock et al., 1997). Intuitively, two elements — such as two members of a potential linguistic construction or two fragments of a visual object — belong together to the extent that the probability of their joint appearance is higher than the product of the probabilities of their individual appearances; coding such elements as a unit results in a more concise representation. It has been conjectured that these principles, which can support structured learning in vision (Barlow, 1990; Edelman et al., 2002a; Edelman et al., 2002b) and in language (Bienenstock et al., 1997; Clark, 2001; Solan et al., 2003b), may provide "common foundations for cortical computation" (Phillips and Singer, 1997). The ADIOS algorithm is based on a criterion for pattern unity that is specifically adapted to the sequential nature of language and to the graph-like data structure used to represent it. It is interesting to note that the entrenchment of pattern- or construction-like units (the degree to which subjects treat them as such) depends on the corpus frequency of the corresponding sub-unit sequences (Harris, 1998), which supports the notion that the patterns postulated by the ADIOS algorithms are the psychologically real.

The implemented ADIOS model has been subjected to extensive tests, some of which focused on the acquisition of artificial languages generated by context-free grammars (CFG), and others — on learning from real natural-language corpora such as CHILDES (MacWhinney and Snow, 1985), ATIS (Moore and Carroll, 2001) and the Bible (Resnik et al., 1999); only a few of these tests can be mentioned here. The CFG experiments involved two ADIOS instances: a teacher and a student. In each of the multiple runs, the teacher was pre-loaded with a ready-made context free grammar (using the straightforward translation of CFG rules into ADIOS patterns), then used to generate a series of training corpora with up to 6400 sentences, each with up to seven levels of recursion. After training in each run i, a student-generated test corpus $C^{(i)}_{learned}$ of size 10000 was used in conjunction with a test corpus $C^{(i)}_{target}$ of the same size produced by the teacher, to calculate precision and recall. This was done by running the teacher as a parser on $C^{(i)}_{learned}$ (precision measured by the teacher's acceptance of the student-generated sentences) and the student – as a parser on $C^{(i)}_{target}$ (recall measured by the student's acceptance of novel sentences not seen during training).[5] The results — nearly

[5]Defining performance in terms of sentence acceptance amounts to testing for the so-called "weak" generativity rather than the

100% precision and about 95% recall – indicate a substantial capacity for unsupervised induction of context-free grammars even from very small corpora (Edelman et al., 2004). Promising performance has also been demonstrated for real-life language. For example, a model trained on the CHILDES data attained a level of performance considered to be "intermediate" for 9th-grade students when subjected to a standard test of English as a Second Language (ESL) proficiency (Solan et al., 2003b).

3.3 Select open questions

The proposed framework places within reach of empirical research many exciting open issues in language and cognition, some of which are outlined next.

Linking psycholinguistics to visual psychophysics. Recent psycholinguistic evidence indicates that listeners and readers routinely settle for "good enough" representations of the linguistic material they face, rather than seeking an exhaustive parse or even just a fully disambiguated semantic interpretation (Bever et al., 1998; Ferreira et al., 2002; Sanford and Sturt, 2002); cf. Figure 2, right. A unified computational approach to vision and to language should help relate these findings to the cluster of phenomena in visual psychophysics known as "change blindness" (Simons and Levin, 1997), which indicate that subjects do not fully parse visual scenes either.

From the "big picture" to the neural mechanisms. Marr and Poggio's framework calls for equal attention to the computation- and the neurobiology-level understanding of cognition. On an abstract, computational level, construction-based approaches — in particular the Chorus of Phrases — readily integrate themselves into the rest of cognition, offering along the way a useful insight into the relationship between the final representational product of language and that of vision. Goldberg's thesis — "constructions all the way down" (Goldberg, 2003) — can be taken to imply that the Chorus of Phrases evoked by an utterance or a text is just about all there is to its interpretation. There is a clear parallel between this stance and the conjecture that in vision the Chorus of Fragments is an adequate, and in fact the only reasonable, bottom line (Edelman, 2002). On the level of mechanism, however, the details have yet to be worked out.

"strong" generativity, under which the derivation/parse trees are compared instead of sentences (Roberts and Atwell, 2003). There is a good reason for this choice: any "gold standard" that can be used to evaluate strong generativity, such as the Penn Treebank (Marcus et al., 1994), invariably reflects its designers' preconceptions about language, which are often controversial among linguists themselves: the derivation trees are always stipulated, never observed. A learner who exhibits perfect weak generativity — that is, who accepts and produces all and only those sentences respectively produced and accepted by the teacher — is, for all practical purposes, perfectly successful.

Between the mechanism and the virtual machine. A crucial question is whether or not it is possible to avoid altogether the need to manipulate constructions dynamically, rather than through a pre-wired network. This is important because the act of binding a variable to a value (or inserting a constituent into a construction) dynamically is deeply problematic in the context of a neural implementation (Edelman and Intrator, 2003). The ability of humans to do algebra or to program computers attests to the existence of some mechanism in the brain that at least creates the semblance of dynamic binding. People, however, must be trained for years before they become good at this kind of symbol manipulation, and it would be prudent to make it a means of last resort in a theory of any cognitive phenomenon that is more mundane than programming in Lisp. The same consideration applies to a related issue, recursion: although it has been recently reaffirmed by some linguists as the epitome of human uniqueness (Hauser et al., 2002), humans are notoriously bad at deep recursion (Gibson and Thomas, 1999). In comparison, shallow recursion, as well as the manipulation of complexity-controlled constructions, can be handled by finite means such as the ADIOS representation (Solan et al., 2003b). These considerations suggest that dynamic binding and deep recursion may both be supported in the brain by a virtual machine that is difficult to build and expensive to maintain and operate, and that is at least once removed from the neural mechanisms that are so good at supporting everyday cognition; cf. (Dennett, 1991, p.209).

4 Conclusions

Computational cognitive science holds that a comprehensive theory of any information processing task must lead to its understanding on several levels of abstraction (Marr and Poggio, 1977). Although distinct, these levels cannot be studied independently, lest theorizing loses touch with psychological and neurobiological reality, or, conversely, the neurobiology becomes too myopic (Edelman, 1999). Accordingly, the framework for computational cognitive linguistics outlined here is informed by the top-down computational and algorithmic principles of context-dependent probabilistic learning and is based on a bottom-up implementational scheme that is ubiquitous in the brain: computing with structured connections, which carry dynamically unfolding neural activation patterns and which support low-dimensional, distributed, redundant, graded representations. The vision of language it offers should boost cognitively oriented theories such as Construction Grammar (Croft, 2001; Goldberg, 2003) and help connect them to rich repositories of computational knowledge (from learning theory, probability and information, and natural language processing) and empirical data (from psychophysics and neuroscience) about the brain.

Acknowledgments

I thank Zach Solan, Eytan Ruppin, David Horn, Roni Katzir, Rick Dale and Bo Pedersen for stimulating discussions, and Michael Spivey, Shweta Narayan and two anonymous reviewers for constructive comments.

References

Aho, A. V., Hopcroft, J. E., and Ullman, J. D. (1974). *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, MA.

Barlow, H. B. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. In *The mechanisation of thought processes*, pages 535–539. H.M.S.O., London.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, 1:295–311.

Barlow, H. B. (1990). Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception. *Vision Research*, 30:1561–1571.

Baum, E. B. (2004). *What is thought?* MIT Press, Cambridge, MA.

Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton University Press, Princeton, NJ.

Bernárdez, E. (1999). Some reflections on the origins of cognitive linguistics. *Journal of English Studies*, 1:9–27.

Bever, T. G., Sanz, M., and Townsend, D. J. (1998). The Emperor's psycholinguistics. *J. of Psycholinguistic Research*, 27:261–284.

Bienenstock, E., Geman, S., and Potter, D. (1997). Compositionality, MDL priors, and object recognition. In Mozer, M. C., Jordan, M. I., and Petsche, T., editors, *Neural Information Processing Systems*, volume 9. MIT Press.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. (1987). Occam's razor. *Information Processing Letters*, 24:377–380.

Bod, R. (1998). *Beyond grammar: an experience-based theory of language*. CSLI Publications, Stanford, US.

Campbell, F. W. and Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. *J. Physiol. (Lond.)*, 197:551–566.

Chatterjee, A. (2001). Language and space: some interactions. *Trends in Cognitive Sciences*, 5:55–61.

Chen, S. and Donoho, D. L. (1994). Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 41–44, Pacific Grove, CA. IEEE Comput. Soc. Press.

Chipere, N. (1997). Individual differences in syntactic skill. *Working Papers in English and Applied Linguistics*, 4:1–32.

Chipere, N. (2001). Native speaker variations in syntactic competence: implications for first language teaching. *Language Awareness*, 10:107–124.

Chomsky, N. (1957). *Syntactic Structures*. Mouton & Co., the Hague.

Christiansen, M. H. and MacDonald, M. C. (2003). Processing of recursive sentence structure: Testing predictions from a connectionist model. in preparation.

Churchland, P. S. and Sejnowski, T. J. (1992). *The Computational Brain*. MIT Press, Cambridge, MA.

Clark, A. (2000). *A theory of sentience*. Oxford University Press, Oxford.

Clark, A. (2001). *Unsupervised Language Acquisition: Theory and Practice*. PhD thesis, COGS, University of Sussex.

Coslett, H. B. (1999). Spatial influences on motor and language function. *Neuropsychologia*, 37:695–706.

Croft, W. (2001). *Radical Construction Grammar: syntactic theory in typological perspective*. Oxford University Press, Oxford.

Cummins, R. (1996). *Representations, Targets, and Attitudes*. MIT Press, Cambridge, MA.

Damasio, A. R. and Tranel, D. (1993). Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Science*, 90:4957–4960.

Dennett, D. C. (1991). *Consciousness explained*. Little, Brown & Company, Boston, MA.

Edelman, S. (1998). Representation is representation of similarity. *Behavioral and Brain Sciences*, 21:449–498.

Edelman, S. (1999). *Representation and recognition in vision*. MIT Press, Cambridge, MA.

Edelman, S. (2001). Neural spaces: a general framework for the understanding of cognition? a commentary on Shepard. *Behavioral and Brain Sciences*, 24:664–665.

Edelman, S. (2002). Constraining the neural representation of the visual world. *Trends in Cognitive Sciences*, 6:125–131.

Edelman, S. and Christiansen, M. H. (2003). How seriously should we take Minimalist syntax? A comment on Lasnik. *Trends in Cognitive Science*, 7:60–61.

Edelman, S., Hiles, B. P., Yang, H., and Intrator, N. (2002a). Probabilistic principles in unsupervised learning of visual structure: human data and a model. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 19–26, Cambridge, MA. MIT Press.

Edelman, S. and Intrator, N. (1997). Learning as extraction of low-dimensional representations. In Medin, D., Goldstone, R., and Schyns, P., editors, *Mechanisms of Perceptual Learning*, pages 353–380. Academic Press.

Edelman, S. and Intrator, N. (2003). Towards structural systematicity in distributed, statically bound visual representations. *Cognitive Science*, 27:73–109.

Edelman, S., Intrator, N., and Jacobson, J. S. (2002b). Unsupervised learning of visual structure. In Bülthoff, H. H., Wallraven, C., Lee, S.-W., and Poggio, T., editors, *Proc. 2nd Intl. Workshop on Biologically Motivated Computer Vision*, volume 2525 of *Lecture Notes in Computer Science*, pages 629–643. Springer.

Edelman, S., Solan, Z., Horn, D., and Ruppin, E. (2004). Bridging computational, formal and psycholinguistic approaches to language. In *Proc. of the 26th Conference of the Cognitive Science Society*, Chicago, IL.

Eliasmith, C. and Thagard, P. (2001). Integrating structure and meaning: A distributed model of analogical mapping. *Cognitive Science*, 25:245–286.

Erman, B. and Warren, B. (2000). The idiom principle and the open-choice principle. *Text*, 20:29–62.

Ferreira, F., Bailey, K. G. D., and Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11:11–15.

Fillmore, C. J. (1985). Syntactic intrusion and the notion of grammatical construction. *Berkeley Linguistic Society*, 11:73–86.

Fodor, J. A., Bever, T. G., and Garrett, M. F. (1974). *The psychology of language*. McGraw Hill, New York.

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291:312–316.

Gallistel, C. R. (1990). *The organization of learning*. MIT Press, Cambridge, MA.

Gibson, E. and Pearlmutter, N. (1998). Constraints on sentence comprehension. *Trends in Cognitive Sciences*, 2:262–268.

Gibson, E. and Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, 14:225–248.

Gibson, J. J. (1957). Survival in a world of probable objects. *Contemporary Psychology*, 2:33–35.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, Chicago.

Goldberg, A. E. (1998). Patterns of experience in patterns of language. In Tomasello, M., editor, *The new psychology of language*, pages 203–219. Erlbaum, Mahwah, NJ.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences*, 7:219–224.

Harris, C. L. (1998). Psycholinguistic studies of entrenchment. In Koenig, J., editor, *Conceptual Structures, Language and Discourse*, volume 2. CSLI, Berkeley, CA.

Harris, Z. S. (1954). Distributional structure. *Word*, 10:140–162.

Hartman, E. J., Keeler, J. D., and Kowalski, J. M. (1990). Layered neural networks with Gaussian hidden units as universal approximations. *Neural Computation*, 2:210–215.

Hauser, M., Chomsky, N., and Fitch, T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298:1569–1579.

Hume, D. (1740). *A Treatise of Human Nature*.

Iliopoulos, C. S. and Smyth, W. F. (1998). On-line algorithms for k-covering. In *Proceedings of the 9-th Australasian Workshop on Combinatorial Algorithms*, volume 6, pages 97–106.

Intrator, N. and Edelman, S. (1997). Learning low dimensional representations of visual objects with extensive use of prior knowledge. *Network*, 8:259–281.

Jiang, T. and Li, M. (1996). DNA sequencing and string learning. *Math. Syst. Theory*, 26:387–405.

Joerges, J., Küttner, A., Galizia, C. G., and Menzel, R. (1997). Representations of odors and odor mixtures visualized in the honeybee brain. *Nature*, 387:285–288.

Joshi, A. and Schabes, Y. (1997). Tree-Adjoining Grammars. In Rozenberg, G. and Salomaa, A., editors, *Handbook of Formal Languages*, volume 3, pages 69–124. Springer, Berlin.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, New York.

Kersten, D. and Schrater, P. R. (2000). Pattern Inference Theory: A probabilistic approach to vision. In Mausfeld, R. and Heyer, D., editors, *Perception and the Physical World*. John Wiley & Sons, Chichester.

Knill, D. and Richards, W., editors (1996). *Perception as Bayesian Inference*. Cambridge University Press, Cambridge.

Lancet, D. (1991). The strong scent of success. *Nature*, 351:275–276. News and Views.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Landsberg, M. E., editor (1995). *Syntactic iconicity and linguistic freezes*. Mouton de Gruyter, Berlin.

Langacker, R. W. (1987). *Foundations of cognitive grammar*, volume I: theoretical prerequisites. Stanford University Press, Stanford, CA.

Lasnik, H. (2002). The Minimalist Program in syntax. *Trends in Cognitive Science*, 6:432–437.

Lodhi, H., Shawe-Taylor, J., Cristianini, N., and Watkins, C. J. (2001). Text classification using string kernels. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 563–569. MIT Press.

Maass, W., Natschläger, T., and Markram, H. (2003). Computational models for generic cortical microcircuits. In Feng, J., editor, *Computational Neuroscience: A Comprehensive Approach*. CRC-Press, Boca Raton, FL. to appear.

MacWhinney, B. and Snow, C. (1985). The Child Language Exchange System. *Journal of Computational Linguistics*, 12:271–296.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. (1994). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Marr, D. (1982). *Vision*. W. H. Freeman, San Francisco, CA.

Marr, D. and Poggio, T. (1977). From understanding computation to understanding neural circuitry. *Neurosciences Res. Prog. Bull.*, 15:470–488.

Miller, G. (1962). Some psychological studies of grammar. *American Psychologist*, 17:748–762.

Minsky, M. (1985). *The Society of Mind*. Simon and Schuster, New York.

Moore, B. and Carroll, J. (2001). Parser comparison – context-free grammar (CFG) data. online at http://www.informatics.susx.ac.uk/research/nlp/carroll/cfg-resources/.

Mumford, D. (1994). Neuronal architectures for pattern-theoretic problems. In Koch, C. and Davis, J. L., editors, *Large-scale neuronal theories of the brain*, chapter 7, pages 125–152. MIT Press, Cambridge, MA.

Mumford, D. (1997). The mathematical modeling of cortical functioning and thought. In *Proc. Norbert Wiener Centennial Conference*, Providence, RI. Amer. Math. Society.

Nakano, Y., Felser, C., and Clahsen, H. (2002). Antecedent priming at trace positions in Japanese long-distance scrambling. *Journal of Psycholinguistic Research*, 31:531–570.

Neisser, U. (1976). *Cognition and reality*. Freeman, San Francisco, CA.

Ochsner, K. N., Chiu, C.-Y. P., and Schacter, D. L. (1994). Varieties of priming. *Current Opinion in Neurobiology*, 4:189–194.

Op de Beeck, H. and Vogels, R. (2000). Spatial sensitivity of Macaque inferior temporal neurons. *J. Comparative Neurology*, 426:505–518.

Op de Beeck, H., Wagemans, J., and Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience*, 4:1244–1252.

O'Regan, J. K. (1992). Solving the real mysteries of visual perception: The world as an outside memory. *Canadian J. of Psychology*, 46:461–488.

Phillips, W. A. and Singer, W. (1997). In search of common foundations for cortical computation. *Behavioral and Brain Sciences*, 20:657–722.

Plate, T. A. (1995). Holographic Reduced Representations. *IEEE Transactions on Neural Networks*, 6:623–641.

Poggio, T. (1990). A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology*, LV:899–910.

Poggio, T. and Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266.

Postal, P. M. (2004). *Skeptical linguistic essays*. Oxford University Press, New York.

Pouget, A., Zemel, R. S., and Dayan, P. (2000). Information processing with population codes. *Nature Review Neuroscience*, 1:125–132.

Pulvermüller, F. (2002). A brain perspective on language mechanisms: from discrete neuronal ensembles to serial order. *Progress in Neurobiology*, 67:85–111.

Pulvermüller, F., Preissl, H., Lutzenberger, W., and Birbaumer, N. (1996). Brain rhythms of language: nouns versus verbs. *Eur. J. Neurosci.*, 8:937–941.

Rao, S. C., Rainer, G., and Miller, E. K. (1997). Integration of what and where in the primate prefrontal cortex. *Science*, 276:821–824.

Reitman, W., Nado, R., and Wilcox, B. (1978). Machine perception: what makes it so hard for computers to see? In Savage, C. W., editor, *Perception and cognition: issues in the foundations of psychology*, volume IX of *Minnesota studies in the philosophy of science*, pages 65–87. University of Minnesota Press, Minneapolis, MN.

Resnik, P., Olsen, M. B., and Diab, M. (1999). The Bible as a parallel corpus: annotating the 'Book of 2000 Tongues'. *Computers and the Humanities*, 33:129–153.

Rissanen, J. (1987). Minimum description length principle. In Kotz, S. and Johnson, N. L., editors, *Encyclopedia of Statistic Sciences*, volume 5, pages 523–527. J. Wiley and Sons.

Roberts, A. and Atwell, E. (2003). The use of corpora for automatic evaluation of grammar inference systems. In Archer, D., Rayson, P., Wilson, A., and McEnery, T., editors, *Proc. of the Corpus Linguistics 2003 conference*. UCREL, University of Lancaster.

Salzman, C. D., Britten, K. H., and Newsome, W. T. (1990). Cortical microstimulation influences perceptual judgements of motion direction. *Nature*, 346:174–177.

Sanford, A. J. and Sturt, P. (2002). Depth of processing in language comprehension: not noticing the evidence. *Trends in Cognitive Sciences*, 6:382–386.

Shamma, S. (2001). On the role of space and time in auditory processing. *Trends in Cognitive Sciences*, 5:340–348.

Shepard, R. N. (1975). Form, formation, and transformation of internal representations. In Solso, R. L., editor, *Information processing and cognition: the Loyola Symposium*, pages 87–122, Hillsdale, NJ. Erlbaum.

Shepherd, G. M. (1992). Modules for molecules. *Nature*, 358:457–458. News and Views.

Simone, R. (1995). Iconic aspects of syntax: a pragmatic approach. In Simone, R., editor, *Iconicity in language*, pages 153–169. John Benjamins.

Simons, D. J. and Levin, D. T. (1997). Change blindness. *Trends in Cognitive Science*, 1:261–267.

Skiena, S. and Sundaram, G. (1995). Reconstructing strings from substrings. *Journal of Computational Biology*, 2:333–353.

Solan, Z., Horn, D., Ruppin, E., and Edelman, S. (2004). Unsupervised context sensitive language acquisition from a large corpus. In Saul, L., editor, *Advances in Neural Information Processing*, volume 16, Cambridge, MA. MIT Press.

Solan, Z., Ruppin, E., Horn, D., and Edelman, S. (2003a). Automatic acquisition and efficient representation of syntactic structures. In Thrun, S., editor, *Advances in Neural Information Processing*, volume 15, Cambridge, MA. MIT Press.

Solan, Z., Ruppin, E., Horn, D., and Edelman, S. (2003b). Unsupervised efficient learning and representation of language structure. In Alterman, R. and Kirsh, D., editors, *Proc. 25th Conference of the Cognitive Science Society*, Hillsdale, NJ. Erlbaum.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In Solla, S. A., Leen, T. K., and Mller, K.-R., editors, *NIPS (Advances in Neural Information Processing Systems)*, volume 12, Cambridge, MA. MIT Press.

Tomasello, M. (1998). Introduction: a cognitive-functional perspective on language structure. In Tomasello, M., editor, *The new psychology of language*, pages vii–xxiii. Erlbaum, Mahwah, NJ.

Tulving, E. and Schacter, D. L. (1990). Priming and human memory systems. *Science*, 247:301–306.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27:1134–1142.

Ward, R., Danziger, S., Owen, V., and Rafal, R. (2002). Deficits in spatial coding and feature binding following damage to spatiotopic maps in the human pulvinar. *Nature Neuroscience*, 5:99–101.

Wiggs, C. L. and Martin, A. (1998). Properties and mechanisms of perceptual priming. *Curr. Opin. Neurobiol.*, 8:227–233.

Wilson, H. R. and Bergen, J. R. (1979). A four mechanism model for threshold spatial vision. *Vision Research*, 19:19–32.

Wittgenstein, L. (1961). *Tractatus Logico-philosophicus*. Routledge, London. trans. D. F. Pears and B. F. McGuinness.

Zadrozny, W. (1994). From compositional to systematic semantics. *Linguistics and philosophy*, 17:329–342.

Appendix A: some useful mathematical concepts

The following is a brief glossary of some of the mathematical concepts used in this chapter to describe the state of the art in the computational understanding of cognition.

Functions. Much of the essence of the idea of learning in cognition is captured by the mathematical notion of a function. Formally, a function is a mapping (a specification of correspondence) from one set, which is called the domain, to another, called the range. In the example of Figure 4, left, the function f maps the element a_1 in the domain A to b_2 in the range B, a_2 to b_1, and so on. The mapping that defines a function must be unequivocal in that a given element in the domain cannot be mapped to more than one element in the range (although any number of elements in the domain can be mapped to the same one in the range, as illustrated in Figure 4, left). Thus, every time a cognitive system learns to attach a valence to a stimulus or associate it with a response, it learns a function defined over the set of possible stimuli. Note that the sets in question may include variables that are internal or external to the system, and are, in general, multidimensional spaces (see below).

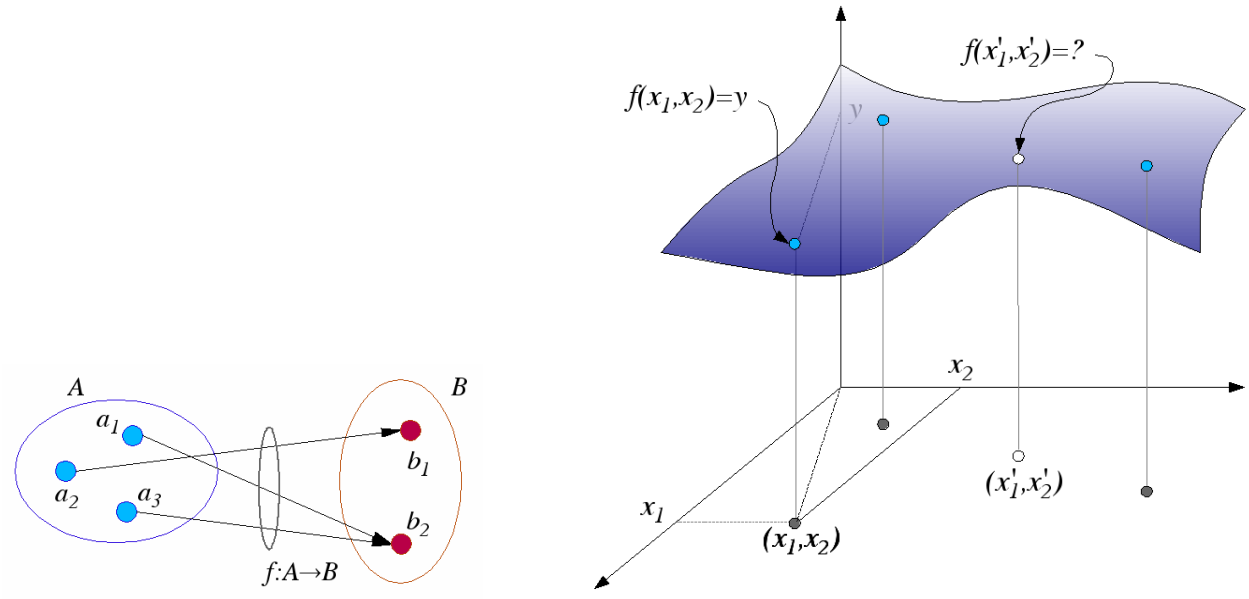

Figure 4: Illustrations of the basic idea of a function (*left*) and of learning a function from examples (*right*); see text for explanations.

Multidimensional spaces. A point in an n-dimensional space can be thought of as an ordered collection of n independent measurements (feature values) that define some entity of interest. For example, the state

(configuration) of a human arm can be described by four numbers: three to specify the angular position of the spherical shoulder joint and one to specify the elbow angle (including the hand introduces many more degrees of freedom and hence many more dimensions). Thus, each configuration of the arm corresponds to a point in a four-dimensional space, and learning to position an arm amounts to learning a function from the space of the relevant muscles to this configuration space.

Learning and function approximation. In this conceptual framework, learning from examples amounts to interpolating an unknown function from the given correspondences between some of the elements in its domain and its range (Poggio, 1990). In the illustration shown in Figure 4, right, the function f maps a two-dimensional space (the horizontal plane) into a one-dimensional space (the vertical axis); it can be thought of as a surface defined over the plane. Given the height of the surface (the value of the function) at several points (filled circles), one can try to determine its value at a new point (open circle). In the context of controlling an arm, for instance, this would amount to a generalization of the previously available control settings to determine a setting for a new target configuration. The mathematical, psychophysical and neurobiological aspects of this approach as applied to visual object recognition are described in (Edelman, 1999).

Dimensionality reduction. The mathematical tools associated with the concept of multidimensional spaces can be applied to the description of brain states and of their evolution over time (Mumford, 1994; Mumford, 1997; Edelman, 2001). In the most straightforward fashion, one dimension is assigned to describe the activity level of each neuron (Churchland and Sejnowski, 1992). This results in a space with many billions of dimensions; apart from the convenience of the mental picture of a brain state as a point in such a space, not much is gained, because of the overwhelming computational complexity of dealing with such high-dimensional spaces. Because certain kinds of brain representations are necessarily high-dimensional,[6] cognition must involve, at various stages, the reduction of dimensionality to a level that is computationally manageable. The dimensionality of a representation can be reduced by projecting it into a lower-dimensional space; for details and applications of this idea, see (Edelman and Intrator, 1997). Note that such a projection is a function, which can be learned from examples (as described above).

Probabilities and statistical inference by density estimation. The most knowledge one can possess about any situation that involves information processing is the joint probability of all the relevant variables, $P(X_1, X_2, \ldots, X_n)$. This profound insight can be traced back to the writings of David Hume:

[6] For example, the retinal signal, whose raw dimensionality runs in the millions (it is equal to the number of axons in the optic nerve), or the motor control signal, whose dimensionality is at least the same as the number of distinct muscles in the body.

> "All kinds of reasoning consist in nothing but a comparison, and a discovery of those relations, either constant or inconstant, which two or more objects bear to each other." (Hume, 1740, Part III, Sect. II)
>
> "An experiment loses of its force, when transferr'd to instances, which are not exactly resembling; tho' 'tis evident it may still retain as much as may be the foundation of probability, as long as there is any resemblance remaining." (Part II, Sect. XII)
>
> "...all knowledge resolves itself into probability..." (Part IV, Sect. I)

Hume's realization of the central and crucial role of *statistical inference* in knowledge generation (that is, learning) has been developed by many others, including his contemporary Thomas Bayes, the pioneering statisticians Karl Pearson and Ronald A. Fisher, and the neurobiologist Horace B. Barlow. Their combined insights led to the modern applications of inference to vision and other senses (Barlow, 1990; Knill and Richards, 1996), as well as to language (Manning and Schütze, 1999).

The conception of visual learning as inference is naturally complemented by the emerging view of perception as statistical *decision making*, stated cogently in the following passage by the originator of the ecological theory of perception, the psychologist J. J. Gibson:

> "...the percept is always a wager. Thus uncertainty enters at *two* levels, not merely one: the configuration may or may not indicate an object, and the cue may or may not be utilized at its true indicative value." (Gibson, 1957)

As a simple concrete example, consider a perceptual system that monitors the values of three features, X_1, X_2 and X_3, which are related to the presence of four possible objects as coded by $X_4 = \{O_1, O_2, O_3, O_4\}$ seen at one of two possible angles, $X_5 = \{A_1, A_2\}$. In this case, the knowledge of the joint probability density $P(X_1, X_2, X_3, X_4, X_5)$ would allow the system to estimate the conditional probability of each combination of object and angle, given the observed feature values: $P(X_4, X_5|X_1, X_2, X_3)$. This information, in turn, would suffice to support optimal decision making on the basis of the maximum likelihood criterion (which combination of values of X_4 and X_5 is most likely in the light of the measurements?). It is important to note that the same tools used to reason about — and to learn — a function from examples are also applicable to probability densities.

The graph data structure. The kind of data that arise in the context of natural language processing, namely, sequences defined over a finite alphabet or lexicon of discrete symbols, can be captured by graphs — a useful data structure that is extensively studied in computer science (Aho et al., 1974). Formally, a graph is specified

by two sets: the vertices, and the edges (which may be directed) that connect some of them. Figure 5 illustrates a simple graph whose vertices are labeled by the symbols {begin, end, I, and, you, read, news, the, run, today, slept}. Traversing some of its directed edges while writing down the encountered vertices yields sentences such as "I read the news today", "you run", "I slept", and "you and I slept". The ADIOS algorithm outlined in section 3.2 uses a graph of this type as its basic data structure.

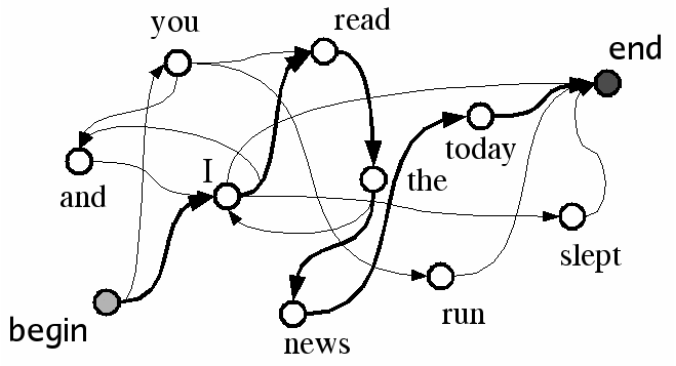

Figure 5: A very simple graph of the kind used as the basic data structure by the ADIOS algorithm.

Appendix B: mathematical tools for computational cognitive linguistics

From the standpoint of methodology, a computationally motivated approach to cognitive linguistics does not imply pitching "mathematics versus psychology" – an expression used by Tomasello as a section heading in his introduction to *The New Psychology of Language* (Tomasello, 1998, p.ix). Rather, the usual tools of mathematical linguistics (such as formal languages, λ-calculus and symbolic logic) should be supplemented by new ones. Some of these, which proved well-suited for analyzing distributed representations in various areas of cognition, are listed below.

Syntax: from constituent trees to string cover. Computational learning theory offers various tools capable of dealing with distributed, potentially over-complete (Chen and Donoho, 1994) representations of sequence data. One of these is string kernels, a representation that tallies the occurrences of specific symbols in specific locations, and supports reasoning about global properties of the sequence probed in this manner and, in particular, about features that can help classify it (Lodhi et al., 2001). Similar methods are increasingly in demand in computational biology, because of the sequential nature of the data in both domains, and, specifically, because of the close analogy between text analysis by the identification of multiple, overlapping local patterns on the one hand, and hybridization approaches to DNA sequencing on the other hand. Recent developments in this field include derivations of the algorithmic complexity of specifying a string by its substrings (Skiena

and Sundaram, 1995; Jiang and Li, 1996; Iliopoulos and Smyth, 1998). This approach is distinct from (and more relevant for our present purposes than) treebank-based parsing (combining multiple local or partial parse trees (Joshi and Schabes, 1997; Bod, 1998)) in that the cover it seeks need be neither precise nor exclusive.

Semantics: from functions to constructions and relations. According to the Chorus of Phrases metaphor (Figure 3), the representation of a sentence by an ensemble of active units can be approximately described as a *relation* (namely, as the subset of units whose activity exceeds some threshold). As in Construction Grammar (Goldberg, 2003), this representation captures both semantic and syntactic information about the input. Interestingly, recent work in computational semantics addressing various problematic aspects of compositionality suggests that systematicity of meaning is better served by defining meaning as a relation over sentence parts (Zadrozny, 1994), rather than as a function of the parts, as stipulated by the classical, Fregean approach.

Acquisition: from parameter setting to structure discovery. The ascendancy of the generative grammar and its accompanying innateness postulate over competing distributional and behaviorist ideas in the 1960s can be ascribed in a large part to the inadequacy of the contemporary statistical inference methods and the perceived inability of association-based learning to handle recursion. Statistics, however, need not be limited to counting word frequencies: in computer science, the integration of advanced statistical inference (including Bayesian methods, the Minimum Description Length principle and other related information-theoretic tools), progress in computational learning theory, efficient algorithms, and cheap hardware led to important conceptual progress, as well as to practical achievements (Manning and Schütze, 1999). Likewise, learning need not be limited to the establishment of pairwise associations: bounded-depth recursively structured patterns can be learned from examples, by efficient algorithms that rely on modern statistical inference (Solan et al., 2003b; Solan et al., 2004); see (Clark, 2001) for an overview of the recent progress in this field.